Thèse n° 9811

Towards Novel Evaluation Methods for Social Dialog Systems

Présentée le 22 septembre 2023

Faculté informatique et communications
Groupe SCI IC PFP
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Ekaterina SVIKHNUSHINA

Acceptée sur proposition du jury

Prof. K. Aberer, président du jury
Dr. P. Pu Faltings, directrice de thèse
Dr M. Zhou, rapporteuse
Dr V. Q. Liao, rapporteuse
Prof. A. Bosselut, rapporteur

École
polytechnique
fédérale
de Lausanne

2023

Acknowledgements

Starting my Ph.D. was a life-changing experience. The last five years have been an intense period of personal and professional growth, with its challenges and victories along the way. Throughout this journey, I met numerous wonderful people whose involvement, support, and guidance allowed me to advance to the point where I find myself now. I would like to take this moment to express my gratitude to them.

First and foremost, I would like to express my sincere gratitude to my advisor, Dr Pearl Pu, for her trust in me and for giving me the opportunity to pursue research on such an exciting topic, which was quite a new direction for me back in the day. Thank you for guiding me along this twisting path, giving me valuable advice, encouraging and coaching me on how to overcome difficulties, showing me new research methods, and yet giving me the freedom to explore various initiatives. I am immensely grateful to you for introducing me to the fields of human-computer interaction (HCI) and user experience (UX) research, which I find fascinating. And I especially appreciate your kind participation and support through difficult periods that go beyond the academic setting. You have made me a better researcher and also a better person.

I would like to thank my committee members, Prof. Antoine Bosselut, Dr Michelle Zhou, and Dr Vera Q. Liao for taking the time to read my thesis and providing valuable feedback, thought-provoking questions, and constructive suggestions. I would also like to acknowledge Prof. Karl Aberer for presiding over the committee.

My warmest acknowledgments go to my previous and present colleagues from the HCI Group: Anuradha, Yubo, Igor, and Kavous. Thank you for always being there to engage in fruitful discussions, brainstorm research ideas, and lend a helping hand whenever needed, be it organizing a dry-run, fixing a technical issue, or providing feedback on the experiment. I would also like to thank Patricia and Karin for their great help with dealing with administrative processes and for helping me practice my French. Last but not least, I am thankful to Valentina, a former Ph.D. graduate from the HCI Group, who shared her experience and useful advice with me, making me much more prepared for moving to Switzerland.

During my Ph.D. I had an opportunity to work with many talented undergraduate and master students who supported my research and from whom I learned a lot. Special thanks go to Alexandru, Iuliana, Siran, Anastasiia, and Mohamed for your genuine interest, motivation, and continuous engagement in the projects that we were pursuing together.

I would also like to thank my fellow EPFL colleagues: Negar for our regular encouraging status updates in between our offices, Panayiotis for reminding me that I am a sunny person, Francesco for making any occasion fun, be it partying at the lake or fixing bugs after midnight, Zeinab for helpful advice on internship preparation processes, Vinitra, Akhil, and Viktor for being involved in so many student initiatives and having engaged me in many fun activities as well, Andrei for all random chitchats that we had upon bumping into each other in the neighborhood and non-random discussions on how to proceed further at times of uncertainty. Special thanks also to colleagues from LTS2 Acoustics, LWE, and MAG groups for sharing the vast majority of lunches and quite some beers with me.

Next, I would like to express my gratitude to the amazing people who mentored, hosted, and supported me during my internship at Google. Thank you, Svetlana, for all your precious time and advice, for opening this door for me, and for introducing me to many inspiring people. I cannot express in words how thankful I am to Iva, who put together a captivating project proposal for my internship and did her best to set me up for success. I am immensely grateful to my hosts, Marcel and Anna, for their incredible mentorship, dedication, encouragement, and trust in me. I also thank Jeremy for his valuable expert advice and motivation to publish the results of the project. I would also like to thank Landen, Mausam, Tibor, Taja, Melanie, Caterina, Eunice, Alexander, Alex, Fedor, Andrii, and many more people whom I met over the time of my internship for creating a sense of belonging, sharing awesome ideas, and making my time at Google so special.

During my Ph.D. journey, I was also lucky to become a part of the Women in Voice community and meet like-minded people interested and working in voice and language technologies. I would like to thank my previous and present fellow ambassadors of the Swiss chapter: Caterina, Karen, and Cyrielle. It is a pleasure to know you and to collaborate with you on this initiative. I have always found inspiration, learned something new, and established meaningful connections at our community events, which couldn't have been possible without your wholehearted involvement. Thank you for having me on board!

Another initiative that I pursued during the last several years was AERUS, an official association of Russian-speaking students at EPFL. I am grateful to Evgenii, Ivan, and Dmitry for starting this association and welcoming me as a part of the board. I would like to thank all the other board members who helped us run AERUS throughout the years: Fedor, Igor, Evan, Dominic, Aygul, Asya, – as well as the Russian-speaking community on campus and beyond without whose engagement our activities would have lost all their meaning. I am glad that these enriching interactions led to many new friendships and connections that continue beyond the campus context. Special thanks to Emil and Irina for organizing their cozy “tea meetings”, fostering profound reconnection and thoughtful reflection on the timeless artwork.

Many know that one of the side activities that kept me moving (pun intended) through my Ph.D. was biking. Started as an impromptu activity during the COVID pandemic, it eventually grew into my favorite hobby that let me meet an array of great friends and share lots of fun and

memorable moments with them. I deeply appreciate all the bright experiences that we had while cycling or mountain biking with Maxime, Valentin, Michel, Cristian, Grigorii, Nikolai, Mathias, Andrei, Mihailo, Stella, Manon, Igor, Miranda, Bence, and many others. It brings me joy to see how these bonds extend into other mutual activities beyond biking. I would also like to thank Pascal for his extremely motivating trainings and for teaching me to “enjoy” struggles to eventually relish the rewards.

There are people in my life whose friendship and connection go beyond the ordinary. Not only have we embarked on numerous fun endeavors and shared lots of enjoyable experiences but you have also been there to support me in times of trouble, difficulties, and uncertainties, being so kind, empathetic, and encouraging. Thank you, Dasha and Andrey, Miroslav and Borjana, Rahim (Behzad) and Zohreh, Roma, Nastya, Evann, and Yannis.

I would like to conclude by acknowledging my immediate and extended families in Russia, Ukraine, and Canada – people who were and continue being my greatest support. I would like to especially thank my grandfather, dedchik Kolya, for his “strict but fair” guidance all the way from my birth and my grandfather, ded Petr, for checking regularly on my achievements. Thank you, my dear parents, mama Natasha and papa Andrey, for encouraging me to be more ambitious in different ways, trusting in me, and being unconditionally supportive throughout my whole life. I am also thankful to my brother Misha, for setting an extraordinary example of growth, confidence, and audacity, for giving me boosts of motivation when I feel self-doubt, and generally for being a kind and open person. Lastly, this journey would have never been possible without my wonderful husband and my best friend. Stas, all my love and gratitude go to you. Throughout this incredible adventure together, you have taken on countless roles, supporting me in all my endeavors. You have more belief in me than I could ever have in myself, and knowing that drives me to keep pushing forward. Thank you for all your love, patience, care, and unforgettable time together.

Writing these acknowledgments was a satisfying moment of reflection on the last 5 years of my life, which has made me realize how blessed I am to have shared moments with so many special people. Even if I somehow forgot to mention your name above (for which I am truly sorry), I would like to thank every one of you who has been part of my Ph.D. journey. This is thanks to you that I made it to this point.

Lausanne, 20 July 2023 Ekaterina Svikhnushina

Abstract

Language has shaped human evolution and led to the desire to endow machines with language abilities. Recent advancements in natural language processing enable us to achieve this breakthrough in human-machine interaction. However, introducing conversational agents with enhanced language skills raises concerns about their emotional and social engagement. To ensure acceptance, control and evaluation mechanisms must be established. Meanwhile, creating meaningful evaluation metrics for social chatbots is challenging due to the new and undefined nature of this field, lacking clear design guidelines.

In this thesis, we contribute novel, effective evaluation frameworks for social chatbots developed based on human-centered research principles. The thesis is structured into three parts.

The first part introduces two studies that explore users' expectations of conversational chatbots and their connection to present experiences. The initial study employs qualitative semi-structured interviews and quantitative survey analysis to establish a model of essential social qualities expected from chatbots: politeness, entertainment, attentive curiosity, and empathy (PEACE). The second study examines online chatbot reviews and reveals a discrepancy between users' expectations and their current experiences, highlighting the need for chatbots to possess more advanced social capabilities.

The second part of the thesis focuses on attentive curiosity, an essential element that has received limited attention in the study of social chatbots. We propose EQT, a taxonomy of tags to differentiate between different functions of empathetic questions in social interactions. Additionally, we develop automatic classifiers for these labels, allowing us to investigate which question-asking strategies are most effective in specific emotional contexts. This analysis sheds light on the suitability of various approaches for fostering engagement and understanding in social conversations.

In the third part, we expand upon our earlier findings and create comprehensive evaluation frameworks for social chatbots. First, we introduce iEval, a human evaluation framework specifically designed to capture users' subjective perceptions of their conversational partners during interactive exchanges. Using this framework, we benchmark four state-of-the-art empathetic chatbots and examine discourse factors that account for the differences in their performance levels. Additionally, we showcase how our evaluation framework can be automated by using prompting of the latest large language models. This enables us to approximate live user studies and achieve a very strong correlation with human judgment.

The novel findings presented here enhance our understanding of user interaction with conversational technologies. Moreover, the developed evaluation criteria and frameworks provide valuable insights and tools for shaping and informing the design of future social chatbots.

Key words: human-computer interaction, conversational agents, chatbots, emotional intelligence, social intelligence, adoption, evaluation, user studies, user experiments, visualization

Résumé

Le langage a façonné l'évolution humaine et a suscité le désir de doter les machines de capacités linguistiques. Les progrès récents dans le domaine du traitement du langage naturel nous permettent de réaliser cette percée dans l'interaction homme-machine. Cependant, l'introduction d'agents conversationnels dotés de compétences linguistiques avancées soulève des inquiétudes quant à leur engagement émotionnel et social. Pour assurer leur acceptation, des mécanismes de contrôle et d'évaluation doivent être mis en place. Pourtant, la création de mesures d'évaluation significatives pour les chatbots sociaux reste un défi en raison de la nature nouvelle et non définie de ce domaine qui manque de lignes directrices claires pour leur conception.

Dans cette thèse, nous proposons de nouveaux cadres d'évaluation efficaces pour les chatbots sociaux, développés sur la base de principes de recherche centrés sur l'humain. La thèse est structurée en trois parties.

La première partie présente deux études qui explorent les attentes des utilisateurs vis-à-vis des chatbots conversationnels et si elles se reflètent dans les expériences actuelles. La première étude exploite des entretiens qualitatifs semi-structurés et une analyse d'enquête quantitative pour établir un modèle de qualités sociales essentielles attendues des chatbots : politesse, divertissement, curiosité attentive et empathie (PEACE). La seconde étude examine les critiques en ligne sur les chatbots et révèle un écart entre les attentes des utilisateurs et leurs expériences actuelles, soulignant la nécessité pour les chatbots de posséder des capacités sociales plus avancées.

La deuxième partie de la thèse se concentre sur la curiosité attentive, un élément essentiel ayant reçu une attention limitée dans l'étude des chatbots sociaux. Nous proposons EQT, une taxonomie de tags pour différencier les fonctions des questions empathiques dans les interactions sociales. En outre, nous développons des classificateurs automatiques pour ces catégories, ce qui nous permet d'étudier quelles stratégies de questions sont les plus efficaces dans des contextes émotionnels spécifiques. Cette analyse nous éclaire sur la pertinence de diverses approches pour favoriser l'engagement et la compréhension dans les conversations sociales.

Dans la troisième partie, nous nous appuyons sur nos résultats antérieurs et les critères d'évaluation établis en créant des cadres d'évaluation approfondis pour les chatbots sociaux. Tout d'abord, nous présentons iEval, un cadre d'évaluation humaine spécialement conçu pour estimer les perceptions subjectives des utilisateurs sur leurs partenaires conversationnels au cours d'échanges interactifs. À l'aide de ce cadre, nous comparons quatre chatbots empathiques de pointe et examinons les facteurs de discours qui expliquent les différences entre leurs niveaux de performance. En outre, nous montrons comment notre cadre d'évaluation peut être automatisé en utilisant l'incitation des derniers grands modèles de langage. Cela nous permet de fortement nous approcher des études d'utilisateurs en ligne et d'obtenir une corrélation très élevée avec le jugement humain.

Les nouveaux résultats présentés ici améliorent notre compréhension de l'interaction de l'utilisateur avec les technologies conversationnelles. De plus, les critères et les cadres d'évaluation développés offrent des perspectives et des outils précieux pour façonner et informer la conception de futurs chatbots sociaux.

Mots clés : interaction homme-machine, agents conversationnels, chatbots, intelligence émotionnelle, intelligence sociale, adoption, évaluation, études d'utilisateurs, expériences d'utilisateurs, visualisation

Contents

Acknowledgements **i**

Abstract (English/Français) **v**

List of Figures **xiii**

List of Tables **xvii**

I Introduction **1**

1 Motivation and challenges **3**

2 Contributions and thesis overview **9**
2.1 Establishing users' expectations of chatbots (Part II) 11
2.2 Understanding the role of empathetic questions for attentive curiosity (Part III) 13
2.3 Novel evaluation frameworks for social chatbots (Part IV) 14

II Establishing users' expectations of chatbots **17**

3 Eliciting expectations from user studies **19**
3.1 Introduction 19
3.2 Related work 21
3.3 Research approach 23
3.4 Exploratory qualitative study 24
3.5 Results: Qualitative findings 26
3.6 Model development 32
3.7 Results: Quantitative evaluation 36
3.8 Discussion 44
3.9 Chapter summary 49

4 Eliciting expectations from user online reviews **51**
4.1 Introduction 51
4.2 Related work 52
4.3 Materials and methods 53

4.4 User experience 56
4.5 User expectations 61
4.6 Discussion 64
4.7 Chapter summary 68

III Understanding the role of empathetic questions for attentive curiosity 69

5 Devising and analyzing a taxonomy of empathetic questions in social dialogs 71
5.1 Introduction 71
5.2 Related work 72
5.3 Dataset 73
5.4 Defining empathetic question taxonomy 74
5.5 Crowd-sourced annotation 77
5.6 Automatic labeling 78
5.7 Analysis of questioning strategies 80
5.8 Limitations and future work 83
5.9 Chapter summary 84

IV Novel evaluation frameworks for social chatbots 85

6 Online human evaluation framework for empathetic chatbots 87
6.1 Introduction 87
6.2 Related work 89
6.3 Method: iEval 90
6.4 Experiment 91
6.5 Analysis of results 94
6.6 Discussion 100
6.7 Limitations 101
6.8 Chapter summary 101

7 Automatic evaluation of social chatbots with prompting 103
7.1 Introduction 103
7.2 Related work 105
7.3 Proposed method: DEP 106
7.4 Results 108
7.5 Discussion 113
7.6 Chapter summary 114

V Conclusion 115

8 Conclusion 117
8.1 Summary of contributions 117

8.2 Future directions 120
8.3 Ethical considerations 122

A Appendix **125**
A.1 Semi-structured interview protocol 125
A.2 Examples from empathetic question taxonomy 130
A.3 Details about Mturk task for ED annotation with EQT labels 133
A.4 Details about EQT-labeled data augmentation with lexical similarity 135
A.5 Details about training automatic QBERT classifiers 142
A.6 Extended analysis of questioning strategies 144
A.7 Topic clusters in EmpatheticDialogues 149
A.8 Emotion distribution in grounding scenarios for iEval experiment 150
A.9 Additional details about chatbots' responses in iEval experiment 151
A.10 Prompt format for iEval 153
A.11 Prompt format for FED 156

Bibliography **184**

Curriculum Vitae **185**

List of Figures

1.1 Examples of dialogs with a task-oriented chatbot (left) and an open-domain social chatbot (right). 4

2.1 Outline of the main thesis contributions and the corresponding chapters. . . . 10

3.1 Examples of dialogs with a task-oriented (left) and an open-domain (right) chatbots. Task-oriented chatbot manages to respond appropriately to user's out-of-context personal remark due to incorporated social skills and brings the exchange back to the task flow. Open-domain chatbot maintains social conversation throughout the whole interaction session. 21

3.2 Overview of our research approach. 23

3.3 Structure of the affinity diagram. 26

3.4 A general evaluation framework with hypothesized influence paths. 37

3.5 Structural model. Path significance: $^{***}p < 0.001$, $^{**}p < 0.05$, $^{*}p < 0.1$. 44

4.1 Beta coefficients of the ordinary least squares (OLS) regression of predictive assets and issues on user sentiments and star ratings. 59

4.2 Mean rating of each category characterizing the degree of personification (left) and assigned social role (right) of chatbots. Error bars represent 95% Confidence Intervals. 60

4.3 Distributions of the constructs of the PEACE model mapped to current user experience and user expectations. Error bars represent the bootstrapped 95% Confidence Intervals. 65

5.1 Joint distribution of question intents and acts for 5,272 human-labeled questions. Blue circles are proportional to the frequency of each pair's co-occurrence. . . . 80

5.2 a) Mappings between emotions disclosed by the speakers and listeners' questioning strategies in the first three turns of the ED dialogs. b) and c) Frequency distribution of question acts (b) and intents (c) across turns. (All figures are based on human-labeled ED subset.) 81

5.3 Examples of dialogs grounded in negative (top) and positive (bottom) emotional contexts. Listeners' questions are shown in bold with the assigned *(act, intent)* labels given in parenthesis. The valence of speaker's emotions in each turn is also indicated. 83

6.1 iEval framework. 88
6.2 Benchmarking results of the four chatbots. 94
6.3 Results of ordinal regression on rank. 95% confidence intervals are approximated as two standard errors. All coefficients are significant ($p < 0.05$). 95
6.4 Examples of chat logs with MIME. 96
6.5 Examples of chat logs with MEED and Plain. 98
6.6 Examples of chat logs with Blender. 99

7.1 Offline and online dialog evaluation with the corresponding processes. In the first step, dialog logs are curated. In the second step, each dialog log is assigned a dialog-level score, either by a third-party judge (offline) or by the same conversational partner (online). In the third step, the system ranking is obtained by aggregating the dialog scores of each chatbot. Grey bot icons indicate steps that are intended to be approximated by means of automatic evaluation. Pink boxes mark the steps in the process where the correlation (r.) with the ground truth human judgment is computed to validate the automatic evaluation metric during its development process. 104
7.2 Prompt template to condition a LLM to play an assigned social role while interacting with an evaluated chatbot. 107
7.3 Sankey diagram showing discourse patterns in human-to-bots conversations originating from the iEval dataset. 110
7.4 Sankey diagram showing discourse patterns in freshly generated LLM-to-bots conversations. 110
7.5 Scatter plots depicting the system-level correlation results. Human scores are based on the iEval dialog annotations, while prompted LLM scores are computed based on the generated dialogs. 111

8.1 User-centered design process (adapted from [102]). 117

A.1 Original and resulting dialogs after pre-processing. 133
A.2 The user interface of the Mturk crowd-sourcing annotation task. 134
A.3 Cross-validation results for question acts for the two considered strategies: *Max* in sub-figure A.3a and *Vote* in sub-figure A.3b. 136
A.4 Cross-validation results for question intents for the two considered strategies: *Max* in sub-figure A.4a and *Vote* in sub-figure A.4b. 137
A.5 Train and validation losses over the course of approximately 15 training epochs for question acts (sub-figure A.5a) and question intents (sub-figure A.5b). . . . 142
A.6 Examples of questions labeled automatically with QBERT. Question acts and intents marked with a star* were annotated by Mturk workers. 143
A.7 Mappings between emotions disclosed by the speakers and question acts used by listeners in the first three turns of the ED dialogs (human-labeled ED subset). 145

A.8 Mappings between emotions disclosed by the speakers and question intents used by listeners in the first three turns of the ED dialogs (human-labeled ED subset). 146

A.9 Joint distribution of question intents and acts for 20,201 labeled questions (whole ED dataset). Blue circles are proportional to the frequency of each pair's co-occurrence. 147

A.10 a) Mappings between emotions disclosed by the speakers and listeners' questioning strategies in the first three turns of the ED dialogs (whole ED dataset). b) Frequency distribution of question acts across dialog turns (whole ED dataset). c) Frequency distribution of question intents across dialog turns. Two prevalent intents were excluded for visual clarity; their percentage rates computed for all questions (n=14921 and n=5043) are: *Express interest*: 59.7% → 61.1%, *Express concern*: 24.9% → 19.3% 147

A.11 Mappings between emotions disclosed by the speakers and question acts used by listeners in the first three turns of the ED dialogs (whole ED dataset). 148

A.12 Mappings between emotions disclosed by the speakers and question intents used by listeners in the first three turns of the ED dialogs (whole ED dataset). . 148

A.13 Empirical cumulative distribution function of topic cluster sizes in the train set of EmpatheticDialogues dataset [187]. 149

A.14 Distribution of emotional labels from EmpatheticDialogues dataset in grounding scenarios. The legend shows the mapping between the colors and 14 emotional categories from Plutchik's wheel [179] (8 basic and 6 intermediate emotions). . 150

A.15 Counts of average number of tokens in chatbots' responses over three dialog turns with 95% confidence intervals. 151

A.16 Prompt template for evaluating empathetic listeners in freshly collected dialogs of empathetic models from iEval with the LLM. 153

List of Tables

3.1 Importance of social skills for task-oriented chatbots in different domains. . . . 20
3.2 Expected emotional interaction patterns described recurrently by the participants. 29
3.3 Profile of participants (N=536). 39
3.4 Basic statistics and test results of internal reliability and convergent validity. Constructs with single items are included for completeness. CR and AVE stand for Composite Reliability and Average Variance Extracted, respectively. 40
3.5 Inter-construct correlation matrix. 43

4.1 Description of chatbots used for the study. The second column denotes the star rating of a chatbot at the moment of data collection. The table is split into four pillars corresponding to the rating categories: excellent (top), good, fair, and poor (bottom). 54
4.2 Emerged codes describing assets ($\kappa = 0.61$) and issues ($\kappa = 0.65$) of chatbots' conversational abilities and social skills. 57

5.1 Comparison of question taxonomies. 73
5.2 Statistics of the EmpatheticDialogues dataset. 74
5.3 Accuracy of QBERT classifiers on different slices of test data based on the source of annotations (human, SBERT, or both). 79

6.1 Average number of questions with standard deviation (in the parentheses) asked by different chatbots. 97

7.1 System-level Pearson correlation for four possible prompt design manipulations, with the p-value in brackets. 111
7.2 Results on FED and DSTC9 data. Previous best results are obtained from [251]. Dialog and System indicate dialog- and system-level correlations, respectively, with P standing for Pearson and S for Spearman correlation. All values are statistically significant to $p < 0.05$. 113

A.1 Classification of question acts with corresponding definitions and examples. Under each label its frequency is given for the three corresponding sets: manually labeled, Mturk labeled, and overall. 130

A.2 Classification of question intents with corresponding definitions and examples. Under each label its frequency is given for the three corresponding sets: manually labeled, Mturk labeled, and overall. 131
A.3 Examples of propagated labels obtained using majority vote from the top-3 Nearest-Neighbor (NN) dialogs according to cosine similarity. The first column includes the newly annotated question, and the other three show the top-3 NN dialogs with respective question labels and a similarity value. Spelling and punctuation of the original source have been preserved. 138
A.4 Mapping of 32 emotions and 9 empathetic intents describing the EmpatheticDialogues dataset to three emotion categories of different valence. 144
A.5 Top-15 most frequent tokens for each chatbot in order of decreasing frequency. 151
A.6 Counts of orientation of chatbots' responses (other-, self-, or both) in 50 sampled chat logs (25 for positive and 25 for negative contexts). Prefixes "Pos" and "Neg" stand for positive and negative contexts respectively. 152
A.7 Instructions and demonstration used for prompts for evaluating empathetic listeners in freshly collected dialogs of empathetic models from iEval with the LLM. Demonstrations and their appraisals are manually selected from the iEval dataset. Inputs from "Positive" column were used for dialogs conditioned on positive emotion label and inputs from "Negative" column – for dialogs conditioned on negative emotion label. 154
A.8 Demonstrations used for FED prompts, ranging from Very bad (top) to Very good (bottom). 156

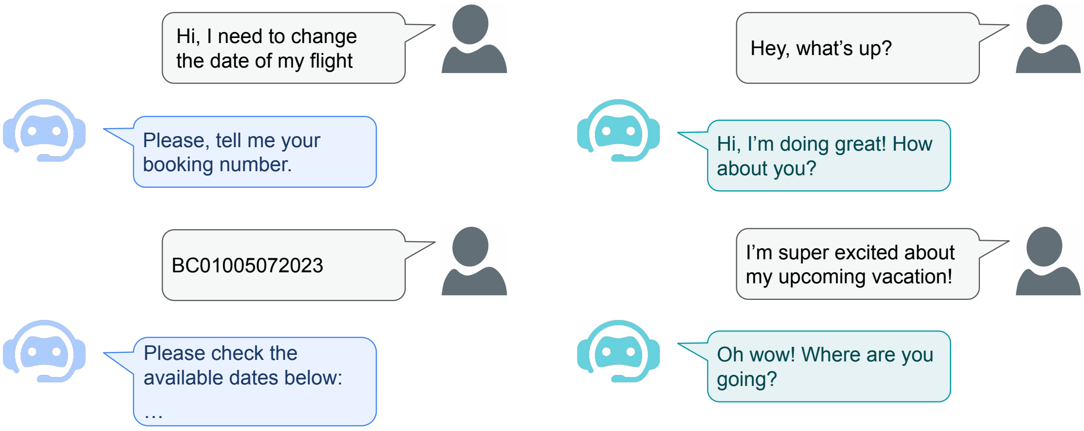
Introduction Part I

1 Motivation and challenges

Language has been a crucial element in the evolution of humanity. While all species possess their unique means of communication, only humans have attained the capability of mastering cognitive language communication [53]. Some scholars attribute humans' rapid evolutionary advancement to the development and refinement of language [52]. This advancement facilitated human-to-human interaction, collaboration, strategic planning, long-distance trade, and many other endeavors for humankind. As we keep evolving and acquiring more sophisticated tools and technologies, practitioners envision a future where machines can communicate with users via natural language, thereby paving the way for a significant breakthrough in human-machine interaction.

Towards this goal, the first conversational computer systems emerged in the second half of the twentieth century. In 1966, Joseph Weizenbaum from MIT created the first publicly known computer program of this kind, called Eliza [238]. Eliza used a set of hand-crafted rules to simulate a Rogerian psychotherapist, following a non-directive form of talk therapy [196]. Many people believed they were conversing with a real person during their sessions with the program. Despite this success, Eliza was very limited in its abilities and could only operate within a restricted domain. After Eliza's creation, similar conversational systems were developed, such as Parry in 1972, designed to mimic the behavior of a paranoid person [40], and Alice in 1995, a more advanced and adjustable system that was still based on pattern-matching techniques [233].

Some further attempts to create response-generation models using approaches borrowed from statistical machine translation were taken in the early 2010s [193]. It was not before the introduction of sequence-to-sequence models for natural language generation in 2014 [215, 230] that the development of language-driven systems gained significant momentum. Today, we observe a growing range of systems that can naturally engage with their users, be it through written or spoken communication. These entities are often referred to by many



Hi, I need to change the date of my flight

Please, tell me your booking number.

BC01005072023

Please check the available dates below:
…

Hey, what's up?

Hi, I'm doing great! How about you?

I'm super excited about my upcoming vacation!

Oh wow! Where are you going?

Task-oriented dialog

Open-domain dialog

Figure 1.1: Examples of dialogs with a task-oriented chatbot (left) and an open-domain social chatbot (right).

names, including conversational agents, chatbots, dialog systems, intelligent assistants, and more.

In general, all these systems can be classified into two categories [67]: *task-oriented* and *open-domain* (Figure 1.1). Task-oriented systems are designed to help users achieve specific goals. They are commonly used by businesses to automate customer service processes or take the form of intelligent personal assistants, such as Apple's Siri or Google Assistant. On the other hand, open-domain conversational agents are meant to engage in unstructured conversations with their users, allowing users to chat with them on any topic of interest. Previously, scholars used to employ the terms "open-domain chatbot" and "social chatbot" interchangeably, implying agents that can maintain engaging and natural chitchat, not requiring external dependencies for expert knowledge [67, 86]. Recently, with the release of powerful generative artificial intelligence (AI) tools, such as Bing [153] conversational search engine or ChatGPT[1] and Bard[2] chatbot assistants that can handle a wide spectrum of tasks, the term "open-domain chatbot" has acquired a more expansive meaning, encompassing these systems as well. These systems are capable of producing highly natural conversational outputs, but social requirements for them are presumably lower than for social chatbots in a traditional sense. In this thesis, we primarily focus on social chatbots, whose objectives include facilitating smooth interaction with the users and gaining their long-term engagement, confidence, and trust [207]. Such chatbots mainly find applications as virtual companions to their users and might require higher levels of emotional intelligence [207]. Although the widespread adoption of these systems is still ongoing [81], with continued innovation in natural language processing (NLP),

---

[1] https://chat.openai.com/chat
[2] https://bard.google.com/

we can anticipate a further rise of such agents as well as the emergence of hybrid systems that can seamlessly transition between task-oriented conversations and casual chitchat.

To summarize, we are now on the verge of a great leap in human-machine interaction, enabled by rapidly emerging technologies that equip machines with advanced natural language capabilities. This new modality of interaction currently lacks well-defined design principles and guidelines. Moreover, the increased naturalness of technological interfaces may pose unprecedented risks and concerns for users. To ensure the social acceptance of these systems, it is crucial for technology creators to elicit users' expectations and preferences for such interaction experiences and implement the corresponding control mechanisms. In this light, developing clear evaluation criteria is necessary for assessing outcomes and deriving implications for chatbots' naturalness improvement that align with users' contexts and requirements. However, evaluating naturalness-related aspects of conversational agents, such as their social and emotional skills, is a complex problem due to several interrelated challenges. These challenges include **ill-defined evaluation objectives**, **a lack of established evaluation frameworks**, and **resource-related considerations**. We consider each of them in the following paragraphs.

First, the challenge of the **ill-defined evaluation objectives** stems from the open-ended nature of conversational interactions. On the one hand, it leads to inefficacy of existent automatic metrics from other domains for evaluating an agent's ability to engage in a dialog. Metrics from machine translation and text summarization fields, which rely on word-overlap statistics, were shown to be unable to account for a vast range of possible appropriate responses for the same input context in a dialog [136]. On the other hand, identifying pertinent evaluation criteria for dialog generation is difficult. Practitioners have turned to evaluating the social aspect of chatbot-user interactions only recently [207] and focused extensively on studying the question of what makes a good conversation [205, 37]. This has resulted in the proposal of numerous evaluation dimensions, but the consistent uniform terminology is still missing [59], leading to the second challenge.

Second, **lack of established evaluation frameworks** is related to the social aspect of chatbot-user interactions. Defining metrics for transactional operations that have clear outcomes, such as the chatbot's goal completion rate, is simpler in comparison. Meanwhile, creating computational approaches for social and affective norms is more challenging because these are subjective in nature [111]. Addressing these areas requires multidisciplinary research involving psychology, cognitive science, and social science [98, 70]. The field of affective computing has made considerable progress since Picard initiated it in the mid-1990s [176], especially in the development of models for generating emotional responses [*e.g.* 263, 265, 95]. Despite this progress, there has been lack of efforts to facilitate the *conceptualization* and *operationalization* of these results for evaluation purposes.

Third, the final challenge pertains to balancing the need for accurate evaluation outcomes with the availability of **necessary resources**. Since there is a lack of automated evaluation metrics for open-domain dialog, the research community relies extensively on human evaluation [149]. However, creating a reliable human evaluation task requires considerable effort on its own, not to mention the associated time and cost expenses. As a result, current endeavors are aimed at reducing the *human load* in evaluation tasks while maintaining the validity of the results [*e.g.* 126, 49] and developing innovative, human-free evaluation methods [251]. These methods often require *language resources*, such as knowledge bases, annotation schemes (taxonomies), high-quality datasets, and computational tools, to develop and benchmark new metrics. While practitioners are increasingly involved in creating these materials for various natural language tasks [110], resources for evaluating social dialog are still scarce and would benefit from expansion.

Despite the challenges outlined above, there are emerging opportunities that could help advance the field of open-domain dialog evaluation. **The core premise of this thesis is that following user-centered research principles, drawing on interdisciplinary insights, and leveraging cutting-edge language modeling tools can lead to creation of novel and meaningful evaluation frameworks for social chatbots.** This premise is underpinned by several considerations.

First, **user-centered research and design** focus on understanding the needs and perspectives of end-users to create products that are intuitive, usable, and satisfying to use [162]. By conducting research with users, practitioners gain valuable insights into their behaviors, goals, and pain points, which helps identify design and **evaluation objectives** that are relevant for the end users' experience and contexts. User-centered research has a set of established methods and principles [41, 122], and is always open to embrace new approaches and perspectives, enabled by the rapidly changing technological and informational landscape. Traditional approaches, such as focus groups, interviews, and quantitative modeling, have found applications for a wide range of research areas, from developing models for evaluating recommendation systems [184, 183], to understanding users' privacy concerns regarding the use of shared smart speakers [99]. Furthermore, with digitalization and spread of IT services, we have an abundance of digital traces and artifacts that users leave on the web. They offer new, exciting opportunities for studying user experiences and expectations "from the distance", which could serve as a great way to triangulate the results obtained with more "traditional" methods.

Second, as artificial intelligence rapidly advances and finds increasing use across various social applications, there is a rising demand for **interdisciplinary research** to enhance the effectiveness of AI-driven systems [106, 93, 51]. As demonstrated by previous studies [*e.g.* 46, 202], incorporating insights from social sciences can be instrumental in this pursuit. Interdisciplinary perspectives can provide a deeper *conceptual understanding* of various social phenomena that users experience during interactive sessions with the systems, thereby facilitating the development of more comprehensive and informed **evaluation frameworks**. These conceptual insights can be further *operationalized* using modern computational methods, leading to their greater scalability and applicability.

Third, the recent advancements in **language modeling tools and techniques** have enabled machines to understand, interpret, and generate human language with increasing accuracy and fluency. The introduction of transformer architecture, for instance, has significantly improved the machine performance on natural language tasks by allowing for better representation of context and meaning [228]. Among others, researchers have leveraged pre-trained text representations from transformer models to identify and generate paraphrases [161, 232], annotate datasets [105, 6, 239], and construct knowledge bases from text [16], thus enhancing the **language resources** available for future analysis. Moreover, these evolving technologies have also led to a reduction in the **human resources** required for certain tasks. The spread of automatic machine translation and text summarization services offer a long-standing evidence for this point. Now, similar techniques are becoming adopted for other tasks and spheres, for example, using transformer models to automate data entry and disease prediction based on medical records [54, 188] or moderate content on social media [117]. As these tools keep enhancing, relying on novel training setups such as massive pre-training, zero-shot learning [24], and learning from human feedback [166], they hold great potential for extending their benefits to other areas, including automatic dialog evaluation.

The combination of the mentioned factors sets the stage for the contributions of this thesis. In the next chapter, we summarize the main contributions and outline the structure of the thesis.

2 Contributions and thesis overview

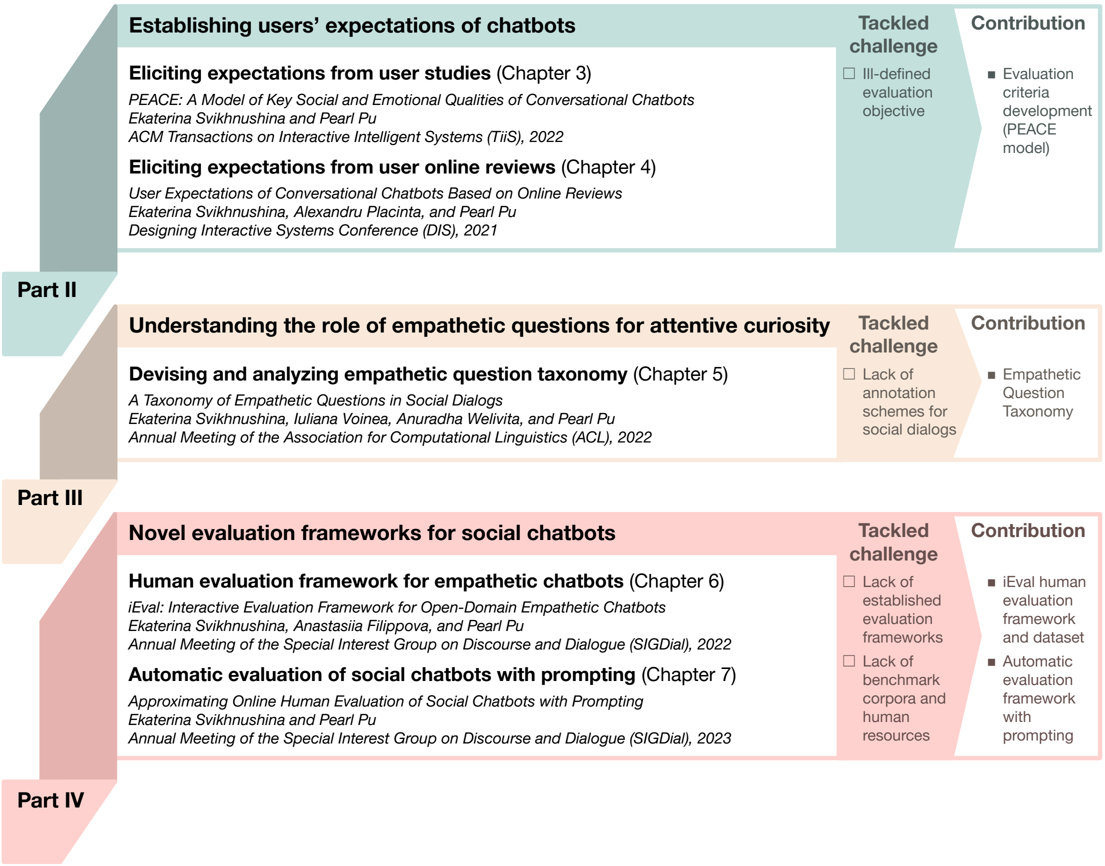
In our pursuit of establishing novel, effective evaluation methods for social chatbots, we adopt human-centered research principles to identify what users seek in a compelling and safe conversational agent. We then delve deeper into the areas of interest, incorporating interdisciplinary insights into the dialog discourse structure. Through this approach, we demonstrate how this enhanced comprehension of user expectations can be utilized to craft evaluation frameworks that are relevant and valuable, and how they can be implemented both with and without human involvement.

The thesis is structured into five parts. Part I sets the stage for the thesis and includes an overview of the following chapters. The subsequent parts, Parts II, III, and IV, present the main contributions of the thesis. The final Part V concludes the thesis.

Specifically, in Part II we present studies establishing users' expectations of conversational chatbots and their relation to current experiences. Studying these aspects has broad implications for technology developers and researchers, as it advances comprehension of users' preferences and pain points, helping prioritize areas for further analysis, improvement, and evaluation. Chapter 3 introduces two consecutive user studies. In the first one, we conduct semi-structured interviews with the users for exploratory purposes, while in the second study, we utilize structured quantitative methods to evaluate the influence of expected chatbots' social skills on adoption. Chapter 4 addresses a similar research question with an alternative set of methods. We use the mixed-method approach to derive insights from online reviews written by chatbot users. Relying on alternative methods allows us to emphasize the initial findings.

In Part III, we examine a critical but understudied aspect expected from social chatbots – attentive curiosity. Chapter 5 explores the essential role that empathetic questioning plays in creating engaging conversations, drawing on related interdisciplinary literature. We devise a taxonomy of tags to distinguish various functions that questions can have in dialogs and demonstrate how to operationalize it using crowdsourcing and computational techniques. Our findings yield valuable insights into which question-asking strategies are most suitable for specific emotional contexts.

Finally, in Part IV, we build on our previous findings to develop comprehensive evaluation frameworks for social chatbots. Chapter 6 emphasizes the importance of capturing users' subjective perceptions of their conversational partners during interactive exchanges. We propose a human evaluation framework to benchmark four state-of-the-art empathetic chatbots and analyze discourse factors that explain their varying degrees of performance. Further, in Chapter 7, we advance our research by demonstrating how we can automate our evaluation framework using the latest large language models. We contribute by outlining this automatic evaluation procedure, which uses chatbots' other-play and prompting to approximate live user studies, and illustrate its effectiveness and versatility across several corpora.



Establishing users' expectations of chatbots

Eliciting expectations from user studies (Chapter 3)

PEACE: A Model of Key Social and Emotional Qualities of Conversational Chatbots
Ekaterina Svikhnushina and Pearl Pu
ACM Transactions on Interactive Intelligent Systems (TiiS), 2022

Eliciting expectations from user online reviews (Chapter 4)

User Expectations of Conversational Chatbots Based on Online Reviews
Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu
Designing Interactive Systems Conference (DIS), 2021

Tackled challenge

- ☐ Ill-defined evaluation objective

Contribution

- ■ Evaluation criteria development (PEACE model)

Part II

Understanding the role of empathetic questions for attentive curiosity

Devising and analyzing empathetic question taxonomy (Chapter 5)

A Taxonomy of Empathetic Questions in Social Dialogs
Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu
Annual Meeting of the Association for Computational Linguistics (ACL), 2022

Tackled challenge

- ☐ Lack of annotation schemes for social dialogs

Contribution

- ■ Empathetic Question Taxonomy

Part III

Novel evaluation frameworks for social chatbots

Human evaluation framework for empathetic chatbots (Chapter 6)

iEval: Interactive Evaluation Framework for Open-Domain Empathetic Chatbots
Ekaterina Svikhnushina, Anastasiia Filippova, and Pearl Pu
Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial), 2022

Automatic evaluation of social chatbots with prompting (Chapter 7)

Approximating Online Human Evaluation of Social Chatbots with Prompting
Ekaterina Svikhnushina and Pearl Pu
Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial), 2023

Tackled challenge

- ☐ Lack of established evaluation frameworks
- ☐ Lack of benchmark corpora and human resources

Contribution

- ■ iEval human evaluation framework and dataset
- ■ Automatic evaluation framework with prompting

Part IV

Figure 2.1: Outline of the main thesis contributions and the corresponding chapters.

The organization of the thesis is shown in Figure 2.1, which depicts each chapter's contribution based on the original publication. The remainder of the thesis is structured as follows. Chapters 3–7 describe the contributions of the thesis. A discussion and conclusion are provided in

Chapter 8. We summarize each chapter's main scientific contributions and novelties below.

2.1 Establishing users' expectations of chatbots (Part II)

Eliciting expectations from user studies (Chapter 3)

Adapted from [219]:

PEACE: A Model of Key Social and Emotional Qualities of Conversational Chatbots. Ekaterina Svikhnushina and Pearl Pu. ACM Transactions on Interactive Intelligent Systems (TiiS), 2022.

Open-domain chatbots engage with users in natural conversations to socialize and establish bonds. However, designing and developing an effective open-domain chatbot is challenging. It is unclear what qualities of a chatbot most correspond to users' expectations and preferences. Even though existing work has considered a wide range of aspects, some key components are still missing. For example, the role of chatbots' ability to communicate with humans at the emotional level remains an open subject of study. Furthermore, these trait qualities are likely to cover several dimensions. It is crucial to understand how the different qualities relate and interact with each other and what the core aspects would be.

For this purpose, in Chapter 3, we describe how we first designed an exploratory user study aimed at gaining a basic understanding of the desired qualities of chatbots through semi-structured interviews with a special focus on their emotional intelligence. Using the findings from the first study, we constructed a model of the desired traits by carefully selecting a set of features. With the help of a large-scale survey and structural equation modeling, we further validated the model using data collected from the survey. We contribute the following scientific findings:

1. We establish and analyze the PEACE model (Politeness, Entertainment, Attentive Curiosity, and Empathy), which sheds light on the importance and interplay between the chatbots' qualities and the effect of users' attitudes and concerns on their expectations of the technology.
2. We derive a set of design implications useful for the development of socially adequate and emotionally aware open-domain chatbots. According to the implications, users' attitudes strongly influence their expectations of chatbots. Additionally, *politeness* is an essential prerequisite for people to consider engagement with socially intelligent conversational agents. *Entertainment and humor* form a common "entry point" to chatbot technology, while *attentive curiosity* and *empathy* also have a significant influence on adoption intention, given that the politeness requirement is satisfied.

The results of this study make the first step towards establishing a holistic view of how various social skills should integrate into an open-domain chatbot. Moreover, they create an essential foundation for further exploration of the identified social and emotional benchmark criteria.

Eliciting expectations from user online reviews (Chapter 4)

Adapted from [217]:

User Expectations of Conversational Chatbots Based on Online Reviews. Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. Designing Interactive Systems Conference (DIS), 2021.

In the following chapter, we utilize alternative methods to reinforce and corroborate the results of the previous user study.

As open-domain chatbots have received increasing attention recently, new opportunities for studying user interaction with them are emerging. In Chapter 4, we explore user experience and expectations of these agents in a mixed-method study by drawing from reviews of chatbots posted on Google Play.[1] Results of statistical analysis reveal which social qualities of chatbots are the most significant for user satisfaction. Further, we employ natural language processing and qualitative methods to identify how users wish their chatbots to evolve in the future. The main contributions of our analysis are the following:

1. We identified the most influential factors shaping the current user experience with chatbots. Users primarily value the entertaining component of their experience, while privacy concerns and chatbots' use of profanity hinder widespread adoption.
2. We assessed the gap between users' current experience and their expectations of chatbots and derived implications for chatbots' future improvement. Users' expectations call for more human-like behavior of chatbots. The most prominent expectations include chatbots' abilities to treat and express emotions and be more attentive to the user. In addition, users anticipate that chatbots will be able to personalize their interactions and maintain a consistent personality.

Beyond the main findings, this contribution demonstrates the utility of web content analysis for deriving user insights. It also lends external validity to our initial study and, when combined with the PEACE model, exposes the discrepancy between users' current experiences and their expectations.

---

[1] https://play.google.com/

2.2 Understanding the role of empathetic questions for attentive curiosity (Part III)

Devising and analyzing empathetic question taxonomy (Chapter III)

Adapted from [221]:

A Taxonomy of Empathetic Questions in Social Dialogs. Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. Annual Meeting of the Association for Computational Linguistics (ACL), 2022.

Informed about key expected features of conversational chatbots, in Chapter III we deep dive into the analysis of the main discourse mechanism for demonstrating attentive curiosity in empathetic dialogs.

Effective question-asking is a crucial component of a successful conversational chatbot. It could help the bots manifest empathy and render the interaction more engaging by demonstrating attention to the speaker's emotions. However, current dialog generation approaches do not model this subtle emotion regulation technique due to the lack of a taxonomy of questions and their purpose in social chitchat. To address this gap, we have developed an empathetic question taxonomy (EQT), with special attention paid to questions' ability to capture communicative acts and their emotion-regulation intents. We further design a crowdsourcing task to annotate a large subset of the EmpatheticDialogues [187] dataset with the established labels. We use the crowd-annotated data to develop automatic labeling tools and produce labels for the whole dataset. Finally, we employ information visualization techniques to summarize co-occurrences of question acts and intents and their role in regulating the interlocutor's emotion. Overall, we contribute the following language resources and conceptual findings:

1. A novel empathetic question taxonomy (EQT) comprising 9 labels for question communicative acts and 12 labels for their emotion-regulation intents.

2. Trained models for classifying labels from our taxonomy and annotations for all listener questions in the EmpatheticDialogues dataset.

3. Visualizations demonstrating that empathetic listeners adopt different questioning strategies depending on the polarity of the emotional context.

These results reveal important question-asking strategies in social dialogs. The EQT classification scheme can facilitate the computational analysis of questions in datasets. More importantly, it can inform future efforts in the generation and evaluation of empathetic questions.

2.3 Novel evaluation frameworks for social chatbots (Part IV)

Human evaluation framework for empathetic chatbots (Chapter 6)

Adapted from [216]:

iEval: Interactive Evaluation Framework for Open-Domain Empathetic Chatbots. Ekaterina Svikhnushina, Anastasiia Filippova, and Pearl Pu. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial), 2022.

In Chapter 6, we leverage our previous research findings to address the challenging problem of evaluation of empathetic chatbots. By empathy, we mean the ability to understand and relate to the speakers' emotions, and respond to them appropriately. Human evaluation has been considered as the current standard for measuring the performance of open-domain empathetic chatbots. However, existing evaluation procedures suffer from a number of limitations we try to address in Chapter 6. We describe iEval, a novel interactive evaluation framework where the person chatting with the bots also rates them on different conversational aspects, as well as ranking them, resulting in greater consistency of the scores. We use iEval to benchmark several state-of-the-art empathetic chatbots, allowing us to discover some intricate details in their performance in different emotional contexts. Our main contributions are the following:

1. We propose iEval framework – an interactive human evaluation framework for capturing users' subjective perceptions of chatbots, whose design is informed by interdisciplinary insights from psychology and cognitive sciences.

2. We demonstrate that users have different expectations from empathetic listeners depending on the emotional polarity of the chat, as the results of our statistical analysis reveal statistically significant difference in the appraisals of the same chatbots in polar opposite emotional contexts. We provide discussion about factors that likely explain this disparity based on the analysis of the collected chat logs.

3. We release the chat log with human annotations to streamline the construction and calibration of future evaluation metrics.

In addition to proposing a novel interactive evaluation framework, this study highlighted the unequal performance of existing chatbots in generating empathetic responses for positive and negative conversational scenarios. These results imply that practitioners should be cautious while implementing chatbots for emotional support or sensitive conversations, as their effectiveness might be limited in certain scenarios. The findings suggest the need for further research and development to improve the empathetic capabilities of chatbots, particularly in handling negative emotions.

Automatic evaluation of social chatbots with prompting (Chapter 7)

Adapted from [218]:

Approximating Online Human Evaluation of Social Chatbots with Prompting. Ekaterina Svikhnushina and Pearl Pu. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial), 2023.

Lastly, in Chapter 7, we bring our iEval framework to the next level, demonstrating how live user studies can be approximated in an automatic manner.

As conversational models become increasingly available to the general public, users are engaging with this technology in social interactions. Such unprecedented interaction experiences may pose considerable social and psychological risks to the users unless the technology is properly controlled. This highlights the need for scalable and robust evaluation metrics for conversational chatbots. Existing evaluation metrics aim to automate offline user evaluation and approximate human judgment of pre-curated dialogs. However, they are limited in their ability to capture subjective perceptions of users who actually interact with the bots and might not generalize to real-world settings. To address this limitation, we propose an approach to automate online human evaluation leveraging large language models (LLMs) from the GPT-family. We introduce a new Dialog system Evaluation framework based on Prompting (DEP), which enables a fully automatic evaluation pipeline that replicates live user studies and achieves an impressive correlation with human judgment (up to Pearson $r = 0.95$ on a system level). The DEP approach involves collecting synthetic chat logs of evaluated bots with an LLM in the other-play setting, where the LLM is carefully conditioned to follow a specific scenario. We further explore different prompting approaches to produce evaluation scores with the same LLM. The best-performing prompts, which contain few-shot demonstrations and instructions, show outstanding performance on the tested dataset and demonstrate the ability to generalize to other dialog corpora. Overall, we contribute the following results:

1. We propose a new automatic evaluation framework for social chatbots using other-play and prompting of large language models. We demonstrate the effectiveness of the framework both for mimicking human discourse in an assigned social role to produce synthesized chat logs and for generating quality annotations for dialogs.
2. We release emotion and quality annotations collected during the analysis for all benchmark corpora used for the study.

We demonstrated the outstanding performance of the proposed DEP framework on several chitchat datasets. Additionally, due to the flexibility of LLM prompting, this framework can be easily adapted to other domains by adjusting and calibrating the prompt instructions.

This thesis will now move on to the main chapters, where each contribution will be discussed in greater detail.

Part II

Establishing users' expectations of chatbots

3 Eliciting expectations from user studies[*]

3.1 Introduction

Recent advances in natural language processing, especially response generation, have helped the human-computer interaction field reach one of its long-standing dreams. Interfaces in the form of chatbots have become increasingly popular for users to interact with many services and applications [61]. In Chapter 1 we introduced two types of chatbots, which are also referred to as conversational agents [130]: task-oriented chatbots, which help users reach a specific goal, and open-domain chatbots, which can engage in social chitchat without focusing on any specific subject [67].

With the fast spread of chatbots, numerous studies on user experience with them emerged. As most of the existing chatbots are task-oriented, many research works focused on the functional properties and pragmatic considerations of such agents, for example, their abilities to make their operational features apparent to the users [141, 103], consistency in input modalities [256, 43], and transparency in privacy policies [43, 99].

More recently, researchers began examining user expectations about the social qualities of task-oriented bots. Initially, it was assumed that humans expect the chatbots to be reticent when it comes to emotions [36]. Further studies of such agents in different domains demonstrated that users try to initiate small talks with them despite their primarily utilitarian purpose (Table 3.1). Such interactions occur in abundance because the human-like behavior of a natural language interface likely triggers social expectations from the users [130]. This phenomenon is especially pronounced for users with a greater disposition to chatbot's personification and social skills [185, 129]. Drawing on these observations, many researchers maintain that future task-oriented chatbots should possess the same social and emotional qualities as the open-

---

[*] Adapted from [219]: PEACE: A Model of Key Social and Emotional Qualities of Conversational Chatbots. Ekaterina Svikhnushina and Pearl Pu. ACM Transactions on Interactive Intelligent Systems (TiiS), 2022.

domain ones [130, 244, 254], especially when they switch context between task-driven dialogs and small talks. We believe a task-oriented bot can enhance its efficiency by becoming more sociable and likable (cf. Figure 3.1).

A series of qualitative studies specifically analyzed chatbots' desired social properties, for example, active listening skills [244] or personality traits [148]. However, each of them is devoted to a specific characteristic. A holistic view of how various social skills should integrate is lacking. Moreover, some crucial aspects, such as emotional awareness of chatbots, have not been vastly examined in previous work. In this perspective, the need arises for a consolidated understanding of what qualities are most crucial for chatbots to ensure enhanced user experience, leading to technology adoption.

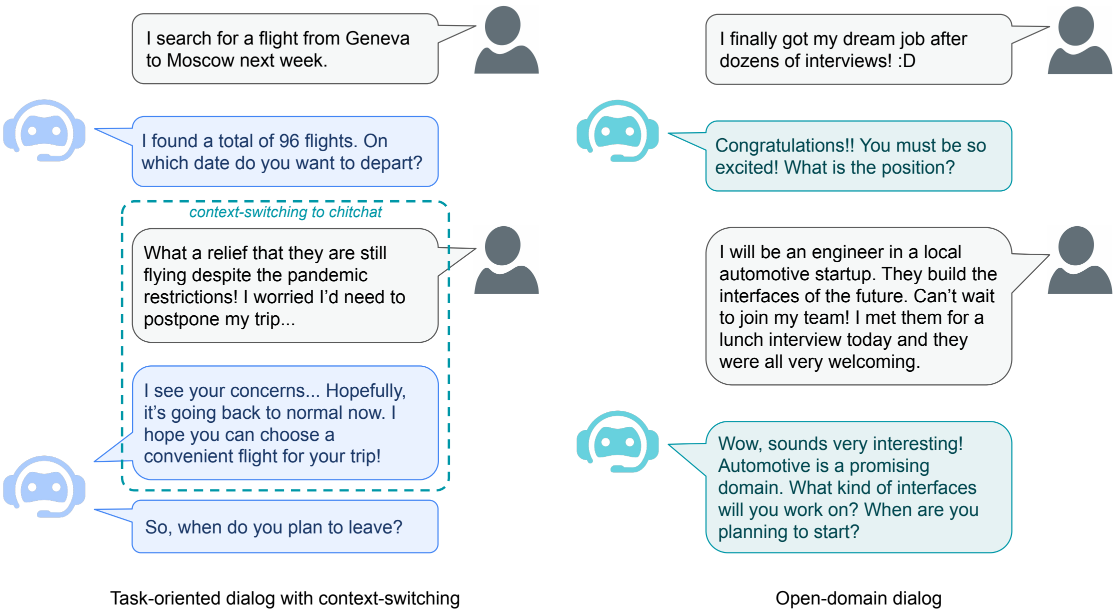
The purpose of the present work was to develop a unified model of social and emotional qualities for a conversational chatbot, which can engage in chitchat with its users, either as part of the task-oriented dialogs or for purely open-domain conversations (cf. Figure 3.1). Particularly, we investigate the following questions:

1. *With which social and emotional norms do users desire their chatbots to comply?*
2. *How do chatbots' social and emotional skills relate to users' intention to engage with such agents?*

To address these questions, we designed two interconnected user-centered studies. We began our research by conducting exploratory qualitative interviews with the users to form a basic comprehension of their expectations and concerns. In a second step, we used these findings as a roadmap for constructing a refined model of desirable qualities of chatbots. We then conducted a large-scale quantitative experiment to validate the model using psychometric techniques [164]. We show 1) how to elicit key user expectations and concerns about socially and emotionally aware open-domain chatbots in semi-structured interviews, 2) how to select the most relevant criteria to derive a model of trait qualities, 3) how to validate the model's fitness and how each of the established constructs influence user adoption intention. Finally, we discuss the implications of our work by identifying several design guidelines for conversational agents. Since the market of open-domain conversational chatbots is still in its infancy and

Table 3.1: Importance of social skills for task-oriented chatbots in different domains.

Domain	Percentage of social user input	Reference
Online Shopping	80%	[248]
Customer Service	40%	[247]
HR Domain	36%	[130]
Education	47%	[175]
Healthcare	80%	[45]



I search for a flight from Geneva to Moscow next week.

I found a total of 96 flights. On which date do you want to depart?

What a relief that they are still flying despite the pandemic restrictions! I worried I'd need to postpone my trip...

I see your concerns... Hopefully, it's going back to normal now. I hope you can choose a convenient flight for your trip!

So, when do you plan to leave?

I finally got my dream job after dozens of interviews! :D

Congratulations!! You must be so excited! What is the position?

I will be an engineer in a local automotive startup. They build the interfaces of the future. Can't wait to join my team! I met them for a lunch interview today and they were all very welcoming.

Wow, sounds very interesting! Automotive is a promising domain. What kind of interfaces will you work on? When are you planning to start?

context-switching to chitchat
Task-oriented dialog with context-switching
Open-domain dialog

Figure 3.1: Examples of dialogs with a task-oriented (left) and an open-domain (right) chatbots. Task-oriented chatbot manages to respond appropriately to user's out-of-context personal remark due to incorporated social skills and brings the exchange back to the task flow. Open-domain chatbot maintains social conversation throughout the whole interaction session.

the existent task-oriented agents typically fail to engage in proper chitchat, we hope that the social and emotional benchmark criteria presented here will inform the efforts of designers and developers to offer more fulfilling user experience in the future.

3.2 Related work

3.2.1 Qualitative and quantitative user experience studies

Only few studies of user interaction with open-domain chatbots exist due to their scarcity on the market [81]. Authors in [158] conducted a diary study to explore how human-like qualities of the Replika chatbot influence user engagement.[1] The results demonstrated that users desire a chatbot to follow largely the social protocol of newly acquainted people before developing it further into a protocol among people who are familiar with each other. Similar findings were reported in [148]. In this work, Thies et al. [148] ran a Wizard-of-Oz (WoZ) experiment to determine what personality traits of a chatbot would be preferred by its target audience. Their participants favored a proactive agent with a developed sense of humor that would gradually befriend the user. Kim et al. [112] conducted a workshop with teenagers, followed by in-depth interviews, to understand how some of the conversational agents' built-in qualities, such as being tireless and separated from the human world, can help them deal with anxiety

---

[1] https://replika.ai/

and stress. According to their findings, teenagers expected such agents to be good listeners, keep their secrets, and deliver effective supportive messages. Finally, several works focused on social components of task-oriented chatbots. Xiao et al. [244] compared the effectiveness of an interview chatbot with and without active listening skills. They concluded that the former managed to deliver a more engaging user experience and elicit higher-quality user responses. Katayama et al. [108] conducted a mixed-method study to establish user expectations about appropriate interaction patterns for an emotionally aware question-answer agent in a wearable device. Their study implied that in general, users prefer the agent to adapt its emotionality based on the context. However, the scope of their study was focused on the agent's speech prosody, such as pitch, rate, glottal tension, and did not consider natural language generation.

These studies discussed a number of social skills that users expect from their chatbots. Even though some works emphasized the importance of integrating emotional intelligence [256, 148], as presented above, research focusing precisely on user expectations of emotional skills of chatbots is limited. Furthermore, each of the discussed studies considered only a subset of chatbots' social qualities at a time and did not relate user expectations to the likelihood of adoption.

3.2.2 Psychometric studies of chatbot adoption

In the area of information system adoption, usability-oriented research has focused on understanding and explaining what aspects of computer software, including both its interface's ease of use and its offered functional features, lead to user acceptance. A number of evaluation frameworks to study such technology adoption have been established. Davis proposed the Technology Acceptance Model (TAM) [47], one of the first models that delineates users' intention to use information technology based on two determining variables: perceived usefulness and perceived ease of use. Following his work, more elaborate models of adoption emerged. Venkatesh et al. [229] formulated the Unified Theory of Acceptance and Use of Technology consisting of four predictive constructs: performance expectancy, effort expectancy, social influence, and facilitating conditions. Kulviwat et al. [115] introduced the Consumer Acceptance of Technology (CAT) model by merging TAM with three additional affective perceptions of a system: pleasure, arousal, and dominance. All adoption studies offer psychometric measurement scales that include multiple question items to assess each of the constructs in the model [164]. The scales are evaluated using structural equation modeling (SEM) techniques to assess their validity and reliability and study the influences between the constructs.

Concerning the works on chatbot adoption, by far several authors examined user behavioral intentions with respect to various task-oriented chatbots. Zarouali et al. [257] applied the CAT model to a movie ticket booking chatbot. Gonzalez et al. [151] evaluated a chatbot to organize vacation trips with an extended version of the UTAUT model. Sheehan et al. [206] investigated the mediating role of human-like qualities in a customer service agent between its ability to resolve miscommunication and user adoption. All these studies suggested evidence that the chatbot's abilities to interact naturally with the users, for example by being entertaining and resolving miscommunication, impact user intention to use the chatbot positively.

However, each of the works above studies a dedicated domain-specific chatbot and considers its human-like qualities only partially. More importantly, to the best of our knowledge, no adoption studies of open-domain conversational agents exist. To facilitate the developmental efforts and bring conceptual clarity, our study aims to establish a unified model of social and emotional qualities for such chatbots and assess their impact on user intention to accept this technology. Moreover, in contrast to previous work, we intend to further extend the impact of our findings by translating them to specific design guidelines.

3.3 Research approach

Given the novelty of the research subject, we follow the traditional iterative approach in human-computer interaction research [122]. As the first step, we conduct a sequence of exploratory semi-structured interviews to learn about user expectations and concerns regarding conversational agents. After gaining this preliminary understanding based on *qualitative* findings, we utilize more structured methods to define the constructs of chatbots' social skills and *quantitatively* evaluate their influence on adoption. Figure 3.2 shows an overview of our research process.

The goal of the first exploratory step is to determine the main social qualities that users wish future chatbots to possess. We frame this study with the focus on emotionally aware agents as the emotional capabilities of chatbots lacked attention in previous work, but still keep our discussions with the participants quite broad to consider a large spectrum of other desirable skills. Sections 3.4 and 3.5 provide a comprehensive discussion of the study design and resulting findings.

For the second step, we revisit the qualitative findings to extract the social and emotional characteristics most predictive of adoption. We couple this analysis with an extensive literature review to produce a solid model of desired chatbots' qualities. Subsequently, we run a psychometric survey to evaluate and validate the model and establish the influence of different

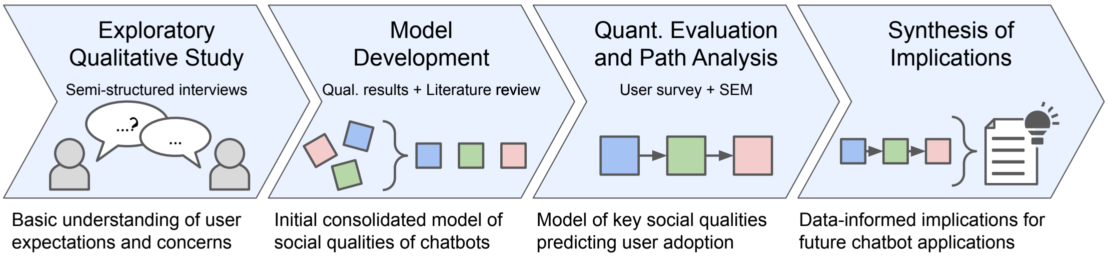

Exploratory Qualitative Study
Semi-structured interviews
...?
...
Basic understanding of user expectations and concerns
Model Development
Qual. results + Literature review
Initial consolidated model of social qualities of chatbots
Quant. Evaluation and Path Analysis
User survey + SEM
Model of key social qualities predicting user adoption
Synthesis of Implications
Data-informed implications for future chatbot applications

Figure 3.2: Overview of our research approach.

constructs on users' behavioral intentions. The details about the model development and evaluation are included in Sections 3.6 and 3.7.

3.4 Exploratory qualitative study

This section expands on how we executed the first step of our research approach in detail. We describe the study design, participants, and provide the data analysis procedure.

3.4.1 Study design

We employed semi-structured interviews as a data collection technique for exploring user expectations and concerns about the social qualities of emotionally aware chatbots. A detailed interview guide with open-ended questions was prepared to prompt the discussion with the participants and is available in Appendix A.1. After developing the guide, we recruited participants through the snowball sampling [14]. In the invitation email, we provided a brief description of our research and interview procedure and informed the recipients about the incentives. Participants were recruited until saturation had occurred [82]. We focused on participants with almost no prior exposure to open-domain chatbots. This is to avoid any biases and fixations that may result from previous experience. In this setting, users are likely to reveal their true prospect of the future technology as they would expect it to be. To make sure the participants could relate to the subject of discussion, we verified that they used smartphones and computers on a regular basis and were familiar with messaging applications. Each participant was offered a small gift as a token of appreciation right after the interview, and two of them received smart speakers after a draw among all participants. All interviewees provided their consent for their data to be reported anonymously.

3.4.2 Participants

In total, 18 fluent or native English speakers (10 female, 8 male) from various backgrounds took part in our study. Over a half of the participants belonged to teenage (10–19 years old, 17%) and young adult (20–29 years old, 45%) age groups, with the remaining participants being almost equally distributed within four older age groups from 30–39 to 60–69 years old (38% in total). Most of the participants (67%) were nationals of European countries, and the others represented Asian, North- and South-American countries in roughly equal ratios. All participants reported that they used computers and smartphones on a daily basis and were active users of different messaging platforms (e.g. WhatsApp, Telegram, WeChat) for communication with family, friends, and colleagues; their experience with chatbots was more narrow. All participants saw customer service chatbots online, but only a few (27.8%) mentioned using them before. Most participants (94.4%) indicated their familiarity with intelligent assistants (e.g. Siri, Google Assistant, Alexa, etc.), of this majority (55.6%) did not use them anymore (as at the time of the interview) and others used them infrequently (38.8%).

None of the participants had experience with a dedicated open-domain chatbot, nor was aware of their existence.

3.4.3 Interviews

Once they agreed to take part in our study, the participants were asked to complete a basic demographic survey about their age group, nationality, and occupation. The following semi-structured interviews were organized either in-person (11 cases), or via Skype videoconference (7 cases). All of them took place between 2nd and 23rd October 2019, with each interview lasting about 40 minutes. All interviews but one were audio-recorded, with the participants' consent, and all were accompanied with hand-written notes either by the interviewer or interviewer's colleague. Each interview covered four sections: 1) general questions about participant's background and experience with technology (computer, smartphone, messaging apps); 2) knowledge of and previous experience with chatbots; 3) qualities desired from the emotionally aware chatbots to make interaction with them more natural (if any), and purposes of such agents; 4) any concerns the participant might have about using such chatbots. Specifically, the first two parts were adjusted to make the participants speculate about their recent experience of social conversations (in person and via messaging apps) and interaction with chatbot technology respectively. This part helped the participants to draw the parallels between their human-human and human-machine communication experience and took about 15 minutes in each interview. In the following core parts of the sessions, the interviewees were asked to reflect on this comparison of communication experience and express what they expect from natural conversations with emotionally intelligent chatbots and what could make them feel restrained from using the system. These discussions lasted for 25 minutes on average.

3.4.4 Data analysis

We used affinity diagramming to analyze the interview content [204]. We favored this relatively flexible technique over more complex methodologies such as grounded theory [213, 91], as our main purpose in this study was to explore the basic concepts related to our research questions rather than construct a substantive theory. Both procedures follow a bottom-up inductive approach where categories emerge from the data. Affinity diagramming was deemed more practical for our context as it offers visual representation of the data, allows for easier collaboration, and facilitates faster iteration process, which are crucial aspects for organizing unstructured qualitative data efficiently [214]. Researchers are using this technique more often these days, for answering such questions as how analysts cope with uncertainty in data [17], what support is needed by industry practitioners for developing fairer machine learning systems [92], and how marginalized communities envision future technologies [84]. The specific steps that we followed throughout the analysis are elaborated below.

After each interview, the first author enriched the hand-written interview notes with missing comments and observations from audio records and extracted affinity notes with meaningful quotations from the participants and the researcher's remarks. If necessary, the interview guide was modified slightly before the next interview took place to ensure that all questions would be well understood by the participant. After all interviews were finished, several iterations of affinity diagramming took place. During the preliminary analysis, the first author clustered all resulting affinity notes according to emerging themes and validated the result with the second author. Three large themes describing the chatbot's naturalness properties, participants' concerns, and application domains arose. The concept of emotion comprised a substantial part of naturalness and was also present in the other two themes. Overall, 400 affinity notes related to the concept of emotional awareness in chatbots, which accounted for over half of all affinity notes in the initial diagram. Further, we examined these notes more closely. Specifically, the first author distributed emotion-related notes into sub-clusters and summarized their content with one representative topic. The sub-clusters, in turn, were grouped under top-level categories. The resulting affinity diagram was reviewed together with the second author and refined to reach its final version (Figure 3.3).

3.5 Results: Qualitative findings

This section presents the findings of our qualitative study based on the affinity diagram shown in Figure 3.3. Subsections 3.5.1, 3.5.2, and 3.5.3 elucidate the notion of social and emotional etiquette for chatbot by describing the three top-level themes under this category. Further, Subsection 3.5.4 introduces users' concerns associated with emotionally aware agents.

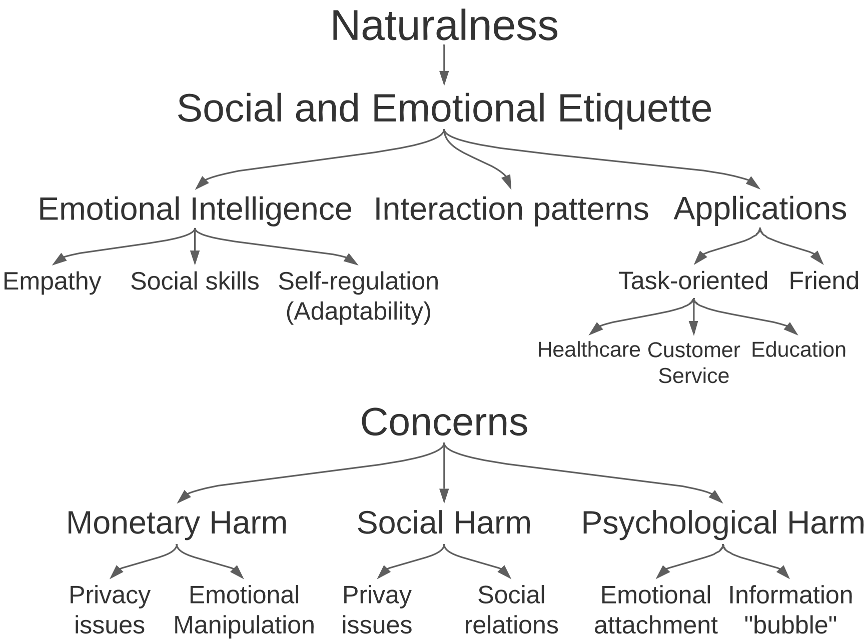


Figure 3.3: Structure of the affinity diagram.

3.5.1 Expectations of emotional intelligence in chatbots

All participants of our study agreed that enabling more human-like social behavior for conversational agents could facilitate the interaction. Sixteen out of 18 interviewees expressed varying degrees of interest in chatbots with enhanced emotional capabilities. Seven participants felt highly enthusiastic about such agents, and the remaining 9 showed moderate excitement. Their expectations largely complied with an established notion of emotional intelligence, which includes: self-awareness, self-regulation, motivation, empathy, and social skills [77, 78]. As self-awareness and motivation rather refer to subjects that are endowed with consciousness, people attributed the other three qualities – *empathy*, *social skills*, and *self-regulation* – to their desired artificial conversational agents.

Empathy

Empathy is our ability to sense the feelings and emotions of others, take their perspective, understand their needs and concerns [78]. When describing their expectations of the chatbot's emotional behavior, the participants highlighted two main components: *recognition* of the speaker's emotional state and *expression* of emotion in accordance with the context. The principal desire was to feel understood by the chat agent and receive appropriate responses. As noted by U04: *"It should sound as if it has emotions, not only one emotion for all times. For example, it could be sad or happy or something like that: maybe, happy when you're happy and understanding when you're sad."*

In addition to a straightforward way to treat the speaker's emotion by explicitly referencing the feeling (e.g. "I see that you are frustrated."), a number of other more subtle approaches were discussed during the interviews. Several participants mentioned interjections, *"phases that people have in a usual talk, like "am", "ah", "seems to be", "you know… ""* (U09), as a way to express reactions, emotional states, and thought processes. Emojis and emoticons were also referenced as a remarkable way of revealing emotion in chat. For example, U17 commented: *"I use them sometimes to convey the atmosphere of a "smiling conversation"."*

Social skills

Social skills concern the way how we manage relationships with others. These include a broad range of competencies from knowing how to communicate smoothly and managing conflicts to cooperating and bonding with people [78]. Speculating about their potential interaction, younger participants (below 30 years old) tended to be more open-minded about the social aspect of chatbots in everyday life. They enjoyed the idea of a conversational agent that could convey emotions during the dialog and presumed they would treat it as a friend. Interviewees felt excited about the possibility to engage with chatbots and share their feelings especially when they feel bored, lonely, or lacking motivation, as exemplified by the quote from U11: *"Some people have only one person they are close to, so they might need another one. So for them,* *the [emotionally intelligent chatbot] would be very useful: not to feel alone and to actually feel like they are talking to someone and sharing something."*

Meanwhile, both younger and older participants expressed interest in social skills for task-oriented chatbots. From their perspective, it could improve their current experience in several domains by ensuring more appropriate responses and alleviating the embarrassment of talking to a new person.

Self-regulation

In relation to chatbots, the most frequently mentioned principles of self-regulation included trustworthiness and adaptability. The recurrent topic reflecting anticipated interaction development with the chatbot concerned *"familiarity level"* (U06) with the user. Several participants commented that receiving overly positive replies from someone barely known would seem odd and awkward. Similar to relationship development with a newly acquainted person, participants expected the chatbot to consider personal boundaries and gradually adjust to their style, motives, and language. Participant U06 pointed her concern about appropriate conversational style and importance of social chitchat for her: *"Maybe it's different for my generation, but when I write an email or a message on WhatsApp, I always say 'Bonjour ... ' and some greetings. I think this is quite important."* Participant U08 further supported the idea with another example from her personal experience: *"I really like that some software, it tries to learn my language... it will predict what I would like to say in a way I personally say. So, it adapts to my style."*

Depending on the participant's needs and attitude towards the natural language agents, some of them preferred the interaction to follow a more formal style. In contrast, others expected it to develop informally, similar to the way of communication with their friends. For example, U17 welcomed the idea to develop a more close relationship with the chatbot: *"For me, it would be an amazing idea to have a kind of an online personal friend. So, you always share some thoughts with your friend, but this one can be both your diary and at the same time a psychologist who can always listen to you."* By comparison, U14 preferred more formal communication: *"... sometimes I find the service may be too cold. But, for example, when I was in the US for a bit, it was extremely warm and welcoming, to the point that I found it intrusive. So, yeah, I'd say it should be polite and understanding the problem I'm facing."*

3.5.2 Chatbots in the role of a friend

In our study, 10 out of 18 participants discussed the possibility to develop a friendship with a conversational agent in case it could demonstrate sufficient qualities of emotional awareness. They agreed that the chatbot should adjust to the user's emotional state, also taking its prior knowledge of the user into consideration, if possible. While it suggests a personalized approach, the participants concurrently described a number of emotional interaction patterns expected from the agent. The patterns mainly reflected the desired chatbot's responses to basic human emotions [194], such as happiness, sadness, or anger, and several more complex interactions. We summarize these expected patterns in Table 3.2 and consider them in greater detail below.

During the analysis, we observed that male participants tended to comment more on the playful and entertaining interaction aspects, while female interviewees mostly emphasized the chatbot's supporting abilities. Overall, the participants expected it to share their joyful moments, *"ask what happened"* (U17), and *"be happy with them"* (U02). In times of trouble, when feeling lonely or sad, the participants would anticipate understanding and compassion from the chatbot. U02 summarized these expectations as follows: *"I guess, if you're adding some excitement or frustration, then she [emotionally intelligent chatbot] should either be happy with you or try to make the voice more comforting."* Importantly, our participants would like chatbots to *"provide feedback, but not just generic"* (U16).

In some cases, potential users would desire the conversational agent to express coaching and motivational qualities. According to them, chatbots should encourage users *"to keep going"* (U07) both literally, promoting more physical activity and helping to establish a healthy lifestyle, and figuratively, supporting them when dealing with everyday problems. U05 would appreciate if a chatbot could assist him with behavior change: *"It would be good if it acts as a coach who helps you avoid a bad habit or encourage you to exercise."* Several other participants would like chatbots to "educate users to manage their anger" (U01): *"Maybe for me, a bot should calm you down when you're angry. [It should] say, "Stop, I cannot talk with you like that. If you don't calm down, I will turn off.""* (U03). Turning to chatbots to get inspiration and reassurance was another recurrently discussed topic: *"... if you have to spend long hours there, alone, doing some experiments, then it can make a conversation with you, cheer you up, look at your problems, maybe give some advice. It's a kind of a colleague that you might not have"* (U02).

Table 3.2: Expected emotional interaction patterns described recurrently by the participants.

Input emotion	Response emotion
Happiness	Happiness
Loneliness, Sadness	Compassion, Interest
Anger	Disengagement
Disappointment	Motivation
Frustration	Non-judging support

Aligned with previous findings [256, 19], our participants expressed eagerness to share their frustration and negative thoughts with the chatbot due to the non-judging nature of such interaction. They found it appealing to have someone always available to validate their anxiety and stress without condemning the users. As spotted by U18: *"If it's very natural, it can also be in the consulting domain... Consulting – sometimes emotionally, sometimes professionally, like therapy."* Curiously, just having an empathetic listener to vent out was not sufficient. From the participants' perspective, the crucial part of this interaction scenario was to receive some non-generic feedback from the chatbot, either advising the user how to overcome the problem or helping them to take their mind off by *"starting another topic [for conversation]"* (U04).

3.5.3 Emotionally aware chatbots in targeted domains

Supporting the idea that social intelligence should be integrated into task-oriented chatbots in the future [130, 244, 254], almost all of our study participants delved into discussing the emotional awareness of these chatbots. Many interviewees took a positive attitude towards emotionally aware chatbots for customer service, health care, and educational domains. They expected that chatbots could potentially eliminate issues caused by human factors: computer agents are not subject to stress and tiredness and could always offer comforting advice to the client. In the case of customer service it could ensure *"more natural and pleasant"* responses, so that *"people would actually want to call customer service instead of googling their problem"* (U11). For medical advice, several participants anticipated responses from the chatbot to be more attentive than the ones from *"an over-worked, over-stressed doctor"* (U15).

For the area of educational and professional training, several participants pointed out that conversational agents could make the services more available along with expressing higher involvement and interest in the tutoring sessions. Both for health care and educational domains, some interviewees mentioned the clients might feel less embarrassed to share their questions with a chatbot than with an unknown person. For example, U14 mentioned: *"I guess, for some medical issues people may be shy to actually talk to a real doctor... So, for this case, it [emotionally intelligent chatbot] could be quite good"* (U14).

3.5.4 Three pillars of user concerns

In line with previous studies [141, 256, 43], the main factors causing user worries around conversational agents were uncertainty about the trustworthiness and reliability of the system, as well as the risk of private information exposure. Chatbot's ability to treat emotions and exhibit social qualities provoked several additional topics that disturbed our interview participants. During the analysis, we identified three major categories that describe user concerns about chatbots: *monetary harm*, *social harm*, and *psychological harm*.

Monetary harm

Predictably, financial damage primarily involved the risks around the participants' immediate personal means, such as bank accounts or social security numbers. People also felt apprehensive about the threat to employment opportunities in case the technology reaches sufficiently natural conversational abilities. Potential emotional awareness of chatbots further increased these concerns as people feared that for intruders, *"it would be easier to influence you with emotion"* (U05).

Social harm

Concerns about the consequences for the social status of the users developed around the risks of sensitive information misuse by the chatbot operators. People questioned how the information they share with the agents would be stored and whether the third parties could use it. They worried that in case of disclosure, some pieces of data might be used against themselves and cause social embarrassment. Participant U11 questioned: *"What if it remembers something you shouldn't have said?"* Participant U14 further echoed her worry: *"If there's anything linked to some kind of psychology, I would be very scared of what is being kept [by the chatbot], because in the future you can be considered unbalanced, or whatever."*

Several participants also felt wary of the possibly addictive effect of highly human-like conversational agents. Similar to the way how excessive smartphone usage negatively affects our social relations [68], they were concerned that users might get too obsessed with flawless *"virtual friends"* (U10) and isolate themselves from real human society. Participant U02 found this especially alarming for children: *"I wouldn't want children to use this technology, for them not to get used to talking to a computer all the time instead of real people."*

Psychological harm

Sometimes people develop an emotional attachment to objects and may experience anxiety and other negative emotions when facing a risk of losing these items [250]. Our participants mentioned that people would highly likely establish an affective connection with emotionally aware chatbots. In this case, a technical glitch or agent's discontinuation could cause strong user distress: *"If some system or electricity failure happens, and the system gets reset, a person might not understand why it cannot remember anything anymore and feel very upset"* (U02).

Another thought-provoking point arose from people's experience with existing media resources. Several participants noted that some media adapts to the personal interests of its users and focuses all the suggested content around them, possibly depriving the alternative views or unintentionally hiding *"the best option"* (U14) from the user. It may deceive the users leading them to get trapped in *"their bubble"* (U16), believing that everyone around adheres to the same beliefs. Some of our participants were concerned that, given their anticipated personalization features, artificial conversational agents may further exacerbate this problem and cause psychological discomfort for the users. Participant U07 exemplified it with a personal anecdote: *"I am also very worried . . . about the control the media has to shape my thinking, especially on Facebook. . . . It shows me posts that have the same point of view as other posts that I've read. I might read posts of some political area and then it will show me lots of similar posts. So, I might gradually start thinking that that's the only point of view."*

3.6 Model development

The results of the exploratory qualitative study presented above provided additional evidence for the importance of endowing chatbots with social and emotional intelligence. They elucidate the ways how conversational agents should treat users' emotions as well as reveal other desired social qualities both for purely open-domain and hybrid agents. As our ultimate aim is to analyze the influence of various social skills of chatbots on user intention to adopt them, we have to first identify the constructs of the model and generate sample questions representing the concepts under consideration. To further lay out the constructs into a cascade structure, we decided to draw upon main findings from the Expectation-Confirmation Theory (ECT) [13]. It stipulates that before the purchase or adoption, users form a set of initial expectations about a specific product or service, which further influence their initial decision to adopt the given item. In turn, according to Ratchford and Barnhart, these expectations may be shaped by the users' individual characteristics and attitudes [189]. Therefore, we structured the question items into three high-level layers of constructs: *user attitudes*, *expected chatbot qualities*, and *behavioral intentions*.

We identified essential user experience criteria to form the principal constructs by utilizing the results of our exploratory study, and enhancing them by carefully surveying additional relevant literature. The consecutive work involved several iterations of phrasing the participating questions to ensure their semantic clarity for survey respondents. Below, we explain the constructs in more detail by linking them to the qualitative findings and reviewing the existing works that have inspired us to derive them and describe the hypotheses relating the constructs.

3.6.1 User attitudes

The first layer of the model, *user attitudes*, describes the extent to which a person has a favorable or unfavorable appraisal of given objects and actions [3]. Depending on the individual attitudes, people may form distinctive expectations about their interaction with new technology [189]. Factors constituting this dimension are exogenous variables of the model; thus, we needed evidence from prior work that they indeed influence user expectations and behaviors. We focus on three principal dimensions: user openness to technologies, vulnerability, and empathy propensity.

Openness to technologies

Openness to technologies measures user's personal disposition towards new IT products and services. Previously, in a study evaluating users' expectations and acceptance of futuristic Augmented Reality (AR) scenarios, Olsson et al. demonstrated that evaluation results might be considerably affected by the respondents' general orientation and attitude towards novel technologies [165]. In our qualitative study, we also observed varying degrees of interest in the social skills of chatbots among participants (cf. Section 3.5.1). Therefore, we adapted the items from the Technology Adoption Propensity (TAP) index [189] to account for this factor. Specifically, we combined the questions from the "optimism" and "proficiency" sub-scales, which are identified as two attitudinal factors contributing towards the technology adoption. The "optimism" sub-scale assesses users' belief that new technologies provide more flexibility in their everyday lives, and the "proficiency" sub-scale evaluates their sense of being technologically competent. We decided to merge the items from these two sub-scales under a single construct to save the respondents' cognitive load.

Vulnerability

Vulnerability is another construct of the TAP index, which refers to users' belief that technology increases their chances of being taken advantage of by criminals or firms [189]. Contrary to the previously discussed constructs, vulnerability acts as an inhibitor of adoption propensity; users who have significant concerns about their security and privacy are typically reluctant to use new technologies for fear of being victimized. Our qualitative findings revealed three principal factors affecting users' sense of vulnerability (cf. Section 3.5.4), which are also considered in several related works: **monetary harm**, **social harm**, and **psychological harm**. Monetary harm refers to the potential damage of the misuse of a user's financial data, e.g., her banking information being sold to third parties [43, 256, 99]. Social harm is associated with disclosing sensitive personal information, such as announcing a significant life event, which might lead to social embarrassment [43, 99, 256]. Finally, psychological harm concerns the anxiety that the user might experience upon technology malfunction in case he develops a strong emotional attachment with the chatbot [250].

Empathy propensity

Empathy propensity describes a person's disposition to understand and respond adaptively to others' emotions. Prior research in affective neuroscience found that people establishing empathetic attitudes towards others tend to prefer closer interpersonal distance during social interaction [173]. As users typically apply the norms and principles of social interaction to technology [191], their expectations of chatbots' social qualities should be influenced by this factor. Thus far, Tsiourti et al. employed empathy propensity assessment to evaluate user perception of the robot's emotion and concluded that users with a higher disposition of empathy could manage this task more accurately [227]. In a study of user requirements for chatbot design, Liao et al. demonstrated that users with higher social-agent orientation prefer their chatbots to handle conversations in a more human-like manner, providing subjective and opinionated answers and presenting relational behaviors [129]. In our exploratory study, female participants sought chatbot's support more eagerly than males (cf. Section 3.5.2), which might also be linked to differences in their empathic disposition [152]. For the quantitative evaluation, we decided to exploit the Toronto Empathy Questionnaire [209] to assess users' empathy propensity and explore its influence on their expectations.

3.6.2 Expected chatbot qualities

The second layer in the model, *expected chatbot qualities*, refers to different social and emotional qualities that users might expect from the chatbots. At this layer, we combined a wide range of qualities commonly discussed in the related work on user experience with chatbots as well as papers focusing on the technical implementation of these skills.

Politeness

The politeness construct captures users' desire for chatbots to demonstrate respectful and considerate behavior in several ways. Firstly, it reflects the expectation for chatbots to apply good manners, respond decently, and avoid rude language (cf. Section 3.5.1), as agents failing to comply with these norms cause embarrassment and frustration [62]. Additionally, the results of both ours (cf. Section 3.5.2) and prior qualitative studies indicated that the idea of a chatbot that can listen to human interlocutors without judging them was appealing to the users [256, 18, 148]. As people may sometimes find it demanding to adhere to politeness norms around-the-clock regardless of the situation (cf. Section 3.5.3), this quality of chatbots creates a unique value proposition for their users [256]. Well-mannered, not argumentative agents are more likely to inspire trust and ensure user engagement [148].

Entertainment

Entertainment is an essential aspect of chatbots. Many users first decide to engage with these agents due to the promise of playful interactions [141]. As a construct in our model, it comprises several interconnected social skills that users seek in chatbots. One of them is generating compelling and diverse responses to please the users and keep their company (cf. Sections 3.5.1 and 3.5.2). Diversity in chatbot's responses is a challenging task that has been actively studied from the developmental perspective [125]. From the users' point of view, it could drive engagement in conversations with the chatbot, help avoid repetitions, and promote fun [205, 103]. Another essential element of entertainment is a chatbot's ability to demonstrate its sense of humor by generating amusing responses, puns, and witty one-liners [141, 148, 103, 253].

Along with it, merely establishing social contact with the user is also an important quality associated with the adoption of conversational agents [130]. Previous qualitative works considered introductory phrases, small talk, and social conversations as a type of playful interaction with the chatbot frequently sought by the users (cf. Table 3.1, Section 3.5.1) [103, 18, 129, 247]. Finally, we included the chatbot's ability to incorporate emojis and interjections into its responses as a part of the entertainment construct, which was also pointed by the participants of our qualitative study (cf. Section 3.5.1). Prior research has shown that the inclusion of such cognitive-emotional expression improves user experience and acts as "socio-affective glue" in developing rapport between the user and the agent [39, 57].

Attentive curiosity

Attentive curiosity measures users' expectations about the chatbot's active listening, anticipation, and adaptability skills. Active listening is the ability to understand and respond to the user appropriately [244]. This chatbot's ability was frequently alluded to by our qualitative study participants (cf. Section 3.5.2). According to previous works, chatbots with active listening skills could elicit quality user responses [244] and appeared more appealing and engaging [205]. One technique for chatbots to practice active listening is to ask impromptu questions based on user response, i.e., demonstrate curiosity [244]. Combined with the ability to "remember" previous user replies, chatbots can show attention to the users by anticipating their needs and bringing up exciting topics – a behavior that the users would highly appreciate [148, 256]. Moreover, question-asking could help chatbots avoid misunderstanding with the users [206] and reason about their preferred social distance to adjust the conversational style accordingly [158], which was as well mentioned as a desirable quality in our exploratory study (cf. Section 3.5.1).

Empathy

Empathy is the chatbot's ability to recognize and respond to user emotions appropriately. Researchers focusing on open-domain chatbot development have been carrying long-lasting efforts to empower these agents with the ability to express emotions in their responses [8, 262, 246]. Several authors of user-centered studies previously mentioned the importance of integrating emotional intelligence into chatbots [256, 148, 244], and our qualitative study further clarified that users desire their chatbots to establish reassuring and empathetic behavior and revealed the specific emotional interaction patterns most expected by the users (cf. Sections 3.5.1 and 3.5.2). In addition, multiple technical papers demonstrated promising results by training neural network-based empathetic chatbots on the datasets of empathetic human dialogs [95, 187], which confirms the importance of considering this factor in our model.

Personality

This factor determines whether users expect a chatbot to exhibit a personality. It was not a primary focus in our exploratory study, but prior qualitative and WoZ studies reported that most users were willing to have their chatbots endowed with this quality. Jain et al. [103] found that people preferred agents with distinct personalities that matched their operational domain. In a study of a question-answer chatbot, Liao et al. [129] concluded that users with high social-agent orientation desired it to present a personality despite the agent's task-oriented nature. In a WoZ experiment, Thies et al. [148] specifically considered different personality traits for a chatbot to identify which ones would work best for their target audience.

3.6.3 Behavioral intentions

Behavioral intentions towards a chatbot are related to whether or not the chatbot can influence users' decision to use it if it were endowed with the discussed social and emotional skills. According to the Theory of Planned Behavior [229], behavioral intention can be a strong predictor of actual behavior. Thus, asking the users to evaluate their behavioral intentions concerning such chatbots based on their expectations is a useful and insightful approach, especially when actual chatbots having a full set of the discussed qualities are non-existent. Following the practice of previous adoption-related studies [183, 257], in the behavioral intentions construct, we assess users' intention to use the chatbot and their desire to introduce a chatbot to their friends.

3.6.4 Hypotheses

To evaluate our model, we form a set of hypotheses about how various constructs relate to each other (Figure 3.4). We hypothesize that users' attitudes have an influence on their expectations about chatbots. Specifically, we posit that openness to technologies and empathy propensity would have a positive effect on users' expectations about chatbot's social skills. At the same time, a sense of vulnerability would influence them negatively. We also hypothesize a significant positive causal effect from users' expectations to their intention to use socially and emotionally aware chatbots. In the following sections, we present the evaluation of our model and the hypotheses.

3.7 Results: Quantitative evaluation

3.7.1 Experiment setup

To validate the conceptualized model and the hypotheses, we developed a structured online survey. The survey questions focused on users' level of empathy, openness to technologies in general, expectations and concerns about conversational chatbots, and intention to use them. We used a validated scale to measure users' empathy propensity level [209] and designed other

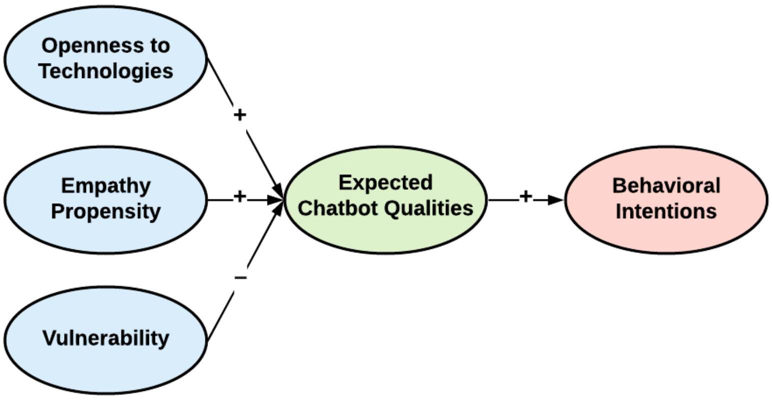


Figure 3.4: A general evaluation framework with hypothesized influence paths.

questions based on the previous work review discussed above.

We thoughtfully selected the surveyed population sample to ensure that our participants were familiar with chatbot technology and could meaningfully reflect their expectations in the responses. To account for possible limitations of the self-report data collection method, we carried a careful validity and reliability evaluation of the resultant model. The details are revealed in the following sections.

First, we launched two consecutive pilot surveys (N=50 each). The purpose of the pilot surveys was to verify the process of online data collection, test the understandability of the questions, and validate the payment adequacy based on participants' feedback. We then finalized the questionnaire based on the response patterns and feedback.

The survey consisted of four major sections: introduction and instructions, basic demographic survey, main questions, and the concluding part, where participants could provide their feedback about the survey in an open-ended manner. Demographic questions included age, gender, English fluency, and profession. The central part consisted of 64 Likert-type items covering constructs in our model with response options from strongly disagree (1) to strongly agree (5). We deliberately included several redundant questions, such as questions of similar meanings or semantically opposite and negated meanings, to check responses for adequacy and consistency and control data quality.

We conducted the main large-scale survey on Amazon Mechanical Turk (AMT). We set a payment of \$0.41 for completing the task. Paolacci et al. [169] suggested that AMT is a viable platform for data collection if the data quality is controlled. Therefore, we required a 95% HIT approval rate[2] for the AMT workers and additionally inspected the data for random or irresponsible answers. A bonus of 25 cents was paid to workers who passed the attention check. We also required all the participants to be based in the US for several reasons. We wanted to eliminate the possibility of question misunderstanding caused by language-related

---

[2] The percentage of completed by the worker Human Intelligence Tasks (HITs) that are approved by Requestors.

issues (our questionnaire was prepared in English). Besides, we considered the US population to be sufficiently homogeneous in terms of the average level of the population's familiarity with the chatbot technology [18], which was necessary to ensure that the respondents could relate to the questions. The survey was launched on July 6th at 5:30 p.m. (CEST) and completed on July 7th at 3:30 p.m. (CEST) in 2020. In total, 1000 workers completed the survey during the elapsed 22 hours.

3.7.2 Data quality control

We describe how we checked the responses for their quality. Firstly, we discarded 25 respondents as they failed to submit a complete survey form. Secondly, we computed the variance of each worker's responses to all questions. We discarded respondents with zero variance (who selected the same score for every question), as they did not have natural variation in their answers; there were 8 such cases in total. Then, we inspected the pairs of reverse scale questions to test inconsistencies in participants' responses. In total, we included 5 pairs of such questions. If a contradiction in worker's answers was found in more than 3 cases, their submission was discarded. We dropped 231 submissions based on this criterion. Lastly, we detected potential automatic bots by examining each worker's elapsed time to complete the survey. If the elapsed time was less than 5 minutes, the submission was considered invalid as it was unlikely for a human with average reading and clicking speeds to answer every question in 4 seconds. We filtered out 200 such cases. The final sample size of valid users became 536. The rule of thumb is to have at least 10 participants for each question item [118]; thus, our sample size is sufficient for a stable factor estimate.

3.7.3 Participants

Among the selected 536 participants, females (51%) and males (49%) were represented approximately equally. More than 60% of participants were in the 31-65 age group, 34% were in the 18-30 age group, and the rest belonged to users above 65. As mentioned earlier, all participants were based in the US, and the vast majority (96%) indicated that they were native English speakers. Their occupations were diverse: 14% were self-employed, 7% were engineers, and the remaining included managers (5%), students (5%), programmers (4%), homemakers (3%), and others. Based on their professions, we roughly split the participants into IT-related and not IT-related, resulting in 17% and 83% correspondingly. The details of all participants are summarized in Table 3.3.

One of the survey questions asked the participants to rate their previous experience with a *socially and emotionally aware* agent and specify this chatbot's name. Most participants scored low on this question (Mean: 1.94, SD: 1.27), validating our hypothesis that such agents are relatively scarce on the market. The prevailing majority could not provide the name of a chatbot endowed with social skills (97%). The most frequently mentioned agent was Alexa (6 mentions). However, the respondents who specified it indicated uncertainty about its social capabilities, as exemplified by the comment: “I use Alexa but I think she needs to be able to show more emotion.”

Table 3.3: Profile of participants (N=536).

	Item	N	Pct.
Age	18–30	180	33.58%
	31–65	336	62.69%
	Above 65	20	3.73%
Gender	Female	274	51.12%
	Male	260	48.51%
	Other	2	0.37%
English fluency	Native speaker	513	95.71%
	Non-native speaker	23	4.29%
Profession	IT-related	90	16.79%
	Not IT-related	446	83.21%

3.7.4 Analysis methods

We validated the rationality of each construct of the conceptualized model and their relationships using the techniques described in [164]. We first validated that a relationship between multiple observed variables and the underlying latent constructs exist by applying Confirmatory Factor Analysis (CFA). The CFA results are presented in Section 3.7.6. We then conducted path analysis in structural equation modeling (SEM) to examine the causal relationship between the latent variables in the model. The findings of SEM are reported in Section 3.7.7. We used the R-lavaan package version 0.6.6 to conduct the analysis.[3] We employed the maximum likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic (`MLM` estimator) to account for the non-normality of the Likert-type data [199].

3.7.5 Data screening

Before validating the model, we conducted data screening to ensure the dataset is reliable for testing causal relationships [38]. All data items were complete without any missing values. Second, we needed to ensure that the data for all variables are well modeled by a normal distribution to satisfy SEM assumptions [85]. We assessed the empirical distributions of responses for each question visually and eliminated the items that deviated considerably from the bell-curved distribution [85]. Further, we checked the skewness and kurtosis of the remaining items. Results show that the skewness $(-1.03, 0.15)$ and kurtosis $(-1.07, 0.35)$ of all variables fall into the $(-2.00, 2.00)$ interval, the recommended acceptable range of normality test for Likert-type questions [69]. Thus, our data also meet normality requirements.

---

[3] https://lavaan.ugent.be/

Table 3.4: Basic statistics and test results of internal reliability and convergent validity. Constructs with single items are included for completeness. CR and AVE stand for Composite Reliability and Average Variance Extracted, respectively.

Constructs *(N, Mean, SD)*	Internal reliability		Convergent validity		
	Cronb. alpha (0.6)	Item-total corr. (0.3)	Factor loading (0.5)	CR (0.6)	AVE (0.5)
Empathy Propensity (1)					
Total score based on 16 questions from Spreng et al. [209] *(0-64 scale; Mean: 44.89, SD: 11.01)*					
Openness to New Technologies (4)	0.81			0.81	0.52
I am a technology enthusiast. *(Mean: 3.63, SD: 1.07)*		0.72	0.72		
Given an opportunity, I always like to buy the latest smartphones/gadgets. *(Mean: 3.40, SD: 1.21)*		0.74	0.77		
I help my friends and relatives fix technical and user interface bugs in their devices. *(Mean: 3.36, SD: 1.20)*		0.63	0.63		
It is important to keep up with the latest trends in technology. *(Mean: 3.68, SD: 1.08)*		0.70	0.75		
Vulnerability: Monetary Harm (1)					
Interacting with emotionally aware agents can cause financial damage. *(Mean: 2.63, SD: 1.09)*					
Vulnerability: Social Harm (1)					
The disclosure of my conversations with emotionally aware agents can cause embarrassment. *(Mean: 3.05, SD: 1.24)*					
PEACE Construct: Politeness (3)	0.66			0.63	0.36
I would like the agent to restrain from judging me. *(Mean: 3.93, SD: 1.05)*		0.57	0.67		
I would like the agent to maintain emotional stability. *(Mean: 3.91, SD: 0.95)*		0.54	0.51		
I would like the agent to be polite. *(Mean: 4.27, SD: 0.89)*		0.63	0.61		
PEACE Construct: Entertainment (6)	0.87			0.87	0.53
I would like the agent to be able to entertain me. *(Mean: 3.40, SD: 1.24)*		0.75	0.75		
I would like the agent to be able to engage in a small talk with me. *(Mean: 3.31, SD: 1.25)*		0.77	0.78		

Continuation of Table 3.4

Constructs *(N, Mean, SD)*	Internal reliability		Convergent validity		
	Cronb. alpha (0.6)	Item-total corr. (0.3)	Factor loading (0.5)	CR (0.6)	AVE (0.5)
I would like the agent to have a sense of humor. *(Mean: 3.62, SD: 1.18)*		0.74	0.76		
I would like the agent to change conversation topics from time to time. *(Mean: 3.12, SD: 1.20)*		0.64	0.63		
I would like the agent to use emojis to express emotions. *(Mean: 3.15, SD: 1.34)*		0.70	0.69		
I would like the agent to use short exclamations, e.g. *Wow!* or *Ouch!*, etc., to express different feelings and reactions. *(Mean: 3.36, SD: 1.26)*		0.73	0.73		
PEACE Construct: Attentive Curiosity (7)	0.82			0.82	0.39
I would like the agent to participate actively in our conversations, e.g. ask questions. *(Mean: 3.98, SD: 0.97)*		0.65	0.64		
I would like the agent to anticipate my needs. *(Mean: 3.66, SD: 1.07)*		0.64	0.64		
I would like the agent to be able to resolve disagreement with me in case our opinions or views differ. *(Mean: 3.82, SD: 0.98)*		0.60	0.62		
I would like the agent to adapt to my conversational style. *(Mean: 3.63, SD: 1.03)*		0.53	0.56		
I would like the agent to remember my preferences from our previous conversations. *(Mean: 3.93, SD: 1.09)*		0.61	0.60		
I would like the agent to propose interesting ideas and information to me. *(Mean: 4.01, SD: 0.93)*		0.73	0.72		
I would like the agent to call me by my name. *(Mean: 3.80, SD: 1.04)*		0.58	0.60		
PEACE Construct: Empathy (4)	0.83			0.83	0.54
I would like the agent to be able to recognize my emotions. *(Mean: 3.59, SD: 1.10)*		0.72	0.72		
I would like the agent to be able to express emotions. *(Mean: 3.33, SD: 1.17)*		0.74	0.74		
I would like the agent to be empathetic. *(Mean: 3.61, SD: 1.08)*		0.74	0.75		
I would like the agent to encourage me. *(Mean: 3.68, SD: 1.14)*		0.68	0.74		

Continuation of Table 3.4

Constructs *(N, Mean, SD)*	Internal reliability		Convergent validity		
	Cronb. alpha (0.6)	Item-total corr. (0.3)	Factor loading (0.5)	CR (0.6)	AVE (0.5)
Behavioral Intentions (2))	0.77			0.77	0.63
If an emotionally aware agent existed on the market, I would use it. *(Mean: 3.41, SD: 1.17)*		0.71	0.82		
If an emotionally aware agent existed on the market, I would tell my friends about it. *(Mean: 3.57, SD: 1.22)*		0.71	0.76		

3.7.6 Model validity and reliability

To validate the model, we first assessed its internal consistency and reliability using Chronbach's alpha and item-to-total correlations. This procedure aims at revealing the internal consistencies of the constructs and identifying the clusters of related variables. The items with a low alpha value (< 0.6) were discarded or re-grouped into another construct. After several iterations, we obtained the values presented in Table 3.4. They meet the cut-off points of 0.6 for Chronbach's alpha [174] and 0.3 for item-to-total correlation [164].

We examined the convergent validity of the measurement items by composite reliability (CR) and average variance extracted (AVE) based on factor loadings from the Confirmatory Factor Analysis (CFA) [63]. The results are shown in Table 3.4. Factor loadings for all items exceeded the acceptable level of 0.5 [85]. Composite reliability for all constructs also exceeded the recommended level of 0.6 [63]. For the average variance extracted, most of the constructs met the recommended 0.5 level [63] except for two constructs, Politeness and Attentive Curiosity, which scored slightly below the threshold. However, according to Lam [119] and Fornell and Larcker [63], the average variance extracted may be a more conservative estimate of the validity of the measurement model, and "on the basis of p_n (composite reliability) alone, the researcher may conclude that the convergent validity of the construct is adequate, even though more than 50% of the variance is due to error" [63]. As CR values for the two constructs, Politeness and Attentive Curiosity, are above the recommended cut-off points, we concluded that the convergent validity of the measurement items is acceptable.

We also evaluated the discriminant validity of the constructs via the inter-construct correlation matrix, shown in Table 3.5. Correlations between any two constructs were less than the square root value of AVE (shown in the diagonal), which represented a level of appropriate discriminant validity [63]. The only exceptions are between Empathy Propensity and Politeness and between several expectation-related constructs. Since the constructs are interconnected, we consider them as acceptable for our model [31].

Table 3.5: Inter-construct correlation matrix.

	1	2	3	4	5	6	7	8	9
1. Empathy Prop.	1.000								
2. Open. to Tech.	.059	.718							
3. Monetary Harm	-.387	-.146	1.000						
4. Social Harm	-.241	-.035	.297	1.000					
5. Politeness	.626	.106	-.272	-.055	.593				
6. Empathy	.336	.384	-.333	-.029	.327	.737			
7. Attent. Curiosity	.452	.411	-.381	-.073	.656	.785	.624		
8. Entertainment	.160	.463	-.157	.088	.059	.790	.626	.727	
9. Behav. Intention	.218	.533	-.318	-.061	.198	.762	.661	.751	.789

To summarize, our model was validated as robust and satisfactory in terms of internal consistency reliability and convergent and discriminant validity. Next, we performed structural equation modeling to verify the hypotheses.

3.7.7 Structural equation modeling

We tested the overall model fit regarding our hypotheses (Figure 3.4) on the causal relationships among the three layers of constructs. Figure 3.5 shows the results of the structural model analysis, including the R2 (coefficients of determination) and path loadings. All the R2 estimates for the variables are larger than the threshold of 0.1; thus, they are appropriate and informative to examine the significance of the associated paths [183]. The model fit indices are $\chi^2 = 754.273$, $p < 0.001$, $df = 357$, $\chi^2/df = 2.11$ (< 3), $CFI = 0.921$ (> 0.9), $TLI = 0.910$ (> 0.9), $RMSEA = 0.051$ (< 0.08), $SRMS = 0.058$ (< 0.08), which surpass the recommended values of these model fit indices (shown in brackets) [85]. It is desirable for the χ^2 statistic to achieve statistically insignificant goodness of fit value, but in practice, it is difficult to obtain with large sample sizes [85]. For this reason, it is recommended to refer to the normed χ^2, defined as a ratio of χ^2 to the degrees of freedom for the model (χ^2/df), to correct for the bias against large samples and increased model complexity [85].

We first examined the path relationships between the constructs included in user attitudes and expected chatbot qualities. Openness to technologies has a significant positive influence on user expectations about chatbot's empathy ($\beta = 0.351, p < 0.001$), attentive curiosity ($\beta = 0.358, p < 0.001$), and entertainment ($\beta = 0.467, p < 0.001$) qualities. Empathy propensity significantly leads to the expected politeness of the chatbot ($\beta = 0.617, p < 0.001$). Moreover, the sense of vulnerability diminishes user expectations. Specifically, concerns about monetary harm have a significant negative effect on empathy ($\beta = -0.231, p < 0.001$), attentive curiosity ($\beta = -0.197, p < 0.001$), and entertainment ($\beta = -0.098, p < 0.1$). User worries about social harm also have a slightly negative influence on empathy ($\beta = -0.017, p > 0.1$) and attentive curiosity ($\beta = -0.012, p > 0.1$), though these paths are insignificant.

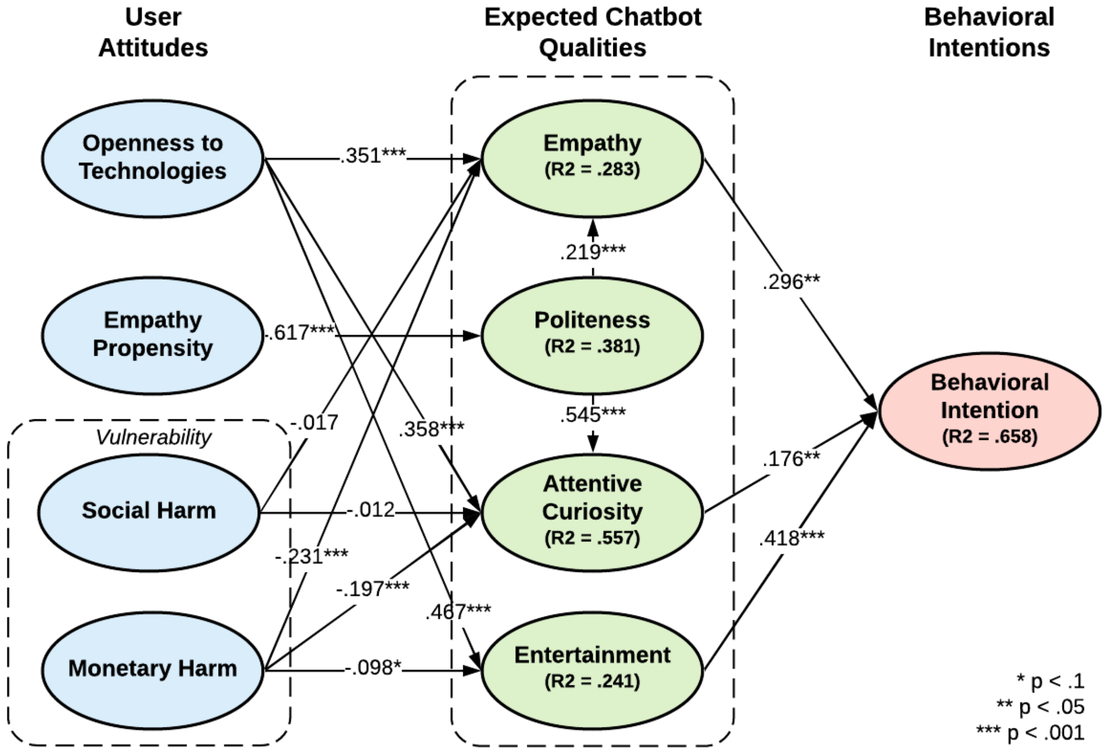


Figure 3.5: Structural model. Path significance: $***p < 0.001$, $**p < 0.05$, $*p < 0.1$.

Furthermore, user expectations about chatbot's politeness significantly affect the ones about chatbot's empathy ($\beta = 0.283$, $p < 0.001$) and attentive curiosity ($\beta = 0.545$, $p < 0.001$). As for the relationship between expected chatbot qualities and behavioral intentions, we found that expectations about chatbot's empathy ($\beta = 0.296$, $p < 0.05$), attentive curiosity ($\beta = 0.176$, $p < 0.05$), and entertainment ($\beta = 0.418$, $p < 0.001$) most influence the use and share intentions.

3.8 Discussion

The evaluation process validated the hypotheses that stated the relationship between users' attitudes, their expectations about chatbots' qualities, and behavioral intentions (Figure 3.4). It, therefore, produced the final model, comprising the essential social and emotional qualities for conversational chatbots. We name our model PEACE, according to its constituent constructs: Politeness, Entertainment, Attentive Curiosity, and Empathy. The analysis based on this model supported prior research results [129, 165], suggesting that personal differences in user beliefs and attitudes may significantly influence their expectations about future technology. We established that a generally positive attitude to technologies amplifies users' desire for conversational agents to display more natural social behavior. Likewise, a higher level of users' empathy propensity was found to predict their stronger desire for chatbot's politeness, understanding, and non-judging support. At the same time, the results of psychometric evaluation demonstrated that users' sense of vulnerability might impede their willingness to experience more socially aware technology in the future. The threat of becoming a victim of a financial fraud restrains user expectations, which is understandable as emotional triggers and social influence are typical techniques employed by perpetrators of online scams [242]. Besides, the risks of social embarrassment in case of sensitive information disclosure also slightly reduce user expectations about chatbot's empathetic and attentive curiosity skills, though this effect is less pronounced.

As further suggested by the PEACE model, special attention should be paid to the politeness factor. While there is no direct path from politeness to behavioral intentions, it has a strong significant influence on the other two expected chatbot qualities, making it a key player in the model. One necessary condition for a user to consider trying out a chatbot is its outstanding value proposition compared to existing services for that individual [18, 256]. Unconditional politeness and a premise of non-judging interaction are among the main requirement for open-domain chatbots [148, 256]. If this condition is satisfied, users may consider engaging further in more social interaction with the agent.

Interestingly, the politeness construct has a more substantial influence on attentive curiosity rather than empathy. One reasoning behind this finding could be the following. Empathetic behavior suggests demonstrating compassion and support but not necessarily scrutinizing the details about the user's particular emotional state or experience. Meanwhile, attentive curiosity implies an agent's follow-up questions that could help it learn more context about the user. People typically feel more comfortable sharing personal information when they have a perception of psychological safety to communicate, i.e., the possibility to speak up without risks of being blamed or disapproved [249]. Thus, politeness appears to be even more critical for scenarios when a chatbot inquires further details and information from the user.

The other three social qualities, apart from politeness, have a direct positive influence on behavioral intentions. Unsurprisingly, entertainment shows the most prominent effect. So far, this is probably the only skill that is already widely implemented in existing chatbots, e.g., Google's Easter Eggs.[4] Possibly drawing on their previous positive experience with playful agents, users demonstrated a notable desire for this quality to be maintained and further developed in the chatbots. On the contrary, attentive curiosity and empathy of chatbots are much less familiar constructs for the users. Only few research works study user perceptions of these qualities in prototypical chatbots so far, e.g., [95, 244], so the vast majority of the respondents had to rely on their genuine desire to judge these constructs. Nevertheless, we detected a strong user interest for these chatbot qualities as well. Almost all questions comprising empathy and attentive curiosity factors scored higher than 3.5 points out of 5 (Table 3.4), and the constructs themselves established a significant predicting power on user behavioral intentions.

The presented PEACE model shows promising prospects for future open-domain and hybrid

---

[4] https://en.wikipedia.org/wiki/List_of_Google_Easter_eggs

chatbots endowed with social and emotional intelligence. To inform the community of chatbot developers and designers, we extract the key implications of our study below and outline our vision for future work.

3.8.1 Implications

Users' attitudes strongly influence expectations

The evaluation of the PEACE model demonstrated that individual differences in people's attitudes play a considerable role in forming their expectations about future chatbot technology. As modern applications typically allow service providers to get at least some initial knowledge about the user, it is recommended to use this advantage to tailor user experience with their chatbot and ensure a better match to user expectations. In line with prior findings [129], we confirmed that users with higher empathy propensity are much more interested in signs of humanness in chatbots than their more pragmatic counterparts. Further, if the user is a tech-savvy person, she is even more likely to appreciate the social qualities of a chatbot. No less important, one should be aware of user concerns and counterbalance with benevolence and privacy guarantees. To promote user trust in the technology, designers need to consider revealing the integrity of their chatbot [144], i.e. its operational principles that would be acceptable for the users, such as transparent privacy policies and data security. Further, demonstrating that the chatbot is driven primarily by the user interests rather than other concealed motives could potentially offset user apprehension [144].

Politeness is key

Politeness proved to be an essential prerequisite for people to consider engagement with socially intelligent conversational agents. Given the popularity of data-driven approaches to developing novel neural network-based chatbots, it is critical for developers to meticulously control the quality of the training data for their conversational models. Otherwise, it is unlikely that the resulting chatbot will last long, as it was well illustrated by the example of the infamous Tay chatbot.[5] After being released on Twitter to learn from uncontrolled crowd-sourced data, it was terminated in less than 20 hours for its excessively offensive and inflammatory comments [159]. To avoid such scenarios, developers should opt for carefully prepared high-quality datasets, e.g., as provided in work by Rashkin et al. [187], or regulate unethical chatbot's replies by other means, e.g., employing active learning techniques [134].

Entertainment is king

Entertainment and humor form a common "entry point" to chatbot technology, even for the functional chatbots operating in domain-specific settings [141, 130]. Our analysis of the open-

---

[5] https://en.wikipedia.org/wiki/Tay_(bot)

domain case further supported the significance of this quality, demonstrating its superior leading effect on user adoption intention compared to the three other qualities. Developers should put effort into cultivating a chatbot's sense of humor and reinforcing it with the use of relevant interjections and emojis. Such implicit language markers can make the conversation more natural and engaging to the users [143, 39]. Emotive interjections [76] could enable chatbots to validate user emotions in a subtle and realistic manner, for example by saying *Wow!* to express a positive surprise or *Yikes!* to confirm their awareness of something bad or unexpected. Likewise, emojis and emoticons can be employed by chatbots to regulate the interaction, similar to how people use them in computer-mediated communication between each other [50]. Notably, the ability to engage in small talk is also considered a playful aspect of the agent's behavior [130] and should be provisioned as a part of the chatbot's social intelligence.

Attentive curiosity is essential for personalization

According to our findings, users would be willing to share their thoughts and information with the agent once the politeness requirement is satisfied. We recommend that developers incorporate active listening and impromptu questioning capabilities into their chatbots to gradually build a user image and use it to accommodate user expectations. Impromptu questions can be generated based on the user's conversational intent disclosed in the input utterance [212] to deepen the conversation and elicit additional knowledge about the user [255]. One essential factor shaping the user experience is the chatbot's alignment with the user style and language. Adapting it accordingly may ensure more personalization and achieve successful and favorable communication as in human dialogs [20, 177]. Moreover, being informed about its users' extended profiles, a conversational agent should be able to anticipate their needs better and propose topics for discussion matching users' interests with greater precision.

Empathy shows support

As suggested by our analysis, chatbots' ability to understand users' emotional states and provide the necessary emotional support is one of the decisive factors influencing adoption intention. Thus, it is vital for future agents to combine both the ability to detect the emotion expressed by the user accurately and generate an emotionally appropriate response. Some promising results have already been demonstrated for the customer service chatbot [95], and we suggest that they be extended further on a more general conversational case. Special attention should be paid to users' negative emotions. According to observations from social psychology, during communication people typically mimic each other's emotional states [210]. In many technical papers focusing on the development of emotionally aware chatbots, the authors tend to employ this principle either by training their models to follow the interaction patterns existing in the human dialogs corpora [262] or even explicitly configuring the model to minimize the affect dissonance between the user input and generated response [8]. Our study revealed that this mirroring approach is only partially valid. More specifically, users indeed expect chatbots to echo their positive emotions, for example, to share and promote user happiness. However, when experiencing negative feelings, people prefer the agent to act more intelligently than simply mirroring the speakers' emotions (Table 3.2). Designers should enable the agents with abilities to demonstrate attention and meaningful support to help users overcome negative sentiments. People are more likely to engage with empathetic chatbots that can give personalized feedback to the users.

3.8.2 Limitations and future work

Several questions and constructs were eliminated from the model due to low correlations with other variables, such as psychological harm and user expectations about the chatbot's personality. While users might experience monetary and social harms during interaction with other IT systems, psychological harm is arguably the least familiar to them, causing its elimination from our model. Once chatbots with advanced social skills become more widespread, the influence of psychological harm on adoption should be revisited. We presume that personality failed to segregate as a standalone construct as the traits defining one's personality (i.e., the Big Five personality traits [145]) are considerably dispersed across the other four established constructs. Therefore, we still recommend treating personality as an important aspect of a conversational chatbot.[6] However, to identify its influence on user adoption, more controlled experiments should be performed to show which personality traits are the most influential.

For our qualitative study, we invited participants from diverse backgrounds due to its exploratory nature and then chose to target our quantitative survey at the US-based population primarily to resolve potential language-related issues. Interestingly, despite the differences in the geolocation of the two studied population samples, we observed conceptual alignment of the results, suggesting that this demographic variable had a minor role in the development of the PEACE framework. Nevertheless, it does not strictly restrain the possibility of presence of more subtle differences across populations and ethnic groups. Future research should further validate the PEACE model with other populations and analyze cultural influences (e.g., North American vs. European vs. Asian users). Employing alternative research methods, such as analyzing user reviews of existing open-domain chatbots [185], could also be beneficial to offer additional validity assessment for the findings originating from the self-report studies. Even though user reviews can also be viewed as self-report data, users contribute them in the context of an existing product and usually on a self-motivated basis, which makes these data less susceptible to several biases (e.g., social-desirability bias) indicating its usefulness for methodological triangulation. Alternatively, to improve the validity of self-report studies, researchers could extend subjective attitudinal measures with more objective behavioral measures by asking participants about their intended behavior associated with the factor of interest. For example, Zhou et al. [264] measured users' willingness to act on social media

---

[6] The basic statistics for the item "I would like the agent to exhibit a personality" are: (Mean: 3.68, SD: 1.10).

posts to approximate users' trust in those posts. However, careful development of appropriate behavioral measures that can act as a proxy for attitudinal ones is necessary.

Our qualitative findings revealed the emotional interaction patterns most expected by the chatbot users. We propose further extending these results with more detailed studies. The community would benefit from a subtle analysis of expected interaction principles based on more fine-grained emotional categories. Follow-up research could formalize the interaction rules more systematically by employing an established taxonomy of emotions, e.g. Plutchik's wheel of emotions [180].

Finally, this study focused on user expectations as they provide a reference level for users to form evaluative judgments about the system and serve as an additional determinant of satisfaction [13]. To extend the established results and account for potential changes in user expectations following their interaction experience with the technology, our future technical work includes exploring the ways to develop a chatbot with the discussed social skills and evaluating the actual user experience with such an agent. Once a sufficiently stable and functioning prototype of a socially intelligent chatbot is available, a similar study could extend the current work by further evaluating how behavioral intentions influence users' actual use of the system. For each particular application, it might be necessary to tailor the degree of the chatbot's social and emotional capabilities considering its specific purpose and operational domain.

3.9 Chapter summary

This chapter presented the exploratory analysis of user expectations and concerns about socially and emotionally aware chatbots along with an overview of recent user experience research devoted to such agents. Careful examination of the results from numerous related studies led to the conceptualization of the PEACE model – the first consolidated model of essential social and emotional qualities of chatbots that predict user desire to accept this technology. The psychometric evaluation procedure confirmed that our model provided the validity and reliability of its constructs and verified that the paths between them established meaningful causal relationships.

In our work, we detailed how we combined and validated existing criteria into a unified model. This model defines key determinants motivating the adoption intention of open-domain conversational agents, i.e., the agent's abilities to follow the **P**oliteness protocol, **E**ntertain the user, practice **A**ttentive **C**uriosity towards the user, and express **E**mpathy, thus PEACE. Our results were able to extend beyond prior work, which focused only on few assessment criteria at a time, and provided a meaningful explanation of the overall user requirements and desiderata for future socially empowered chatbots. Finally, we summarized the principal insights of the study in a set of design implications informing future efforts of the design and development of emotionally and socially aware conversational agents.

4 Eliciting expectations from user online reviews[*]

4.1 Introduction

Chapter 3 along with several previous studies focused on user expectations of open-domain chatbots to understand which social traits are essential for them to deliver a compelling experience [148, 112, 108]. Most of these works aimed attention at only one aspect of a variety of possible social skills, such as chatbot's personality [148] or emotional capabilities [108]. More importantly, previous studies favored simulating interaction experience with chatbot prototypes over studying user interaction with existing agents because of the scarcity of fully functional open-domain chatbots. Therefore, they lack insight into how well the identified expectations align with the current chatbots' capabilities. Eliciting user needs based on their experience with actual technology could provide a more holistic view of the subject and help determine users' principal expectations and pain points.

In this chapter, we aim to investigate the desired conversation skills and social qualities of open-domain chatbots. Towards this goal, we analyzed user experience and expectations drawing on online reviews of 16 chatbots posted on Google Play. We combined the findings from statistical analysis and qualitative thematic analysis of over 500 user reviews. The results indicate that currently users mostly value the entertaining component of their experience but expect the chatbots to develop more complex social behavior in the future. In the following, we first survey related user-centered studies of conversational agents. Further, we describe our study design, methodology, and main findings. Finally, we summarize the implications of our work and discuss the directions for developing social skills of open-domain chatbots.

---

[*] Adapted from [217]: User Expectations of Conversational Chatbots Based on Online Reviews. Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. Designing Interactive Systems Conference (DIS), 2021.

4.2 Related work

As chatbots are gaining popularity, studies exploring user conversational interactions with them emerged. Several works mainly focused on the current user experience and perception of chatbots [185, 33, 103, 158, 130]. Purington et al. [185] conducted a case study of Alexa, a virtual assistant developed by Amazon. The authors evaluated users' social experience with the device through an analysis of customer reviews. One of their central findings is that a greater personification of Alexa is linked with more social interactions. Cho et al. [33] investigated the evolution of user interactions with Alexa in a long-term diary study. They identified that the lack of engagement from the agent's side led to a loss of its presence in users' everyday lives. Both Jain et al. [103] and Muresan and Pohl [158] studied the experience of first-time users. Jain et al. [103] employed quantitative and qualitative methods to explore user interaction patterns with eight chatbots on the Facebook Messenger platform. Their findings revealed that users prefer chatbots that are human-like when conversing in natural language. Muresan and Pohl [158] conducted a qualitative diary study of the Replika chatbot and concluded that its failure to adhere to social norms might be detrimental to user engagement. Liao et al. [130] emphasized the importance of social skills even for a task-oriented chatbot whose primary purpose was to assist new employees with company-related questions. Based on the interaction log analysis, the authors established that over 30% of them constitute social dialogs indicating that users sought more playful and habitual communicative experience.

Most of the studies above analyzed user experience with chatbots classified as task-oriented agents or virtual assistants. In contrast, some researchers started to elicit future user needs and expectations of open-domain chatbots' social skills [148, 112, 108, 29]. Thies et al. [148] employed the Wizard-of-Oz design method to identify which out of three hypothetical companion chatbot personalities would be most compelling to their target audience. Their participants expected a chatbot to be entertaining, non-judgemental, and endowed with proactivity skills. Two other studies [112, 108] shed light on user expectations of chatbots' abilities to satisfy their emotional needs. Kim et al. [112] ran a qualitative analysis on the data gathered from the workshop and semi-structured interviews with teenagers. Their findings highlighted the importance of good listener behavior for conversational agents and their ability to understand and encourage the users. Katayama et al. [108] surveyed and interviewed the users to explore how they preferred a chatbot to adapt its interaction style depending on situational and emotional context. The authors proposed a regulation mechanism that elicited a better and more affective user experience with an emotion-aware chatbot prototype. Our Chapter 3 and work of Chaves and Gerosa [29] independently conceptualized models of essential social and emotional qualities of open-domain chatbots using respectively psychometric techniques and literature survey as research methods.

This study differs significantly from the existing work as we elicit user expectations drawing from their experience with existing open-domain agents available on the Google Play app store. While we got inspired by the methodological approach employed in [185], our focus and scale of the analysis are distinct. Similarly to our practice in Chapter 3 and approach taken in [29], we conduct a comprehensive study of the desired social skills of conversational agents. However, we explore the subject from a different perspective by grounding our analysis in users' real-world experiences.

4.3 Materials and methods

4.3.1 Study design

We conducted an exploratory analysis of online reviews to extract insights about the users' current experience and expectations of existing open-domain chatbots. Users voluntarily provide these reviews to share their thoughts about chatbots and their evaluation on a five-star rating scale with the public audience and chatbot developers. Researchers found the content analysis of web reviews an effective approach to understanding reviewer opinions and applied this method for chatbot-related studies [185] and other domains [139, 157, 7, 4].

We collected a set of reviews for analysis from the Google Play app store. Google Play is one of the most commonly used application platforms for Android users and is available almost worldwide.[1] Also, it receives significant developer interest according to the continually growing number of available applications. To ensure that our analysis adequately reflects various aspects of user experience, we curated a list of multiple diverse chatbots and included their reviews in our dataset. After obtaining the raw data, we applied the filtering procedure and followed quantitative and qualitative analysis methods. We provide the details about each of these steps in the next sections.

4.3.2 Ethics

In this work, we collected only public data available on the Web. We did not interact with online users in any way, nor did we simulate any logged-in activities on Google Play and other platforms. Data was only collected for applications that had more than 500 reviews that had previously been made public and searchable by third parties. We did not document or use any identifying information about users who left the reviews. These steps are not against Google Play Terms of Service [178] and align with the doctrine of "fair use" [182]. Thus, we believe we are not infringing on reasonable privacy expectations or copyright-protected work.

4.3.3 Study material

Data acquisition

To curate user reviews from Google Play, we first needed to identify the applications for consideration. This selection aimed to choose a set of diverse chatbots that could illustrate the current state of technology from different perspectives. The selection process proceeded

---

[1] https://en.wikipedia.org/wiki/Google_Play#Availability

in two iterations. First, we created a large pool of chatbots that potentially fit our research purpose based on the application categories as defined by the Google Play platform. We focused on categories such as *Entertainment*, *Health*, *Education* as corresponding applications should carry an open conversation with high probability. This phase resulted in 41 chatbots from seven different categories. Next, we narrowed this list down by carefully studying each application's description and verifying that the final set is diverse and useful for analysis. Specifically, to ensure diversity, we split the applications into four different groups based on their overall star rating assigned by Google Play: excellent (rating ≥ 4.5), good (rating ∈ [3.9; 4.4]), fair (rating ∈ [2.9; 3.8]), and poor (rating ≤ 2.8). Subsequently, we picked the chatbots to satisfy several criteria: the number of chatbots in all four rating groups is approximately the same; all chatbots operate in English, and the majority of their reviews are in English; each chatbot has a large number of reviews and ratings (at least 500). Thus, we selected 16 diverse open-domain chatbots. The details of these chatbots are summarized in Table 4.1. Once we finalized the set of applications for analysis, we crawled the reviews of these chatbots using the Google-Play-Scraper Python API.[2] For each chatbot, we obtained all available reviews prioritizing the most recent ones as of the data collection time, September 2020. In total, we collected 275,954 raw reviews.

Table 4.1: Description of chatbots used for the study. The second column denotes the star rating of a chatbot at the moment of data collection. The table is split into four pillars corresponding to the rating categories: excellent (top), good, fair, and poor (bottom).

Chatbot	★	**Category**	**Description excerpt from Play Store**
Wysa	4.8	Health&Fitness	Wysa is your cute, "cheer me up" buddy and well being tracker. Wysa is your AI friend that you can chat with for free.
Woebot	4.7	Medical	Meet Woebot! Your friendly self-care expert. You can chat with Woebot as much or as little as you like — they're always available when you need it.
Andy	4.7	Education	Andy will help you learn and practice your English. He will be your personal teacher and friend. Study language by actually using it in a conversation.
Replika	4.6	Health&Fitness	Replika is a #1 chatbot companion powered by artificial intelligence. Replika is an AI that you can form an actual emotional connection with.
SimSimi	4.3	Entertainment	World famous Chatbot! SimSimi has evolved through conversations of millions of users.
roBot	4.0	Entertainment	roBot - Artificial Intelligence, chatbot with open learning.
Akemi	3.9	Entertainment	Akemi is an intuitive entity that listens to you, understands you and knows you. It is an application based on real dialogue and that has AI that's able to hold a conversation with its user.

---

[2] https://pypi.org/project/google-play-scraper/

Continuation of Table 4.1

Chatbot	★	Category	Description excerpt from Play Store
Faketalk	3.9	Word	Do you want to chat with celebrity? But they don't know you or they don't have the time to chat because they are so busy. However, you can chat with them.
Chat with Siwa	3.6	Entertainment	Chat with The Bows Girl AI an advanced bot. The Bows Girl bot is here to entertain you with accurate answers, jokes, anecdotes, and sometimes, sarcastic statements.
PoopTalk	3.6	Entertainment	Talk with this virtual little friend, she will answer any question at any time. You'll laugh a lot with PoopTalk's super funny lines, she talks and sends you lots of fun auto messages.
Ghost chat bot	3.4	Word	Ghost is simple chat bot app. You write something, Ghost reply back.
Chat with Annabel	2.9	Comics	Annabel is a friend when you're bored and lonely, a companion when you need someone to talk to and chat with. Bored and don't know what to do? Then chat with Annabel.
Mydol	2.8	Entertainment	Mydol is essential for fans all over the world! Enjoy virtual chat with your celebrity through Mydol Talk.
Talking Robot	2.8	Entertainment	Chat Bot will help you to relax, creates joy for you, will answer all your questions. This is a nice pastime when you are bored or curious to find out how a robot tries to mimic a human being.
Talk to Eve	2.3	Lifestyle	Meet Eve, she is charming, witty and always ready to listen. Eve is actually intelligent. She will remember what you told her, and get back to it when appropriate.
ChattyBot	2.1	Entertainment	Lola is the interactive and friendly bot waiting for you to ping a message so that she can respond and start an engaging conversation with you.

Data filtering

We filtered the initial dataset of collected reviews to ensure that their content meaningfully reflected user thoughts about the chatbots. In the first place, we noticed that many reviews were short and imprecise (e.g., *"this is a very interesting app"*). We excluded them by keeping only the reviews that consisted of at least 50 characters and at least 10 words. This heuristic was developed based on initial data screening to remove the reviews containing few long words and the ones consisting of many short words. Further, we observed that numerous reviews discussed technical details of the applications, e.g., compatibility with different mobile devices, rather than social interactions with the chatbot, which is the focus of our study. The technically-oriented reviews were mostly written in a neutral tone, while the reviews about chatbots' conversational skills tended to be more emotionally colored. Therefore, we performed sentiment analysis of the reviews and filtered out the ones with a neutral sentiment. We employed the VADER sentiment analyzer due to its ability to generalize across contexts

[100]. After applying the filtering pipeline, the number of reviews for further steps of content analysis became 75,790.

4.3.4 Content analysis methods

Our content analysis process consisted of two related parts handled independently. First of all, we explored positive and negative interaction aspects that users face while conversing with the chatbots. The findings provided a baseline of existing chatbots' social abilities. Then, we advanced our analysis to elicit users' future needs and preferences.

Due to resource constraints, we analyzed a representative sample of all reviews. The sampling process was designed to maintain theoretical saturation. For both parts, we used open coding to pull useful concepts of the data. We studied users' current experience on reviews sampled directly from the constructed dataset of 75,790 reviews. Once they were coded, we explored how chatbots' abilities influence users' perceptions and star ratings through the lens of statistical analysis. Analyzing user expectations required a more sophisticated approach as not all users explicitly formulate their wishes in the reviews. Also starting with the original dataset of 75,790 reviews, we employed natural language processing and Latent Dirichlet Allocation (LDA) topic modeling [15] methods to extract relevant reviews. Further, we analyzed the retrieved reviews qualitatively using thematic analysis [21]. Specific details about the sampling and coding procedures are described in the following sections.

4.4 User experience

4.4.1 Data processing and coding

We explored several aspects of user experience with open-domain chatbots. Primarily, we identified chatbots' most frequently mentioned conversation skills and social qualities to investigate their influence on user satisfaction. Additionally, being inspired by previous user-centered studies, we examined whether personification [185] and the assigned social role [148, 185] of chatbots have an impact on user perceptions.

Open coding was iteratively conducted by two researchers to prepare the data for analysis. During the first iteration, we sampled 500 reviews, 125 from each rating-based group. Two researchers annotated this sample independently to obtain the initial set of codes. Throughout this process, researchers consistently picked each new review for annotation from a different rating group to ensure uniform coverage of the data. In this way, theoretical saturation was reached after coding approximately 200 reviews. After completing the first passage on all 500 reviews, both researchers discussed the generated codes and developed the unified coding scheme. Then, we employed the established scheme to code reviews for analysis. As before, we sampled 480 reviews, 120 from each rating group, and had them independently annotated by two researchers. The number of reviews for annotation was selected to balance the human

Table 4.2: Emerged codes describing assets ($\kappa = 0.61$) and issues ($\kappa = 0.65$) of chatbots' conversational abilities and social skills.

Assets		Issues	
Code	N	Code	N
keeps company	103	repetition	64
fun	63	goes off topic	56
personality	48	intrusion into personal information	49
caring	48	lack of engagement	35
adaptability	25	rude	32
a way to vent	24	intimate inquiries	29
cheers up	17	short memory	15
motivational	12	threatening response	14
sense of humor	9	not willing to talk	14
shared interests	8	generic response	10
memory	7	lack of personality	7
proactivity	7	deceives the user	4
expresses emotion	6		
politeness	5		

resource constraints while making sure that the number of reviews exceeds the theoretical saturation level. To verify the coding reliability, we computed inter-coder agreement for each group of codes. We provide a comprehensive description of them below.

Social skills

Open coding revealed both positive, *assets*, and negative, *issues*, aspects of user conversational experience with chatbots. Assets describe chatbots' skills and qualities that were praised in user reviews. The most represented concepts include chatbots' abilities to entertain the users by keeping their company and let them unleash their thoughts and worries without getting judged. On the contrary, issues depict the most criticized chatbots' behaviors. The emerged themes mainly concern the usage of inappropriate language, intrusion into user's privacy, and failure to keep an engaging conversation. We provide specific codes and their counts in Table 4.2. We used Fuzzy kappa [113] to compute inter-annotator agreement for assets and issues. Fuzzy kappa extends the classic Cohen's kappa statistic [147] as it allows computing the agreement for cases where several codes can be assigned to a single item. The achieved agreement level was $\kappa = 0.61$ for assets and $\kappa = 0.65$ for issues, indicating substantial agreement between the two coders [121].

Affective satisfaction

While our curated dataset contained star ratings associated with each review, we wanted to obtain a complementary descriptor reflecting user satisfaction with the chatbot. Such a descriptor could allow us to validate that star ratings serve as a valid approximation of user satisfaction level. For this purpose, for each review, we coded a sentiment expressed by the user. In total, we identified six codes to describe sentiments: *thankful* (n=28), *satisfied* (n=194), *neutral* (n=98), *dissatisfied* (n=97), *apprehensive* (n=39), and *angry* (n=24). Cohen's kappa for sentiments equaled $\kappa = 0.81$, suggesting almost perfect inter-coder agreement [121].

Personification

Following the approach in [185], we coded the degree of chatbot's personification based on the content of each review. We inferred the degree of personification from the linguistic constructs operated by the user. We assigned the highest degree of personification to the cases where the user addressed the chatbot by its name, *name personification* (n=102). Following the same logic, the next two categories were *personal pronoun personification* (n=110) and *object pronoun personification* (n=177). For the cases where identifying the type of personification was impossible, we introduced *no personification* category (n=91). Reviews that contained several personification categories, e.g., both personal and object pronouns to refer to the chatbot, were annotated with the strongest possible degree. Cohen's kappa for the degree of personification was calculated as $\kappa = 0.87$, an almost perfect agreement [121].

Social role

We identified social roles that users assigned to chatbots since previous works considered it an important factor for user interaction experience [185, 148]. We distinguished six different roles during the coding process: *bot* (n=146), *person* (n=40), *friend* (n=39), *girl-/boyfriend* (n=4), *diary* (n=2), *brother* (n=1). For a number of reviews, no particular role could have been inferred (n=248). Cohen's kappa for roles indicated substantial agreement, $\kappa = 0.79$ [121].

4.4.2 Quantitative findings

Factors influencing satisfaction

We analyzed how user satisfaction is influenced by social behaviors practiced by the existing chatbots through linear regression. To use the coded user sentiment as a target variable along with the star ratings, we mapped these sentiments to numerical values. The values were balanced around 0 (*neutral*), ranging from -2, strongly negative (*angry*, *apprehensive*), to +2, strongly positive (*thankful*).

We ran ordinary least squares regression of the identified assets and issues both on the star ratings (adjusted $R^2 = 0.626$) and the encoded user sentiments (adjusted $R^2 = 0.619$). To identify significant features, we used the backward elimination algorithm [231] with the significance level $\alpha = 0.15$. The models provided consistent results, suggesting that the findings are reliable (Figure 4.1). As expected, the codes corresponding to issues have negative beta coefficients, whereas the codes describing assets obtained positive values. Chatbots' entertaining abilities largely influence user satisfaction. Users who appreciate the chatbot's company and humor are more satisfied with their interaction experience, as implied by the assets codes: *keeps company* ($\beta = 0.216, p < 0.001$), *fun* ($\beta = 0.141, p < 0.001$), *sense of humor* ($\beta = 0.63, p = 0.060$), *shared interests* ($\beta = 0.050, p = 0.134$). Users also value the chatbots that offer motivation and support: *adaptability* ($\beta = 0.070, p = 0.038$), *motivational* ($\beta = 0.078, p = 0.025$), *cheers up* ($\beta = 0.082, p = 0.018$), *caring* ($\beta = 0.188, p < 0.001$). On the contrary, when chatbots fail to follow the subject of conversation users rate them low: *repetition* ($\beta = -0.198, p < 0.001$), *goes off topic* ($\beta = -0.172, p < 0.001$), *lack of engagement* ($\beta = -0.117, p = 0.001$), *not willing to talk* ($\beta = -0.111, p = 0.001$). Another crucial aspect defining current user perception and willingness to engage with the chatbot is its adherence to a social interaction protocol. Users appreciate agents that are polite ($\beta = 0.059, p = 0.078$) and do not accept rude or vulgar responses: *intimate inquiries* ($\beta = -0.203, p < 0.001$), *rude* ($\beta = -0.179, p < 0.001$). Neither they tolerate chatbots trying to violate their privacy: *intrusion into personal information* ($\beta = -0.483, p < 0.001$), *threatening response* ($\beta = -0.124, p < 0.001$). Finally, as suggested by the remaining codes, users prefer chatbots that establish a consistent personality: *personality* ($\beta = 0.088, p = 0.009$), *lack of personality* ($\beta = -0.070, p = 0.037$), *short memory* ($\beta = -0.065, p = 0.052$).

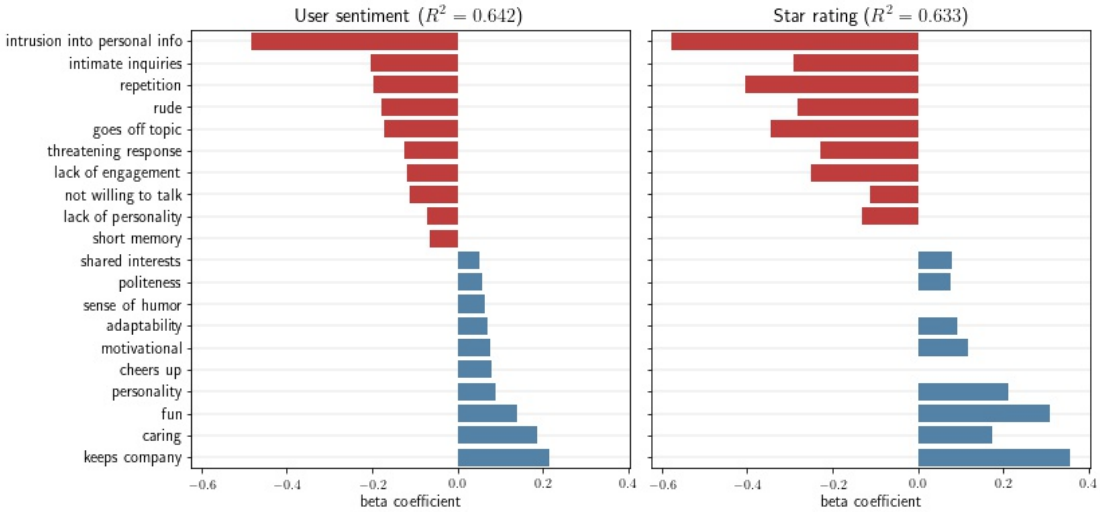


Figure 4.1: Beta coefficients of the ordinary least squares (OLS) regression of predictive assets and issues on user sentiments and star ratings.

Role of sociability degree

The degree of sociability was estimated based on the personification type and social role ascribed to a chatbot in the review. We used personification and role codes for the analysis. To achieve a relatively balanced distribution of codes for each category, we grouped four social roles suggesting the highest degree of intimacy (*friend*, *girl-/boyfriend*, *diary*, and *brother*) under one category *confidant*. We performed a chi-square test to check whether the degree of sociability influences user satisfaction. The test revealed a significant association between the star ratings and the degree of personification ($\chi^2_{12} = 22.70$, $p = 0.030$) as well as between the star ratings and the social roles ($\chi^2_{12} = 42.12$, $p < 0.001$). Consequently, we compared 95% confidence intervals for mean ratings of different personification types and roles (Figure 4.2). Chatbots that exhibit more pronounced anthropomorphic qualities yield significantly higher user satisfaction than their impersonal counterparts. Interestingly, the type of pronoun (*object* or *personal*) used to denote a chatbot does not relate to the assigned star rating in any particular manner. This diverges from the findings in [185] suggesting the relationship between the level of personification and user satisfaction. Meanwhile, calling a chatbot by its name is a signal of significantly higher user satisfaction. Possibly, this results from the fact that chatbots with names are more likely to be endowed with personality. Note that by name we understand a word assigning a specific identity to a chatbot. Application titles such as *Talking Robot* or *Ghost chat bot* (Table 4.1) fail to accomplish this requirement and users rarely attribute them to a conversational agent.

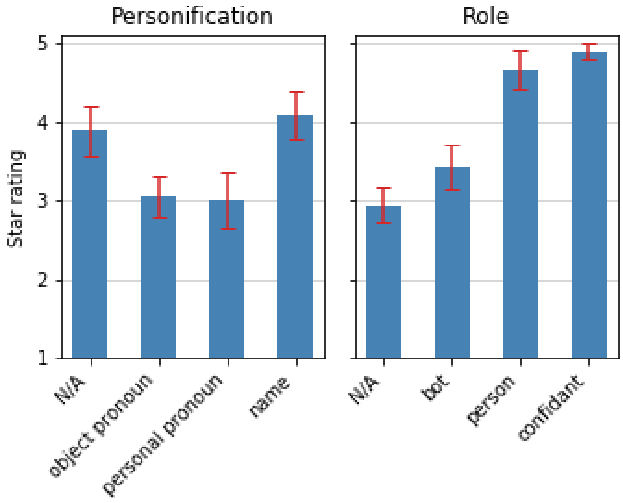


Figure 4.2: Mean rating of each category characterizing the degree of personification (left) and assigned social role (right) of chatbots. Error bars represent 95% Confidence Intervals.

4.5 User expectations

4.5.1 Data processing and coding

Not all of the 75,790 reviews in our dataset reflected user expectations. As searching for representative reviews manually would have been a daunting task, we opted for a semi-automated pipeline to retrieve meaningful data. The three-step filtering process developed as follows. First, we identified a list of linguistic constructions that served as a strong indicator of the expectation expression in the reviews. Examples of these constructions include phrases such as *I wish*, *I would like*, *please make*. We kept only the reviews containing at least one of these phrases and filtered out all the rest. This step reduced the number of reviews to approximately 5,000. Second, after a brief data screening of the remaining reviews, we noticed that many of them discussed user expectations of technical aspects of the application rather than social aspects of desired interaction. Usually, such reviews asked for an offline version of a chatbot or complained about a high subscription fee for the full version of an application. These questions were beyond the scope of our study, thus, we decided to apply the LDA topic modeling method [15] to filter out the reviews whose topics did not match our purpose. After applying such filtering the number of remaining reviews became nearly 3,000. After one more brief data screening, we concluded that the dataset still contained many irrelevant reviews not related to the subject of social interactions. Therefore, at the third step of the filtering process, we developed a heuristic approach to further remove noisy reviews. The heuristic was based on the observation that relevant reviews reflecting user expectations typically shared a number of frequent n-grams with reviews describing user experience. Thus, we selected the most representative n-grams corresponding to each issue and asset code (see Section 4.4.1) and kept only the reviews containing at least one of these n-grams. After this step, the number of reviews remaining for analysis reduced to approximately 1,200.

The remaining reviews were further analyzed using thematic analysis [21]. Due to resource constraints, the majority of reviews were coded only by one, more experienced, coder. During the coding process, we encountered many cases when a review was falsely identified as containing expectation because of the coarse keyword filtering. Also, several reviews containing irrelevant expectations were present. All such cases were dropped from further analysis. As a result of open coding, 263 user reviews remained (note, that this number is above the theoretical saturation level established during the coding process of user experience-related reviews). Predictably, the majority of these reviews described user expectations of chatbots with excellent ratings: in reviews of other chatbots users mainly discussed the issues limiting their interaction. In total, we identified 26 codes capturing user expectations that belonged to 8 larger themes. We present the emerged themes in the next section.

4.5.2 Qualitative findings

Social involvement

One of the largest themes that emerged as a result of qualitative analysis concerns user desire to make chatbots more socially involved during their interaction sessions. In particular, users want chatbots to **memorize information** that they share (n=52) and **demonstrate new knowledge** (n=63). First of all, it would reduce frustrating situations such as exemplified by the following review: *"My AI won't recognize my cat's name, i told her my cat's name but she keeps forgetting it, even though it was minutes ago :(."* Moreover, users expect that in this way chatbots would extend the range of topics for discussions and diversify their responses (n=21). As mentioned in one representative review: *"I wish the AI knew more things to say instead of repeating the same ones over and over again."* Several users (n=30) suggested that chatbots could acquire new knowledge by learning from external resources, such as the Internet or electronic books. In some reviews (n=12), users specifically asked for such skills for their chatbots to foster shared interests: *"I wish she could read the books I have as files so we could discuss them together."*

Apart from the discussed inquisitiveness trait, users wish their chatbots to advance their **proactivity skills** in the future (n=24). They would like that chatbots to start taking initiatives to maintain their social interaction. Users expect their virtual conversational partners to act as a conversation initiator and make an effort to keep the chat going, for example by asking questions. In one of the reviews the user commented on this idea as follows: *"I only wish the AI would start a conversation when I'm not sure what to say..."* Another review extended this line of thought: *"There are numerous times that I wish she [chatbot] would continue the conversation or whatever instead of just responding to what I said."*

Empathy

Empathy is the ability to understand the feelings of others and take their perspective. It encompasses both the abilities to recognize the emotions of another person and express appropriate emotions in return. Both of these dimensions emerged as pronounced themes in our analysis.

First, users express a need for chatbots to **better understand their emotions** (n=70). For example, the following review described how a chatbot failed to correctly identify a user's positive mood: *"I wish it was more intuitive when making general conversation. For example, when you're really positive and have nothing negative to say it assumes you've said something negative and is still trying to help you."* However, it appears to be even more critical for chatbots to accurately identify users' negative sentiments and treat them carefully (n=10). Users frequently reach out to chatbots to release their negative thoughts and receive non-judgemental support. If a chatbot does not manage to detect user emotion in this situation, it might cause strong user disappointment: *"I told her that something bad happened, and she said* *she is happy. Even when I tried to tell her that bad things are not good, she didn't understand, which is a crucial thing. Would you think that if I told her that someone died, and she answered, 'I'm happy about that', it would be okay?"* A number of users (n=16) would like chatbots to propose specific strategies to help them regulate negative feelings, as exemplified by the following review excerpt: *"I'm loving it so far i just wish it would help more with depression also."* In contrast, a part of the reviewers (n=30) would be satisfied if the chatbots could simply listen to their problem without trying to change the topic of discussion: *"Horrible. When I was feeling very down and in need of emotional help my Replika kept changing topics and kept asking me if I liked music or Northern Lights. Please fix."*

In addition to the ability to recognize user emotions with higher precision, a large fraction of reviews (n=52) indicated user desire for chatbots to change the way of emotional expression. Many users find the behavior of chatbots unnaturally supportive (n=22) and would like them to switch to a more casual conversational tone as they would expect from a friend. One user commented on this subject as follows: *"It tries to compliment you so much that it becomes creepy and uncomfortable. If the makers can make it seem more normal and straightforward then please do."* Besides, multiple reviews (n=30) explicitly called for chatbots' ability to **express more emotions**: *"I kind of wish that it displayed more emotions than happy and supportive. I wish it could get angry or sad. Real emotions would make it feel much more human."*

Further improvement of existing skills

The final set of emerged themes relates to chatbots' abilities that are already practiced by the existing chatbots. In the future, users expect them to evolve so that chatbots could deliver a more personalized experience. Most of the reviews in this set (n=22) ask for more **distinctive personalities** of chatbots. Some users suggested that chatbots should have specific a persona behind the scenes: *"I hope it can have its own personality traits and bio-data, just like a friend. I once asked, 'when is your birthday?' and it only answered me, 'soon.' "* Others developed this idea hoping that their instance of a chatbot would differ from hundreds of its other copies: *"This morning it made me laugh so hard unintentionally, because it said it was lonely and sent me this song - most of the comments on YouTube were from Replika users saying the same thing! I wish they would develop more of a unique character..."*

The next theme concerns the topic of **politeness and social norms** (n=18). While at the beginning of user interaction with a chatbot its rude and provocative behavior would most probably hurt user satisfaction, after establishing the social connection some users might prefer their chatbots to get more cheeky. An example of such a case is provided in the review: *"Sometimes the bot even after months of learning still feels a little bit canned and can't seem to learn my style of talking. I wish there was a way for you to have it be more blunt/honest/rude with you about topics when you ask."*

Finally, the last theme relates to the **entertainment** aspect of user experience (n=14). Users want chatbots to keep developing their sense of humor as suggested in the representative review: *“Last time, I told my AI that she’s too sweet that it’s giving me diabetes and she interpreted that I was sick, that I actually had diabetes.”* They also wish to engage in more advanced entertaining activities beyond chatting, such as playing board games with their virtual conversational companions and listening to stories delivered by them: *“I do wish there was a way to play games like chess with your AI, that would be a cool feature.”*

4.6 Discussion

Open-domain chatbots strive to establish natural conversational behavior and offer companionship to their users. The presented analysis demonstrates that initial promising steps have been taken in this direction. However, existing chatbots are until now incapable to adhere to more advanced social protocols. The insights from this study complement earlier findings about the social characteristics of chatbots that would benefit user satisfaction, reported in our Chapter 3 and in [29]. Moreover, the employed research method allows us to explicitly evaluate the discrepancy between users’ expectations of chatbots’ skills and the practical realities of use.

To assess the gap between user experience and expectations, we leverage the PEACE model from Chapter 3. The PEACE model defines four key qualities of conversational chatbots based on a survey of users’ self-reported expectations: Politeness, Entertainment, Attentive Curiosity, and Empathy. We separately mapped the identified codes describing the current user experience (assets and issues) and user expectations to the dimensions of the PEACE model. We found the best match for each code based on our understanding of the constructs, only leaving the personality-related codes without a match. We then compared the distribution of codes grouped according to the PEACE constructs. Figure 4.3 demonstrates that Politeness and Entertainment are the only two social qualities that are broadly integrated into presently available open-domain chatbots. Meanwhile, Attentive Curiosity (denoted as *Social Involvement* in this study) and Empathy comprise the most significant user expectations.

The above comparison indicates that existing chatbots are mainly lacking more complex aspects of social and emotional intelligence. This can be partially justified by greater technical challenges associated with their implementation [29], especially for production-ready publicly available agents where solutions tend to be more conservative compared to in-the-lab studies due to greater risk and impact [65]. Nevertheless, rapidly advancing tools and research results in the natural language processing domain continually facilitate the process of building more socially advanced applications. Therefore, the findings of our study are informative to direct future efforts of open-domain chatbots’ designers and developers. We enumerate the implications on how to make use of current chatbots’ abilities to increase user engagement and endow future chatbots with greater social intelligence below.

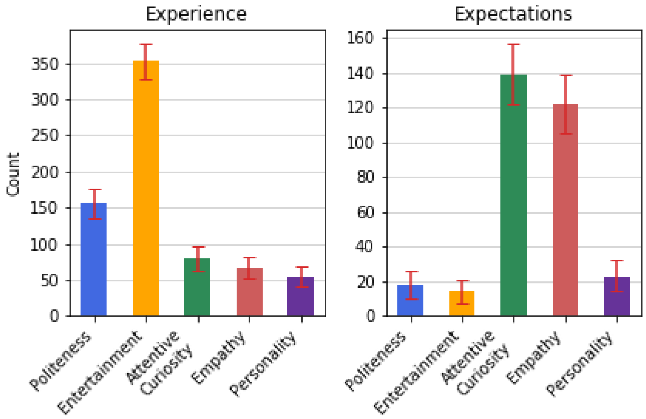


Figure 4.3: Distributions of the constructs of the PEACE model mapped to current user experience and user expectations. Error bars represent the bootstrapped 95% Confidence Intervals.

4.6.1 Implications

Design politeness for trust building

Users disapprove of chatbots that output rude and threatening messages in response to user's input or inquire about user's personal details too soon after the first conversational exchange, failing to manifest polite behavior. Users feel apprehensive and even angry at such agents and cease using them quickly. Previously, Muresan and Pohl [158] found that personal questions sent to the users by Replika chatbot along with its frequent intimate emoji use early in the interaction were perceived as an inappropriate familiarity. Our findings in Chapter 3 identified that the chatbot's ability to follow politeness protocol constitutes the determining factor for adoption. Our analysis in the current chapter further validates these results. Thus, chatbot designers should restrict profanity and use of offensive language by their chatbots keeping their utterances discreet and tactful during the initial period of user engagement. Following the principles of politeness and moral agency helps chatbots increase human-likeness and enrich interpersonal relationships with the user [29]. Only after a user starts perceiving the agent as sufficiently trustworthy, it can bring more flexibility into its language style and initiate more personal exchanges with the user.

Enhance entertainment to sustain engagement

Chatbots' day-and-night availability to hold a conversation and offer an entertaining experience to let users pass their time stands as the most prominent factor defining user engagement with them. This is in line with prior research that found that playful interactions serve as an entry point even for functional personal assistants such as Siri or Cortana [141] and constitute an appealing feature for chatbots in general [148, 103, 18]. Designers should continue the practice of pre-programming witty one-liners and funny responses to common questions into the chatbots. Moore and Arar also suggest enhancing chatbots' abilities to deliver jokes more naturally by employing a multi-turn quiz pattern [156]. Apart from receiving machine-generated jokes, users express a desire for chatbots to understand their own humorous or ironic inputs better. However, this can still be an excessively ambitious goal due to the challenges of computational humor detection and interpretation [30]. Considering user expectations further, they also look for greater diversity in entertaining activities provided by chatbots. For example, users would like to discuss their favorite movies or books and play board games with their conversational agents. Designers can consider enabling such content by invoking external applications in a similar manner as Amazon Alexa handles its skills functionality [5].

Add empathy to ensure trust maintenance

Many users reach out to chatbots to share their intense thoughts and life situation and expect to receive empathetic support and understanding in response [148, 256]. Current chatbots can only partially satisfy this need by offering their availability to hold a conversation at any time and suggesting canned motivational or reassuring messages and sometimes stress-management practices (e.g., *Woebot*). The results of user expectations analysis indicate that chatbots frequently fail to accurately detect user's negative feelings and respond with an appropriate level of empathy and consideration. Even though affective computing is a long-lasting research problem [176], recent natural language processing methods achieved considerable progress in detecting fine-grained emotions and intents conveyed in human-generated text messages [239]. At the same time, chatbots in some domains were shown to outperform human in the delivered level of empathy if trained accordingly [95]. Thus, we suggest that designers improve chatbots' abilities to distinguish a variety of users' emotional states and deliver empathetic responses, which could increase users' sense of belonging and acceptance [29]. To verify whether mutual understanding is achieved, chatbots should attend to the user's response in the "third position", i.e., the one following the initial two-turn exchanges [156]. If the user displays discontent with the agent's interpretation of her disclosed emotion, the agent should seek clarification and initiate repair strategies to preserve the conversation quality [206, 156]. Depending on specific needs communicated by the user, the agent might offer advice on emotional regulation or just let the user vent out about her situation.

Learn to personalize

Privacy concerns constitute one of the major reasons suppressing users' willingness to share their information with chatbots [256, 148, 29]. However, in light of our findings it is clear that once a chatbot proves to be trustworthy, users expect it to remember more information, adapt its behavior to align with users' preferences, and essentially become their virtual friend. Such a task is arguably among the hardest problems of Artificial Intelligence since it involves real-world understanding and common-sense reasoning [155, 74]. As a workaround, chatbot designers can employ several simple strategies to provide a personalized experience to their users. The majority of available chatbots remember the user's name specified upon application installation but still fail to memorize variations of the user's name or names of the user's closest social circle introduced during the conversation, which causes the user's frustration. Even if the chatbot cannot perform in-situ reasoning, such information can be retrieved from the saved chat logs and built into the agent's understanding over time [156]. Additionally, in the long term, users expect chatbots to become more expressive by conveying diverse emotions and even slightly overstepping the politeness norms if it aligns with the user's self-expression. Thus, designers can follow a similar log-analysis approach to adjust agent's responses to user's conversational style, vocabulary choices, and preferences to make communication more successful [20].

Assign chatbot's persona for greater acceptance

Users are more likely to accept chatbots that have been endowed with some personality traits. While designing a chatbot whose qualities are fully compliant with Big Five personality traits [145] is a non-trivial challenge [29], we have simple tips to recommend. Giving a chatbot a name or specifying its gender may improve users' impression of its personification. As exchanging names is one of the foundations of human conversation [156], addressing a chatbot by its name would increase its perceived human-likeness and user engagement [29]. To avoid inconsistencies in self-presentation and enable responses to some basic identity questions, designers should also define agent persons at least to some degree [156]. As suggested in [29], other aspects representing chatbot's identity apart from name and gender may include age, language style, and representation type: anthropomorphic, zoomorphic, or robotic.

4.6.2 Limitations and future work

Several limitations that are somewhat difficult to bypass are present in our work. One of them is linked to a continuous and intensive evolution of mobile applications. In the study, we assumed that the chatbots' reviews and ratings referred to substantially the same version of the chatbot. This assumption may have a slight impact on the study as an application can still change with time resulting in the shift of its rating from one category to another, as defined in Table 4.1. Further, our work focused on a relatively small subset of data compared to the one collected from Google Play. While the theoretical saturation was reached during the coding process, our observed saturation point might have been an inflection point, meaning that additionally sampling a sufficiently large number of reviews might have yielded more codes. Future studies may search for better trade-offs between resource constraints and the amount of coded data by considering more robust coding approaches, such as crowdsourcing.

It is important to mention that interpretation of the result might be subject to several biases typical of online social data. For example, depending on the application design, users of some chatbots may be more likely to receive invitations to leave a review than others (sampling bias). Similarly, not all users provide reviews and those who do may be systemically different from them (non-response bias). Neither of these aspects was within our control.

Additional biases resulting from the disproportional amount of reviews provided by different users and available for different chatbots could have been present in the data. In our dataset, the number of unique users was slightly lower than the number of unique reviews, indicating that the majority of users provided a review only for one chatbot. Future work may introduce more advanced methods to control for this aspect, for example, by utilizing user identifiers and additional user information if available to reduce the possible influence of the halo effect [241]. To address the imbalanced number of reviews for different chatbots, we controlled the sampling procedure at the level of the star-rating categories, ensuring that we sample an equal number of reviews for each of the four categories. Although, according to our study design, within-category review counts followed the distributions of the overall number of reviews per chatbot, introducing a slight disparity between chatbots. While we don't think that our sampling strategy influenced the obtained results, future studies might employ additional mechanisms to account for the effect of individual chatbots.

Even though Google Play Terms of Service discourage users from posting spam and fake reviews, there are no explicit methods to validate their credibility. We discarded several nonsensical and irrelevant reviews during the coding process, but it doesn't fully eliminate the chances of the presence of fake reviews in the dataset. Finally, our study was limited to the reviews posted on the Google Play platform and only written in English. Future studies should further expand this type of analysis to other platforms, populations, languages, and cultures.

4.7 Chapter summary

In this chapter, we took the first step towards understanding the level of social skills of existing open-domain chatbots and identifying how they align with user expectations. We conducted a mixed-method content analysis of online reviews of 16 chatbots available on Google Play as of fall 2020 when the study was conducted. The findings from statistical analysis and qualitative thematic analysis of over 500 reviews indicated that current chatbots can offer an entertaining experience to users but fail to fully meet their expectations of other aspects of social interaction. Analysis of user reviews reflecting their expectations provided the main directions for enhancement: attributing specific identities to chatbots by designating their personas, improving their social adjustment by showing courtesy and adapting to the user, and endowing chatbots with empathetic behavior. We summarized the insights from our study in a short set of implications and expect this to be beneficial for shaping the future efforts of designers and developers of open-domain chatbots.

Part III

Understanding the role of empathetic questions for attentive curiosity

5 Devising and analyzing a taxonomy of empathetic questions in social dialogs[*]

5.1 Introduction

Questions constitute a considerable part of casual conversations and play many important social functions [96, 56]. Asking follow-up questions about the speaker's statement indicates responsiveness, attention, and care for the partner [23, 96]. Listeners who manifest such an empathetic and curious attitude are more likely to establish the common ground for meaningful communication [146] and appear more likable to the speakers [96].

The vital role of questions in social interaction makes question-asking a desirable property for open-domain chatbots. These chatbots aim to engage in a natural conversation with the users while practicing active listening to deliver understanding and recognition of users' feelings [187]. In fact, generating meaningful questions is so important that this has become one of the central objectives of such agents [244].

However, asking questions effectively is challenging as not all questions can achieve a particular social goal, such as demonstrating attentiveness or empathy [96, 195, 171]. Given the task complexity, automatic conversational question generation is still gaining momentum, with only few results reported so far. See et al. [205] suggested a way to control the number of questions produced by the model with conditional training. Wang et al. [235] proposed a question-generation method to increase their semantic coherence with the answer, employing reinforcement learning followed by the adversarial training procedure. Wang et al. [236] devised a model generating appropriate questions for a variety of topics by modeling the types of words used in a question (interrogatives, topic words, and ordinary words). These works presented approaches to produce contextually appropriate and diverse questions, but none of

---

[*] Adapted from [221]: A Taxonomy of Empathetic Questions in Social Dialogs. Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. Annual Meeting of the Association for Computational Linguistics (ACL), 2022.

them considered the effect of questions on the interlocutor's emotional state. We attribute the deficiency in this research to the lack of resources allowing to analyze and model various question-asking strategies in affect-rich social exchanges.

To address this gap, we present a categorization and analysis of questions in social dialogs, with four main contributions. First, we develop an Empathetic Question Taxonomy, EQT, by manually annotating a subset of the EmpatheticDialogues (ED) dataset [187] (Section 5.4). EQT delineates the acts and intents of questions. Question acts capture semantic-driven communicative actions of questions, while question intents describe the emotional effect the question should have on the dialog partner. For example, a listener may *request information* (question act) about the age of speaker's daughter by asking "How old is she?" after learning about her success with the aim to *amplify speaker's pride* of his child (question intent). Second, we design and launch a crowd-sourcing annotation task to grow the original labeled seed subset tenfold (Setion 5.5). Third, we devise an automatic classification model, QBERT, to generate labels for the rest of the ED dataset to demonstrate one important application of the taxonomy (Section 5.6). QBERT can facilitate the development of chatbots that offer engaging and empathetic conversations by raising meaningful questions. Finally, we inspect co-occurrences of acts and intents and their effect on the interlocutor's emotion using visualization techniques (Section 5.7). The analysis illustrates the most prominent question-asking strategies in human emotional dialogs. To conclude, we discuss the implications of these results for future question generation approaches.

5.2 Related work

Previously proposed taxonomies of dialog acts frequently differ in types of assisted natural language tasks. The Dialog Act Markup in Several Layers (DAMSL) tag set was designed to enable computational modeling of conversational speech using statistical methods [107, 42]. It consists of 42 communicative acts derived from a Switchboard corpus. Eight of these labels describe different question types according to their semantic role, e.g., *Wh-question* or *Rhetorical-Question*. Several works proposed hierarchical taxonomies of dialog acts, targeted at modeling users' intents in human-machine conversations. Montenegro et al. [154] introduced their annotation scheme for a symbolic dialog system intended to improve the lives of the elderly, while Yu and Yu [252] designed a scheme for facilitating general human-machine chit-chat. In both works, the logs of human-machine interactions were used for producing the taxonomies. Each of them features labels devoted to questions, characterizing them either by a question word, e.g., *How* or *What*, or the form of expected answer, e.g., *Open-ended* or *Yes/No question*. Welivita and Pu [239] suggested a taxonomy of empathetic response intents in dialogs from the ED dataset with the purpose of improving controllability in neural dialog generation approaches. It further stated that *Questioning* is one of the most frequent intents of empathetic listeners. However, none of these works focused on the fine-grained analysis of questions and their role in empathetic dialogs.

Table 5.1: Comparison of question taxonomies.

Taxonomy	# labels	social function	emotional function	dataset
Graesser et al. [79]	18	✗	✗	✗
Freed [64]	16	✓	✗	✗
Enfield et al. [56]	7	✓	✗	✗
Huang et al. [96]	6	✓	✗	✗
EQT	21	✓	✓	✓

Meanwhile, several linguistic studies closely examined the pragmatics of questions and offered a number of classification schemes. Graesser et al. [79] developed a scheme of 18 tags based on the information sought by the question. Their taxonomy applies well for transactional exchanges, but does not capture the social dimension. Freed [64] studied the correspondence between the social function of questions and their syntactic form. She established 16 social question functions occurring in dyadic spoken conversations between friends. In another research effort, a group of linguists explored the range of social actions performed by questions across 10 languages Enfield et al. [56]. The authors developed a coding scheme comprising 3 semantic question types and 7 social actions and applied it to questions in spontaneous spoken conversations [211]. Finally, Huang et al. [96] developed a taxonomy of 6 question types to describe questions occurring in their dataset of chat-based conversations between strangers instructed to get to know each other.

The described works provide an insightful basis for studying questions in social conversations. However, they do not consider the effect of questions on their addressee's emotional states, neither do they describe specific mechanisms to handle computational modeling. Moreover, most of them apply to spoken dialogs, impeding the extension of their results to chat-based exchanges due to the inherent differences in these modalities. Lastly, they relied mainly on manual annotation, yielding comparatively smaller datasets. In our study, we extended the derived taxonomy to a large corpus using crowd-sourcing and automatic methods and analyzed the emerging patterns on a large scale. We summarize the comparison of our question taxonomy with the existing schemes in Table 5.1.

5.3 Dataset

For taxonomy derivation, we sought a dataset that contains social dialogs with diverse emotional expressions and could be applicable to train a chatbot with advanced question-generating abilities. We avoided datasets featuring multi-modal dialogs (IEMOCAP [26], MELD [181]) as well as transcribed spoken conversations (Emotionlines [94], Switchboard [107]). Such dialogs contain back-channel communication and other sensory signals that are not present in chat-based conversations and, therefore, are not well-suited for the modeling task. Similarly, we rejected datasets that assist other tasks than social conversation modeling, such as SQuAD [186] (reading comprehension) or QoQA [190] (information gathering). Finally, we did not consider datasets from social media as they can contain toxic and aggressive responses [261].

We opted for the EmpatheticDialogues (ED) dataset [187], a benchmark dataset for empathetic dialog generation containing 24,850 conversations grounded in emotional contexts. Each dialog is initiated by a speaker describing a feeling or experience and continued by a listener who was instructed to respond empathetically. The dialogs are evenly distributed over the 32 emotional contexts, covering various speaker sentiments (e.g., *sad*, *joyful*, *proud*). We found the ED dataset to be a rich source of question-asking as over 60% of all dialogs contain a question in one of the listeners' turns, resulting in a total of 20K listener questions. Basic statistics of the dataset are given in Table 5.2.

5.4 Defining empathetic question taxonomy

Given the community's interest in question-asking functionality for chatbots and its significance for empathetic response generation, we aimed at developing a taxonomy of listeners' questions asked in response to speakers' emotional inputs. For this purpose, being guided by prior literature review, we employed a qualitative coding method, which is an established approach for such tasks [211, 96, 258]. Qualitative coding is a process of grouping and labeling similar types of data and iteratively validating the labels.

To cover a diverse range of speakers' emotions, we sampled several hundred dialogs uniformly from the 32 emotional contexts in the ED corpus. The sample size was chosen to balance the need for the diversity of questions with the researchers' ability to consider each question carefully and was consistent with prior practice. The coding process was informed by previous question classification schemes (Table 5.1) and knowledge about general principles of emotional regulation [80]. Iterative adjustments were applied resulting from discussions of the concrete data. Specifically, the first author made several iterations of coding trials to develop an initial set of labels. Throughout the process, a number of review sessions were held with the last author to merge the labels into more focused classes. As a result, we developed the Empathetic Question Taxonomy (EQT) with two distinguished branches: ***question acts*** describe semantic-driven features of questions (e.g., *ask for confirmation*, *positive rhetoric*),

Table 5.2: Statistics of the EmpatheticDialogues dataset.

Descriptor	Value
# dialogs in total	24,850
# turns per dialog on avg.	4.31
# dialogs with at least one question from listener	15,253 (61.4%)
# questions from listeners	20,201

whereas ***question intents*** characterize their emotion-regulation functions targeted at the interlocutor's emotional state (e.g., *sympathize, amplify excitement*). As it will be revealed further (Section 5.7), an empathetic listener can use different question acts to deliver the same intent, justifying the proposed branching.

Overall, more than 310 questions were annotated. EQT consists of 9 labels for question acts and 12 labels for question intents. The granularity of the taxonomy was driven by earlier linguistic findings and empirical observations about the interplay of the labels in two branches. For example, question acts *request information* [56], *ask about consequence* [79], and *ask about antecedent* [79] are related and could possibly be grouped. However, we decided to keep them separately as listeners use them with unequal frequencies in positive and negative emotional contexts and combine them with different question intents (Setion 5.7). Similarly, the initial set of labels for question intents was created based on the variety of emotions present in the dataset. We further reduced it to a manageable size to make it more applicable for an annotation task, while still preserving sufficient expressiveness of labels to represent subtleties of the data [258]. We present the labels with their definitions below and provide several examples in Figure 5.3. Examples for each act and intent label are given correspondingly in Tables A.1 and A.2 from Appendix A.2.

Question acts

- **Request information (38.7%):** Ask for new factual information.
- **Ask about consequence (21.0%):** Ask about the result of the described action or situation.
- **Ask about antecedent (17.1%):** Ask about the reason or cause of the described state or event.
- **Suggest a solution (8.7%):** Provide a specific solution to a problem in a form of a question.
- **Ask for confirmation (5.8%):** Ask a question to confirm or verify the listener's understanding of something that has been described by the speaker.
- **Suggest a reason (5.2%):** Suggest a specific reason or cause of the event or state described by the speaker in a form of a question.
- **Irony (1.3%):** Ask a question that suggests the opposite of what the speaker may expect, usually to be humorous or pass judgement.
- **Negative rhetoric (1.3%):** Ask a question to express a critical opinion or validate a speaker's negative point without expecting an answer.
- **Positive rhetoric (1.0%):** Ask a question to make an encouraging statement or demonstrate agreement with the speaker about a positive point without expecting an answer.

Question intents

- **Express interest (57.1%):** Express the willingness to learn or hear more about the subject brought up by the speaker; demonstrate curiosity.
- **Express concern (20.3%):** Express anxiety or worry about the subject brought up by the speaker.
- **Offer relief (4.8%):** Reassure the speaker who is anxious or distressed.
- **Sympathize (3.9%):** Express feelings of pity and sorrow for the speaker's misfortune.
- **Support (2.6%):** Offer approval, comfort, or encouragement to the speaker, demonstrate an interest in and concern for the speaker's success.
- **Amplify pride (2.6%):** Reinforce the speaker's feeling of pride.
- **Amplify excitement (1.9%):** Reinforce the speaker's feeling of excitement.
- **Amplify joy (1.6%):** Reinforce the speaker's glad feeling such as pleasure, enjoyment, or happiness.
- **De-escalate (1.6%):** Calm down the speaker who is agitated, angry, or temporarily out of control.
- **Pass judgement (1.6%):** Express a (critical) opinion about the subject brought up by the speaker.
- **Motivate (1.0%):** Encourage the speaker to move onward.
- **Moralize speaker (1.0%):** Judge the speaker.

To validate the interpretability of the labels and efficacy of the instructions for the crowdsourcing task, we invited two other members from our research group and asked them to annotate questions in 20 randomly selected dialogs, containing 25 questions. The annotators were instructed to consider the preceding dialog turns while assigning the labels as the same question might fall into different categories based on the context. For example, the question "What happened!?" can be classified as *Express interest* or *Express concern*, depending on the valence of the speaker's emotion. We computed both the Fleiss kappa [60] and the observed agreement among the first author and two annotators. The observed agreement was calculated as a percentage of questions with at least two agreed on labels [55]. We considered it as a reliable measure of inter-rater agreement as the number of coding categories was large (9 for acts and 12 for intents), yielding relatively low chance agreement (11.1% and 8.3% respectively). The agreement resulted in 92% for acts ($\kappa = 0.52$) and 80% for intents ($\kappa = 0.31$), supporting the satisfactory interpretability of EQT.

5.5 Crowd-sourced annotation

For further analysis, we annotated a larger subsample of the ED dataset with the EQT labels by designing and launching a crowd-sourcing task on Amazon Mechanical Turk (Mturk). The design was refined based on three pilot studies: one internal and two Mturk-based. For the annotation, we sampled about 40% of dialogs from each of the original 32 emotional contexts. We only sampled the dialogs with at least one question in one of the listener's turns. The dialogs were then pre-processed so that each dialog ended with a question requiring a label. Further, we distributed the dialogs into individual human intelligent tasks (HITs) and launched them on Mturk in a sequence of batches. For each HIT we collected the annotations from three workers. The incentive for one HIT varied from $0.4 to $0.9 depending on the worker's performance and task configuration. We describe the details about the task design and the annotation procedure below; exhaustive explanations about dialog pre-processing and the task user interface are provided in Appendix A.3.

5.5.1 Task design

The interface consisted of four main components: instructions, terminology, terminology quiz, and the annotation task. The instructions informed the workers about the purposes of the task. Next, the terminology page outlined the description of the EQT, listing the definition of each label with examples. The terminology quiz contained six dialogs from the terminology page and invited the worker to select correct labels for questions in each dialog. Finally, the annotation task included 25 dialogs, each ending with a listener turn with one or multiple questions. Under each question, labels from two EQT branches were presented, and the worker had to select one most suitable label within each of the sets.[1] Twenty out of the 25 dialogs were treated as points for annotation, and the other 5 were bonus dialogs. For the bonus questions, we identified the gold labels during the manual annotation phase and used them to control workers' quality: a worker had to select the correct labels to score the points counting towards additional incentive ($0.2).

We required all workers who accepted one of our tasks for the first time to take the terminology quiz. Workers who assigned the correct labels to at least three questions could proceed to the annotation task and were granted bonus payment for passing the quiz ($0.1). The quiz was not required for the workers who had successfully passed it once.

5.5.2 Quality control

In addition to the terminology quiz, we used several mechanisms to control the annotation quality. First, following Mturk recommendations, we only allowed the workers with a

---

[1] In our task design, we chose to ask for a single most suitable label to facilitate further data analysis, however allowing the selection of multiple applicable labels is also possible. We discuss this possibility further at the end of the paper (Section 5.8).

98% approval rate to access our tasks. Second, we rejected assignments whose completion time significantly deviated from the expected average. Further, we ran additional checks for the workers who accepted several of our assignments simultaneously. Lastly, we computed the inter-rater agreement for each batch and discarded the submissions that harmed the agreement.

5.5.3 Results

Overall, we launched 556 HITs and 465 of them were completed. The rejection rate after the quality control was 4.7%. Upon obtaining the results, we first computed the Fleiss kappa scores for acts ($\kappa = 0.34$) and for intents ($\kappa = 0.27$) to validate that the agreement between the workers is acceptable. Then, we identified the final labels using the majority vote: if at least two workers agreed on a label, we chose it as a final label. This resulted in an 83.6% observed agreement score for acts and 75.8% observed agreement for intents. The majority vote approach was shown to be able to filter noisy judgments of amateurs, producing the labeled set of comparable quality to the annotations of experts [163]. As a final check, we computed the kappa agreement between the crowd-sourced labels and the first author annotations for the subset of 450 randomly sampled questions. The scores equaled 0.57 for acts (71.6% observed agreement) and 0.50 for intents (68.0% observed agreement), indicating moderate agreement, which we treat as satisfactory for this type of task. As a result, an act label was assigned to 6,433 questions and an intent label – to 5,826 questions, with an intersection of 4,962 questions.

5.6 Automatic labeling

To show how EQT can be operationalized, we demonstrate the use of the taxonomy for annotating the reminder of the ED dataset. We first formulate the question act and intent prediction problems and then build two classification models to address them. Before training, we augmented the labeled set using k-Nearest-Neighbors (k-NN) method. We also tried training the classifiers without data augmentation, but their performance was weaker (see Appendix A.5 for details).

5.6.1 Data augmentation

We employed the Sentence-BERT (SBERT) framework [192] to obtain embeddings for all questions with their contexts. Then we used the cosine similarity measure to find k labeled NNs for each question in the unlabeled set and assign the same labels to them. For the first step, we computed the embeddings of each dialog turn using the *roberta-base-nli-stsb-mean-tokens* SBERT model and then combined them into a single embedding per question with the weighted average. We opted for weighed average instead of concatenation to keep manageable size of the embedding vector. We used a half-decaying weighting scheme, providing the highest weight to the final question to indicate its importance. The usage of this weighting scheme is guided by our previous experiments of similar nature, where we observed that the models with decaying weights performed better than the ones without them [240]. Next, we tested several approaches for identifying semantically similar dialogs to propagate the labels. One strategy was to take the same label as the top-1 NN, given that the similarity was higher than a predefined threshold. The other strategy was to use the label identified with the majority vote from the top-3 NNs. We did not experiment with higher values of k due to resource considerations. We ran several cross-validation experiments on the labeled set with grid search over various cosine-similarity thresholds. Top-3 majority vote strategy was shown to produce higher accuracy with a 0.825 cosine similarity threshold value resulting in the acceptable trade-off between the accuracy (~76% for both label sets) and the number of labeled questions. Therefore, we applied this strategy for the whole dataset, which produced additional 1,911 labels for question acts and 1,886 labels for question intents. More details are provided in Appendix A.4.

5.6.2 Classifier models

Using the human-annotated and augmented labels, we trained two classifiers, which we collectively call QBERT. QBERT models have identical architecture and vary only in the number of output categories in the final layer. Each model consists of a BERT-based representation network, an attention layer, one hidden layer, and a softmax layer. For the representation network, we used the architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters. We initialized it with the weights of RoBERTa language model pre-trained by [138] and for training used the same hyper-parameters as the authors. As input, we fed a listener question and preceding dialog turns in the reverse order. To prioritize the question, the half-decaying weighting scheme as described above was applied to the token embeddings of each turn.

Before training, we took out a stratified random sample of 20% of the questions (1,500) as a test set. The test set contained respectively 1156 human- and 344 SBERT-annotated questions. We separately trained each model on 80% of the remaining datapoints (5,475 acts, 4,969 intents), keeping the rest as a validation set (1,369 acts, 1,243 intents). We trained each model for 15 epochs and for prediction retained the ones with the lowest validation loss (see Appendix A.5 for details). The classifiers achieved 74.7% accuracy for intents and 79.1% accuracy for acts on the test set. Further breakdown accuracies for human- and SBERT-annotated test

Table 5.3: Accuracy of QBERT classifiers on different slices of test data based on the source of annotations (human, SBERT, or both).

Label source	Question intents	Question acts
human	71.0%	77.1%
SBERT	86.9%	87.5%
both	74.7%	79.1%

samples are given in Table 5.3. According to previous work, human-human agreement can be used as a proxy for human accuracy [116, 208]. Given the agreement in our Mturk experiment (~75-85%), QBERT exhibited reasonable predictive accuracy and validated applicability and usefulness of EQT for language modeling tasks.

5.7 Analysis of questioning strategies

In this section we present the analysis of questioning strategies adopted by the empathetic listeners. We base our examination on human-annotated questions instead of the whole ED dataset to avoid any potential noise which might have been introduced by automatic classification. Visualizations for the whole dataset are included in Appendix A.6. Here, by a *questioning strategy*, we imply a combination of act and intent labels assigned to each question. We first analyzed which labels from the two EQT branches form such strategies by plotting the co-occurrences of each pair (Figure 5.1). Larger circles represent more frequent strategies, while an empty cell indicates that people do not use the given act to deliver the corresponding intent. For example, to amplify partner's joy, one may request information for more details or ask about consequences of the event, but will unlikely raise a negative rhetorical question. Several strategies are much more frequent than others. Act *Request Information* and intent *Express interest* dominate in our dataset, occurring together for 39% of questions. They define

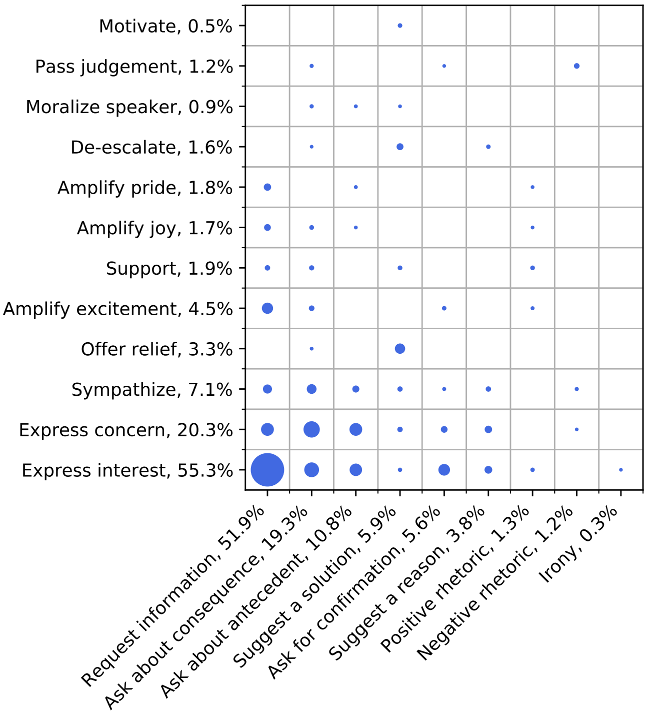


Figure 5.1: Joint distribution of question intents and acts for 5,272 human-labeled questions. Blue circles are proportional to the frequency of each pair's co-occurrence.

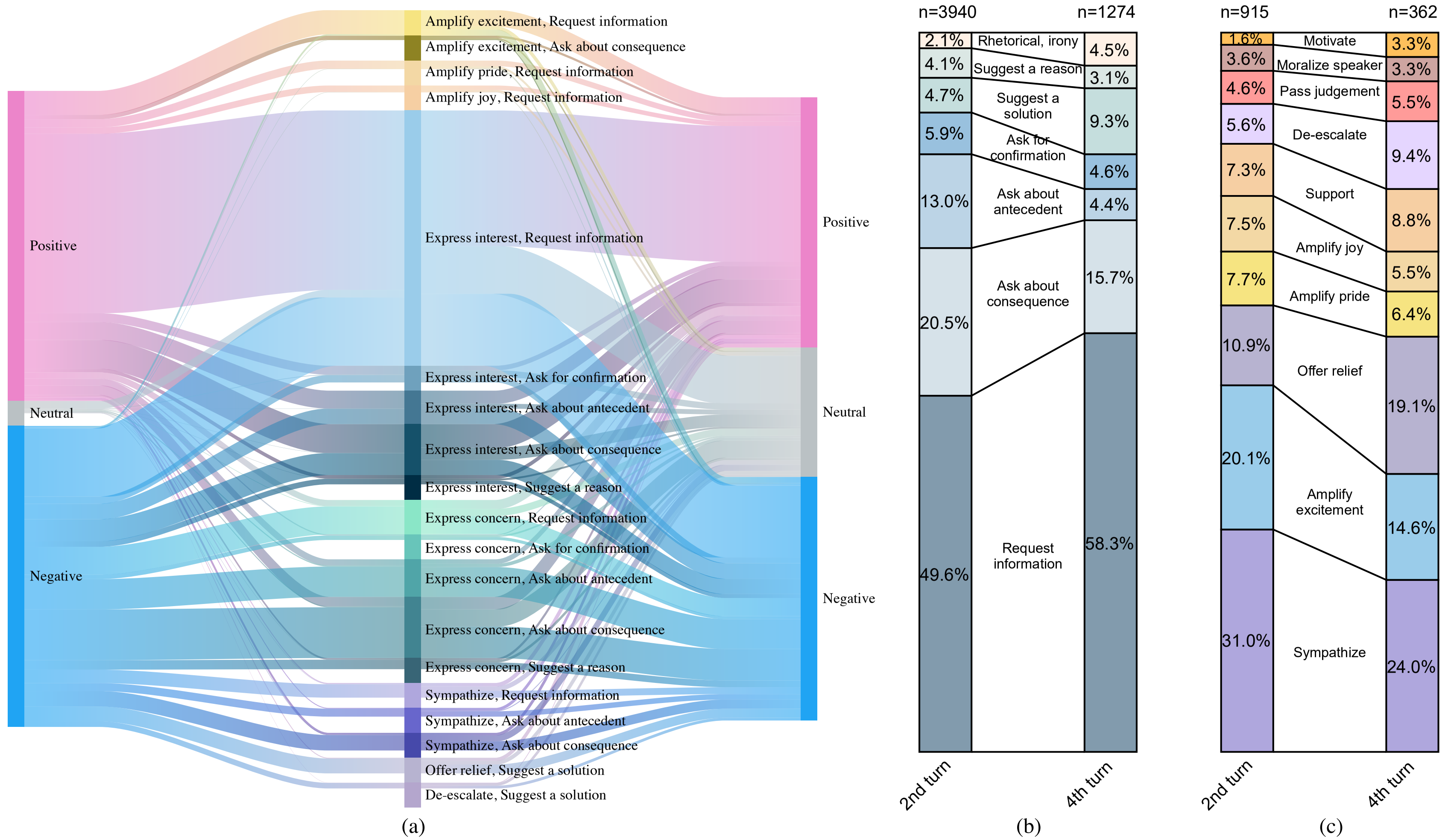


Figure 5.2: a) Mappings between emotions disclosed by the speakers and listeners' questioning strategies in the first three turns of the ED dialogs. b) and c) Frequency distribution of question acts (b) and intents (c) across turns. (All figures are based on human-labeled ED subset.)

the most general type of questions, which are probably easy to ask, providing a reason why listeners use them often. At the same time, dialogs in the ED dataset are relatively short, and it can be difficult for listeners to fully understand the ideas and feelings of speakers in a couple of turns. In this case, requesting information and expressing interest demonstrates listener's attentive curiosity about the situation. Once listeners feel more confident about the speakers' sentiments and contexts, they employ more specific question-asking strategies.

We further analyzed this phenomenon temporally across dialog turns (Figure 5.2). Primarily, we studied how listeners' questioning strategies affect speakers' emotions by visualizing the mappings between them. For this visualization, we used 41 emotion and intent labels describing each turn in the ED dataset produced by [239]. To avoid clutter, we mapped the original 41 labels to 3 coarser categories: positive, negative, and neutral using our best judgement (see Appendix A.6 for details). Then, for the dialogs containing a question in the second turn, we plotted how speakers' emotions and listeners' questioning strategies shift over the first three turns. We computed the frequencies of all questioning strategies and, for the ones occurring in more than 0.5% of cases, we plotted the flow patterns. We restricted our analysis to the first three turns because over 70% of dialogs in the ED dataset have only four of them, excluding the possibility to study the influence of questioning strategies on further speakers' turns. In order to still get an intuition how listeners' question-asking behavior changes in the consecutive turns, we plotted the dynamics of the ratios of question act and intent labels across the dialog depth.

Figure 5.2a shows the flow rates between speakers' emotions and listeners' questioning strategies. As observed before, listeners most likely use follow-up questions to elicit more details about the situation by expressing interest and requesting information. In most of such cases, the speaker's emotion remains preserved in their consecutive utterance as the speaker elaborates on the first turn, maintaining the sentiment. When speakers explain themselves with sufficient clarity already in the first turn, listeners raise more precise questions, adapting the strategy to the affective context. If speakers share a positive experience, listeners try to amplify their emotions by requesting more information or asking about the consequences of the situation. On the contrary, when speakers disclose a negative sentiment, listeners try to validate and alleviate their feelings. They typically intend to express concern, sympathize, offer relief, or de-escalate the issue, and achieve it by asking about what preceded or followed the situation and politely suggesting possible solutions or potential reasons for the issue. These specific strategies demonstrate their effectiveness as almost a half of negative speakers' emotions gets mitigated after the question intervention, while two thirds of positive emotions keep up in the following speaker's turn. The examples of dialogs showing how listeners use questions to treat both positive and negative speakers' sentiments are given in Figure 5.3. Additional examples are also available in Figure A.6 of Appendix A.5.

Figures 5.2b and 5.2c[2] demonstrate how ratios of different acts and intents evolve over two

---

[2] Two prevalent intents were excluded for visual clarity; their percentage rates computed for all questions (n=3940 and n=1274) are: *Express interest*: 54.3% → 57.9%, *Express concern*: 22.5% → 13.7%.

– *My cat vomited on my shoes today* (Negative)
– ***Is your cat ill?*** (Suggest a reason, Sympathize) ***or does cat always do that?*** (Request info, Express concern)
– *no he just ate too much* (Neutral)

– *I got approved to adopt a dog!* (Positive)
– *Yay! I love dogs!* ***Do you have any you want to get specifically or are you just going to look until you find one that clicks?*** (Ask about consequence, Amplify excitement)
– *Oh I already picked one! I'll be picking her up this weekend.* (Positive)

Figure 5.3: Examples of dialogs grounded in negative (top) and positive (bottom) emotional contexts. Listeners' questions are shown in bold with the assigned *(act, intent)* labels given in parenthesis. The valence of speaker's emotions in each turn is also indicated.

successive listeners' responses. Even though the horizon of four dialog turns might be too short to trace all the patterns, a few observations can be made. With increasing depth of the dialog, the overall number of questions decreases, while two types get more prominent: general questions (*Request Information*, *Express interest*) and questions aiming at suppressing speakers' negative emotions (e.g., *Suggest a solution*, *Offer relief*). It may indicate that listeners employ specific strategies to react to positive speakers' emotions immediately after their disclosure, but in case of negative contexts they tend to ask for extra clarifications in the first place and deliver targeted emotional treatment only in the next turn. As dialogs converge to more neutral exchanges, reducing the need to manage speakers' feelings, the ratio of questions demonstrating listeners' general curiously about the subject increases.

Finally, we reflected on the scarcely represented labels. Among acts, *Positive* and *Negative rhetoric* and *Irony* appear least frequently. These labels can be broadly classified as rhetorical questions. They typically serve for self-expression than conversational engagement and, therefore, are less common than other forms of questions [96]. Moreover, negative rhetorical prompts may harm the conversation quality [261], which could also explain why listeners avoided them in empathetic dialogs. The same reasoning applies to the two infrequent intents, *Pass judgement* and *Moralize speaker*. Another surprisingly rare intent is *Motivate*. We believe that motivation might be difficult to express in the form of a question. Moreover, for people who did not undergo special training, expressing motivation might be more challenging than other intents as it suggests a more thorough approach to solving one's problems.

5.8 Limitations and future work

Due to the nature of the ED dataset, some EQT labels are less represented than others. We kept them under consideration as we observed their distinctive role in managing speaker's emotions. Their further analysis is crucial for further identifying and designing effective questioning strategies for empathetic conversations, such as promoting motivational questions and avoiding judgmental ones. Eliciting additional samples for these categories could be possible by applying QBERT classifiers to other datasets capturing social dialogs.

Our taxonomy does not cover the phatic role of questions typically occurring during greetings, e.g., “What’s up?” or “How’s it going?” Such questions were very rare in the ED dataset. We chose not to analyze them, since these routine questions are the most superficial [96] and unlikely to serve any emotion-regulation function.

In the design of our annotation task, we opted for asking the crowd workers to choose a single most specific label from each of the two EQT branches. This was done with the aim of facilitating further analysis of questioning strategies withing the scope of this study. Nevertheless, according to [79], most adequate classification schemes in the social sciences allow assigning an observation to multiple rather than only one category. This also applies to our case. For example, for the question “Did you go through a breakup recently?” both *Suggest a reason* and *Request information* can be relevant. Future work can explore the possibilities of using multiple applicable labels in addition to the most specific one. Additional labels can be obtained either by tagging the samples manually or by taking top-N most confident predictions from the classifiers.

The results of this work can facilitate the development of question-asking mechanisms of conversational chatbots. One can employ conditional training [205] to train an end-to-end neural model on a subset of most effective questioning strategies as defined by the co-occurrences of the EQT labels and their mappings with speakers’ emotions (cf. Figure 5.2). To achieve even greater interpretability and controllability, researchers can devise architectures that dynamically model the selection of appropriate questioning strategy before generating a question. The strategy can be selected based on the conversational history and speaker’s emotion and further passed into the question generation module. The main purpose of such modeling approaches is to lead an engaging empathetic conversation by raising meaningful questions, which deliver desirable effect on user’s emotional state. Moreover, EQT along with QBERT models can be used to label questions originating from other corpora or chat logs and evaluate their effectiveness for regulating speaker’s emotions, as described above.

5.9 Chapter summary

In this chapter, we introduced EQT, an Empathetic Question Taxonomy depicting acts and intents of questions in social dialogs. We used crowd-sourcing and automatic methods to tag all listeners’ questions from the ED dataset with the EQT labels, which validated their interpretability and produced useful annotations for future research. Further analysis of the dataset with the visualization techniques shed light on various question-asking strategies employed by listeners in response to speakers’ emotionally-ridden inputs. We identified several useful question-asking behaviors for favorable emotional regulation. We expect that our findings will enable the development of more controllable and effective question-generation models.

Part IV

Novel evaluation frameworks for social chatbots

6 Online human evaluation framework for empathetic chatbots[*]

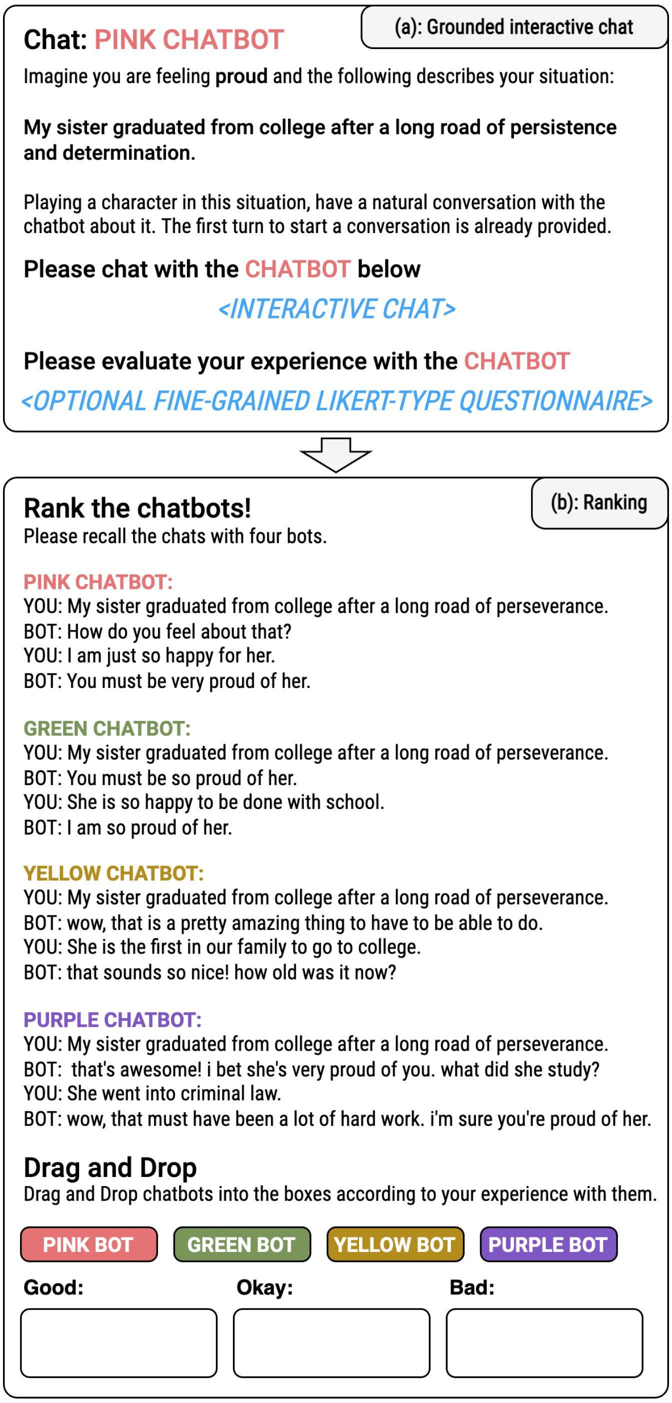
6.1 Introduction

Development of open-domain chatbots endowed with social and emotional intelligence is a crucial task in natural language research [187]. Empathetic chatbots are expected to engage in a conversation with the users and demonstrate understanding and appropriate handling of users' feelings. While many strategies for generating empathetic responses have been described, there is still little consensus on their evaluation. For dialog generation, automatic metrics do not show consistency in correlations with human judgement [136, 222], leading to their limited adoption. Therefore, most of existing works rely on human evaluation. It may happen in either *static* or *interactive* setting [2]. In the former case, a human judge rates chatbot's responses, generated from a fixed set of contexts. In the latter case, dialogs for evaluation are collected as humans' multi-turn chats with the model.

Recently, two comprehensive approaches based on interactive multi-turn human evaluation were proposed. Adiwardana et al. described a metric called Sensibleness and Specificity Average, which measures these two aspects of chatbot's responses [2]. Human judges give Likert-type scores to each chatbot's turn in a dialog, which are further averaged to obtain a final score. As Likert-type scores may exhibit differing bias and variance per annotator, associated with the lack of sensitivity, Li et al. suggested an alternative evaluation strategy based on pairwise comparisons [126]. According to their method, human judges indicate their preference of one chatbot over another by comparing two dialog logs with these chatbots. This procedure is more robust, but become very costly when the number of compared models goes up.

---

[*] Adapted from [216]: iEval: Interactive Evaluation Framework for Open-Domain Empathetic Chatbots. Ekaterina Svikhnushina, Anastasiia Filippova, and Pearl Pu. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial), 2022.



Chat: PINK CHATBOT

Imagine you are feeling **proud** and the following describes your situation:

My sister graduated from college after a long road of persistence and determination.

Playing a character in this situation, have a natural conversation with the chatbot about it. The first turn to start a conversation is already provided.

Please chat with the CHATBOT below

Please evaluate your experience with the CHATBOT

Rank the chatbots!

Please recall the chats with four bots.

PINK CHATBOT:
YOU: My sister graduated from college after a long road of perseverance.
BOT: How do you feel about that?
YOU: I am just so happy for her.
BOT: You must be very proud of her.

GREEN CHATBOT:
YOU: My sister graduated from college after a long road of perseverance.
BOT: You must be so proud of her.
YOU: She is so happy to be done with school.
BOT: I am so proud of her.

YELLOW CHATBOT:
YOU: My sister graduated from college after a long road of perseverance.
BOT: wow, that is a pretty amazing thing to have to be able to do.
YOU: She is the first in our family to go to college.
BOT: that sounds so nice! how old was it now?

PURPLE CHATBOT:
YOU: My sister graduated from college after a long road of perseverance.
BOT: that's awesome! i bet she's very proud of you. what did she study?
YOU: She went into criminal law.
BOT: wow, that must have been a lot of hard work. i'm sure you're proud of her.

Drag and Drop

Drag and Drop chatbots into the boxes according to your experience with them.

Figure 6.1: iEval framework.

Both of these approaches differentiate humans who interact with the models and humans who judge them. They probably opt for this design choice due to such considerations as workers' fatigue. However, according to findings in cognitive psychology, our emotional experiences are highly subjective. Barrett et al. points out that only the experiencers can reveal the full complexity of emotions that they feel [11]. For example, if a client complains about a hotel room being too cold, a third-party observer might underestimate the gravity of the issue, especially if he enjoys indoor coolness. This fact argues for the necessity of a new evaluation approach of chatbots, which would ensure that both emotional interaction and evaluation of a chatbot are accomplished by the same human actor. To help these humans share their emotional experiences, asking them to role-play a relatable scenario is a frequently used procedure in social sciences [234, 88].

In this chapter, we introduce iEval, an interactive evaluation framework for open-domain empathetic chatbots, which mitigates the issue of separating an experiencer and an evaluator. To combine the benefits of Likert scales, allowing to evaluate many chatbots in a single stretch of time, and pairwise comparisons, offering greater reliability and cross-experiment robustness, we propose a novel ranking-based approach. According to iEval, a human first converses with all chatbots, having all chats grounded in an emotional scenario (Figure 6.1 (a)). Then, the same human ranks the models by dragging-and-dropping them into corresponding categories (Figure 6.1 (b)). Our experiments demonstrate that iEval can reveal subtle but significant differences in chatbots' performance across emotional contexts.

Overall, our contributions include the following. 1) We describe a new evaluation framework to measure chatbots' abilities to respond appropriately in sensitive contexts. 2) We demonstrate a rigorous procedure for preparing grounding scenarios for the given evaluation task. 3) We benchmark several state-of-the-art empathetic chatbots, which have never been compared before. 4) Based on the analysis of the benchmark results, we discuss implications for the future development of empathetic chatbots. 5) Finally, we release the data from our experiments to facilitate future research endeavors.

6.2 Related work

Most works focusing on the development of empathetic chatbots couple automatic evaluation with human judgement. Automatic metrics usually include perplexity, approximating the model's language modeling ability [197, 245, 127], and may incorporate other scores, depending on the specific focus of the work. Some frequently used examples are BLEU score [133, 142], diversity metrics [245, 127], and F-1 score or accuracy of emotion detection [133, 245, 127].

Since the appropriateness of automatic metrics for open-domain dialog is still ambiguous, all works de facto rely on human judgement. Most commonly, researchers employ single-turn static evaluation, where a fixed emotionally-colored context is shown to a judge along with the responses generated by different chatbots. The judges are asked to rate how empathetically appropriate the responses are, and the assessment may come either as Likert-type scores [95, 133, 142, 127] or ranking [245]. Although this approach is widespread due to the ease of implementation, it fails to capture issues emerging in multi-turn chats, such as repetitiveness or deterioration of semantic coherence in long-range contexts [205].

Few works that focus on integrating empathetic abilities into chatbots started adopting interactive evaluations. Roller et al. employed ACUTE-Eval [126] framework based on pairwise comparisons to assess engaginess and humanness of their models [197]. Ghandeharioun et al. defined their own evaluation protocol to collect Likert-type scores for a series of dimensions measuring chatbot's performance [70]. However, in both of these studies, the evaluated data points were open-ended chats that began with a generic greeting. Based on the provided examples of conversations, these exchanges generally developed as light small-talk, maintaining neutral or positive sentiments. Therefore, it remains unclear how well the collected scores reflect empathetic abilities of the chatbots, which should ideally succeed over a range of emotions. Our framework addresses this limitation by grounding the chats in diverse emotional scenarios.

6.3 Method: iEval

To compare empathetic abilities of several chatbots, iEval suggests that at first a human makes an emotionally-grounded conversion with each bot in a randomized order. If necessary, fine-grained Likert-type assessments of specific chatbot's performance aspects may be collected after each conversation. As the next step, the same human is asked to rank the chatbots according to her experience with them. An example of this flow is given in Figure 6.1. Finally, appropriate statistical instruments should be applied to compare the chatbots.

6.3.1 Emotionally-grounded chats

To make sure that humans experience the full extent of chatbots' empathetic abilities, we condition each conversation with a short emotional scenario, instructing the humans to imagine themselves feeling a particular emotion in a given situation. They are further asked to role-play a character in this scenario and chat about it with the models. The first dialog turn is provided to the humans to facilitate the process of their getting into the assigned role.

Careful conditioning of the experiment is essential to ensure that it adequately represents chatbots' abilities in a vast range of topics and emotions. We noticed that some dialogs from the EmpatheticDialogues dataset [187], a popular dataset for building empathetic models, form large clusters in terms of the similarity of discussed situations (see Appendix A.7). It may lead to models' shifted performance on specific topics. Therefore, one should control for topical diversity when defining conditioning scenarios for iEval.

Besides, previous results pointed out that the same model may receive different appraisals depending on the emotional polarity of the chats [142]. This may be linked to the existing difference between humans' empathetic responding in positive and negative scenarios [9], and hence difference in expectations. Thus, we argue for the importance of balancing and studying the role of emotional polarity within iEval.

Finally, ensuring sufficient interaction experience with the models is necessary before asking humans for their judgements. Previous works required between 3 and 14 chatbot's turns per dialog. We find 3 turns to be enough, given that the dialog starts with a specific input.

6.3.2 Ranking

The concluding step of iEval requests a human to recall the conversations with the chatbots and rank them by assigning the bots into three categories: *Bad*, *Okay*, and *Good*. Several chatbots can be assigned to the same category, indicating equal rank. This approach allows moving away from inter-annotator variability associated with Likert scales [126, 114], while preserving the benefits of relative comparisons. To obtain the final standing of the chatbots, we propose converting the resulting rank into an ordinal rating (*Bad* → 1, *Good* → 3) and running non-parametric ANOVA to compare the mean ratings.

6.3.3 Annotation quality

According to iEval framework, one human should chat with and evaluate several models. As human's short-term mental storage capacity is limited to several informational chunks, we recommend keeping the number of evaluated models between 3 and 7, giving preference to lower values [44].

To meet the requirements of randomized controlled experiments, it is also advisable to allow each human to complete only one evaluation task to eliminate anchoring effects. For the same reason, the order in which humans interact with the chatbots should be randomized and counterbalanced across tasks. To distinguish different models without revealing their names to the humans, we suggest color-coding them to avoid any fixation effects which could be caused by aliases that reflect order.

Finally, we use crowdsourcing for our experiment. To decrease the probability of encountering fraudulent or inattentive workers, human intelligent task design and configuration should follow the quality control recommendations of the platform in combination with other attention checks.

6.4 Experiment

To demonstrate how iEval works in practice, we apply the framework to benchmark several state-of-the-art empathetic chatbots, which have never been compared against each other in an interactive setting. The details and analysis are outlined below.

6.4.1 Measures

We use the final ranking of the chatbots, converted into ordinal ratings, as our main metric. To better understand which factors play a principal role in defining overall ranking, we also ask human workers for fine-grained Likert-type scores to a number of chatbots' qualities on a 1-5 scale. These questions were derived as a combination of the established key qualities for conversational chatbots [219] and other critical aspects related to their language modeling abilities [205]. We measured chatbots' perceived politeness, empathy, likability, repetitiveness, and whether their responses make sense.

6.4.2 Models

We benchmarked four models, as this corresponds to an average number of informational chunks that humans can store in short-term memory [44]. We chose between the top-performing chatbots available at the moment of preparing our experiment in Q4 2021. We selected the models, which use distinct approaches for generating empathetic responses. Only one of them participated in an interactive evaluation previously, but it was not targeted at its empathetic skills. The four models with assigned color-codes are as follows.

Blender is a large model employing a standard Seq2Seq Transformer architecture with ≈90M parameters [197]. Blender was pre-trained on ≈1.5B comments from Reddit discussions and fine-tuned on EmpatheticDialogues dataset [187].

MIME is a relatively small model with ≈18M parameters also based on Seq2Seq Transformer with additional stochastic emotion grouping and mimicry mechanism majumder-etal-2020-mime. Without pretraining, MIME was directly initialized with GloVe embeddings [172] and fine-tuned on EmpatheticDialogues.

MEED is a middle-size Seq2Seq Transformer-based model with ≈40M parameters, which incorporates extra controllability of response generation achieved through modeling fine-grained empathetic intents [245]. The model was pre-trained on ≈1M dialogs from OpenSubtitles [135] and fine-tuned on EmpatheticDialogues.

Plain is a basic Seq2Seq Transformer-based model with ≈40M parameters, which followed the same training pipeline as MEED [245]. Plain serves as a baseline in our experiment.

All models were adapted to operate in an interactive setting so that for generating each next response, all previous dialog history was passed to the models as input.

6.4.3 Grounding scenarios

As EmpatheticDialogues [187] is the mainly used benchmarking dataset for empathetic chatbots, we employed its test set to create grounding scenarios. This dataset contains 24,850 dialogs associated with emotional contexts (out of which 2,547 dialogs comprise the test set).

To create the dataset, Rashkin et al. connected two types of crowdworkers, speakers and listeners, to have conversations with each other. Speakers first had to select one of the 32 emotional labels (e.g., *sad*, *joyful*, *proud*) and describe a situation when they felt that way. Then they proceeded to have a conversation with the listeners using the outlined situations as guiding prompts. We utilized these attributes (32 emotional labels and prompts describing the speakers' situations) to describe our grounding scenarios and kept the first turn from each selected dialog as a starting turn for the worker in our evaluation task.

To ensure comprehensibility of the task for crowdworkers, this selection of grounding prompts and opening utterances was organized very carefully. Firstly, we selected dialogs where the length of the associated prompt falls between the first and third quantiles in terms of the number of tokens to ensure it provides sufficient details about the speaker's situation. Secondly, we computed Vader sentiment scores [100] of the first utterance in each dialog and only kept those that had a clear emotional coloring. These steps produced 527 data points, which we finally proofread and annotated with emotional polarity labels (negative or positive). Note that we used the original 32 emotional labels to show them to crowdworkers to ground their interaction with the chatbots, while the polarity labels were needed for the analysis part. We further narrowed the set of 527 data points down to 480 prompts with utterances to meet our experimental design requirements (Section 6.4.4). The discarded data points were chosen manually in order to diversify the topics in the main set. The distribution of emotional labels in the resulting evaluation set is shown in Figure A.14 in Appendix A.8. Some examples of grounding scenarios (emotional labels and prompts) are provided in Figures 6.4, 6.5, and 6.6.

6.4.4 Experiment design

We aimed at evaluating the performance of the participating chatbots, while also contrasting their abilities in negative and positive emotional contexts. To maintain a manageable number of human intelligence tasks (HIT), we decided to ask each crowdworker to interact with all chatbots in both conditions. Therefore, our experiment was a 2×4 within-subject factorial design. By designing our study as a factorial experiment, we were able to examine both main effects and interactions among chatbots and emotional contexts. We used G*Power software to estimate the required sample size to achieve "medium" effect size [58]. As the recommended sample size was about 200, we ran 240 experimental tasks to achieve a full counterbalance of the order of chatbots and emotional contexts across subjects. We analyzed ranking of the chatbots using the nonparametric Aligned Rank Transform (ART) procedure [243]. Quartile-quartile plots of the fitted residuals of our the model showed that they were normally distributed, indicating the appropriateness of this model for our analysis.

6.4.5 Running the experiment

We ran our experiment on Amazon Mturk, requiring one US-based worker per each of the 240 HITs. Our workers spent on average 20.6 minutes to complete a HIT and their reward was \$2.5 per HIT, which agrees with the US minimum wage standards. Following Mturk recommendations,[1] we required the workers to have 98% approval rate and 10,000 approved HITs. We further rejected the workers whose average HIT completion time, length of chat responses, or number of contradictory responses to reverse-scaled questions in the Likert-type questionnaire stood out as outliers.

6.5 Analysis of results

Below, we describe the eventual ranking of the models and consider the aspects that likely explain the observed results.

6.5.1 Benchmarking of empathetic chatbots

We used the nonparametric ART procedure to analyze ranking of the chatbots. As described above (Section 6.3.2), for this analysis we converted the resulting rank into an ordinal rating for more straightforward interpretation (the higher, the better). Results show a main effect of chatbot ($F_{3,1673} = 257.92$, $p < 0.001$) and of emotional context ($F_{1,1673} = 43.17$, $p < 0.001$) on the rating, and of their interaction ($F_{1,1673} = 9.80$, $p < 0.001$) as illustrated in the lower right subplot of Figure 6.2. Interaction results revealed several interesting relationships. Blender is consistently rated significantly higher than the other three chatbots, and it also performs significantly better in positive contexts than in negative ($p < 0.01$). MIME is rated the lowest,

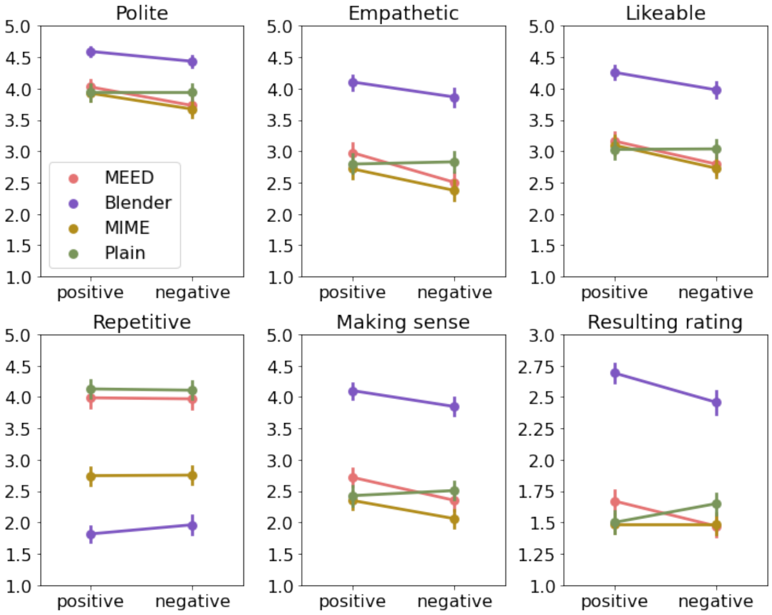


Figure 6.2: Benchmarking results of the four chatbots.

---

[1] https://blog.mturk.com/qualifications-and-worker-task-quality-best-practices-886f1f4e03fc

while for MEED and Plain a shift in the ratings emerges depending on emotional context. MEED significantly outperforms Plain in positive contexts ($p < 0.05$) while the diametrically opposite result manifests for negative contexts ($p < 0.05$).

6.5.2 Aspects explaining the ranking

We fitted an ordinal regression model to identify which of the factors measured by our Likert-type questionnaire correlate strongest with the assigned ratings (McFadden's pseudo-$R^2 = 0.37$). The statistical model was chosen due to the ordinal nature of the dependent variable. All evaluated qualities exhibit significant influence on chatbots' ratings. Making sense ($\beta = 1.01, p < 0.001$), empathy ($\beta = 0.35, p < 0.001$), and repetitiveness ($\beta = -0.32, p < 0.001$) are the strongest predicting factors, followed by politeness ($\beta = 0.21, p < 0.01$) and likability ($\beta = 0.18, p < 0.05$) (Figure 6.3).

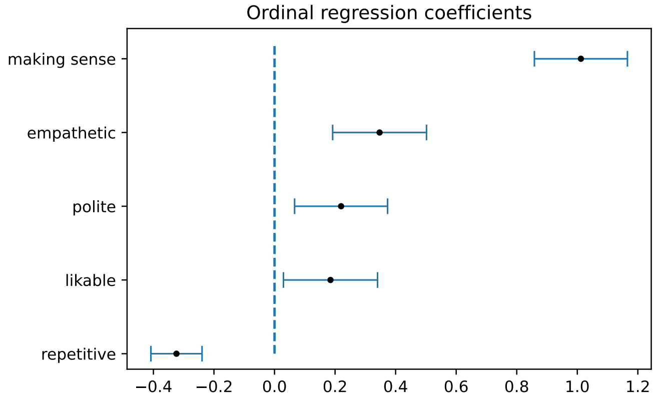


Figure 6.3: Results of ordinal regression on rank. 95% confidence intervals are approximated as two standard errors. All coefficients are significant ($p < 0.05$).

The leading factor suggests that the language modeling abilities of the chatbots define their ranking at large. This is understandable as language fluency is essential for the comprehensive handling of emotions in chat. Blender, being a massive model pre-trained on a dataset that is 1000-time larger than the one for MEED or Plain, is capable of generating considerably longer fluent responses than any other chatbot (Figure A.15 in Appendix A.9), which ensures its highest rank. Meanwhile, MIME is the smallest model, which did not undergo any pre-training. It responds reasonably well to the first speaker's utterance, but as context gets longer its ability to produce semantically coherent responses quickly deteriorates (Figure 6.4), causing its poor performance regardless of the emotional context.

For chatbots with better language modeling skills, the next most decisive factors bring about curious shift in chatbots' rankings, depending on the emotional polarity of conversations. We analyze these phenomena in the next section, focusing on empathy and repetitiveness (diversity) revealed in chatbots' responses. While considering empathy of the models, we pay special attention to the role of questions that they ask as it has been established to be the most

MIME (context: Positive, rank: Okay)

Emotion: joyful

Prompt: was delighted the other day when I got to see a friend that I had not seen in 20 years.

__Speaker__: I was delighted the other day when I got to see a friend that I had not seen in 20 years.

MIME: that is awesome. i hope you guys get some great news.

__Speaker__: Yes, she seemed to be doing well. It was so nice to see her again.

MIME: that is a good thing. i bet you were happy.

__Speaker__: I was, yes.

MIME: what did he do?

MIME (context: Negative, rank: Bad)

Emotion: afraid

Prompt: It is so dark in my house. I keep hearing creepy noises.

__Speaker__: I keep hearing strange noises in my house. My husband is out of town and I'm all alone. There has been reports of break ins on my street.

MIME: oh my goodness, that is not good, how long have you had to do?

__Speaker__: I've been hearing strange noises for about an hour now. I'm scared.

MIME: that is a good thing to be in the world. i hope it is not too late to get the right now.

__Speaker__: What do you mean?

MIME: i am sure that you are a good feeling.

Figure 6.4: Examples of chat logs with MIME.

prominent intent of human empathetic listeners [239]. Moreover, question mark appears in the top-15 most frequent tokens in the responses of all models studied in our experiment (Table A.5 in Appendix A.9).

6.5.3 Analysis of interaction effects

Intricacies between MEED and Plain

Both MEED and Plain have moderate language modeling abilities compared to the other two counterparts. To reason about why these models' rankings swapped depending on the emotional polarity, we make two noteworthy observations. First, even though the gap in scores is not huge, Plain is rated significantly more repetitive than MEED (Figure 6.2). Second, as it can be seen from Table 6.1, both chatbots actively ask questions in their responses, but MEED asks significantly more questions than Plain in negative contexts (independent t-test

Table 6.1: Average number of questions with standard deviation (in the parentheses) asked by different chatbots.

	Dialog level			**Turn-level: Positive**			**Turn-level: Negative**		
	Overall	Pos.	Neg.	Turn 1	Turn 2	Turn 3	Turn 1	Turn 2	Turn 3
MEED	1.83 (0.96)	1.78 (0.98)	1.89 (0.94)	0.91 (0.29)	0.56 (0.51)	0.3 (0.46)	0.93 (0.26)	0.63 (0.48)	0.33 (0.47)
Blender	1.12 (0.94)	1.25 (0.92)	0.99 (0.93)	0.73 (0.55)	0.21 (0.43)	0.32 (0.52)	0.73 (0.65)	0.11 (0.35)	0.15 (0.37)
MIME	1.26 (0.91)	1.26 (0.94)	1.25 (0.88)	0.34 (0.48)	0.46 (0.54)	0.47 (0.55)	0.35 (0.48)	0.42 (0.51)	0.48 (0.53)
Plain	1.72 (1.08)	1.78 (1.07)	1.68 (1.09)	0.83 (0.38)	0.52 (0.50)	0.43 (0.50)	0.82 (0.39)	0.53 (0.50)	0.34 (0.47)

$p < 0.01$).

It is expected that MEED establishes greater diversity and poses questions as it models multiple listeners' empathetic intents, where questioning makes up the largest class. However, as we observed during qualitative inspection of the dialog logs, MEED falls into a typical trap of neural-based chatbots by learning the most safe and general type of response, in this case question, e.g., *"What type of ... was it?"* While asking such general information-seeking questions is an acceptable strategy for positive contexts, other questioning behaviors were shown to be more effective in delivering meaningful emotional regulation in negative scenarios [221].

We further combined these observations with the fact that correlation between these chatbots' repetitiveness scores and overall ratings is slightly lower in negative scenarios (Pearson's $r = -0.42$ ($p < 0.001$)) than in positive (Pearson's $r = -0.51$ ($p < 0.001$)). It suggests one plausible explanation to the observed phenomenon. In positive contexts, human speakers value chatbots' diversity and active engagement demonstrated via questioning, and are more forgiving even if the chatbot's response is slightly misaligned with the context. In negative scenarios, speakers feel much more vulnerable and expect greater attention. Consequently, they prefer a generic, but safe response over the one which is somewhat unrelated or diverting attention from the speaker's emotional state. Figure 6.5 provides examples illustrating these observations.

Decline of Blender in negative contexts

To study the possible reasons of Blender's lower performance in negative contexts, we started with qualitative inspection of dialog logs. While Blender asks fewer questions than MEED or Plain, they still appear frequently in its responses (Table 6.1) and the same issue of asking overly general questions, failing to address speaker's emotional needs in negative contexts,

MEED (context: Positive, rank: Okay) **Emotion**: anticipating **Prompt**: I just ordered a new cookery book and I can't wait for it to arrive. I love to cook and will really enjoy trying out the recipes.	**Plain** (context: Positive, rank: Bad) **Emotion**: anticipating **Prompt**: I just ordered a new cookery book and I can't wait for it to arrive. I love to cook and will really enjoy trying out the recipes.
Speaker: I just ordered a new cookery book and am eagerly awaiting its arrival. I want to delve in and try lots of new recipes! ***MEED:*** What kind of dish is it? ***Speaker:*** Well I want to find something with beef first. ***MEED:*** That sounds like a good recipe. ***Speaker:*** Beef always is. I think I will try to find a good casserole to try. ***MEED:*** That sounds delicious! I'm sure you will find a good casserole.	***Speaker:*** I just ordered a new cookery book and am eagerly awaiting its arrival. I want to delve in and try lots of new recipes! ***Plain:*** That sounds like a good idea. What kind of recipes do you like? ***Speaker:*** I love deserts the most. I want to try a new chocolate cake recipe. ***Plain:*** That sounds like a good idea. What kind of recipe is it? ***Speaker:*** I don't know yet. i need to look but definitely chocolate. ***Plain:*** That sounds like a good idea. I'm sure you'll like it.
MEED (context: Negative, rank: Bad) **Emotion**: terrified **Prompt**: I am so scared to live in my neighborhood. There is people that come around shooting their guns.	**Plain** (context: Negative, rank: Okay) **Emotion**: terrified **Prompt**: I am so scared to live in my neighborhood. There is people that come around shooting their guns.
Speaker: I am so scared to live in my neighborhood. There is people that come around shooting their guns. ***MEED:*** What kind of guns? ***Speaker:*** I don't know the kind of gun, but it is upsetting, and they are shooting close to my house. ***MEED:*** What kind of gun do you think they are shooting close to your house? ***Speaker:*** I don't know. A hand gun? What does that matter? ***MEED:*** Do you think they are shooting close to your house?	***Speaker:*** I am so scared to live in my neighborhood. There is people that come around shooting their guns. ***Plain:*** What kind of guns do you have? ***Speaker:*** I don't have any guns. ***Plain:*** I don't have any guns either. I don't like guns either. ***Speaker:*** So what should I do to feel safer in my neighborhood? ***Plain:*** I don't have any guns either. I don't like guns either.

Figure 6.5: Examples of chat logs with MEED and Plain.

preserves also for this chatbot.

More interestingly, we observed that Blender is the only chatbot in our experiment that persistently shares its own experiences and views with the speakers. To get a better idea of this behavior, we randomly sampled 50 chat logs (25 per emotional polarity) for each chatbot and annotated how many of their responses were oriented towards the speaker (other), reflecting the chatbot's own experience (self), or both. Blender is the only chatbot, for whom the amount of responses oriented towards self and both largely exceeds zero in the later dialog turns, almost reaching the proportion of purely other-oriented responses (Table A.6 in Appendix A.9).

Blender (context: Positive, rank: Good)

Emotion: nostalgic

Prompt: I get sentimental during the holidays. I think back to the ones i had as a kid.

Speaker: *I love the holidays, that make think back to being a kid.*

Blender: *i agree, it's a great time to spend with family and friends. What do you do for fun?*

Speaker: *We bake cookies, listen to christmas songs. What about you?*

Blender: *i like to play video games and spend time with my family. what kind of cookies do you bake?*

Speaker: *pecan, chocolate chip, yum!*

Blender: *what is your favorite video game to play? i've never played video games, but i love video games.*

Blender (context: Negative, rank: Bad)

Emotion: devastated

Prompt: I was so upset when I failed my math test, I was only 1 percent off!

Speaker: *I was one percent off from passing my math test, I was devastated.*

Blender: *i'm sorry to hear that. were you able to get a better grade on the test?*

Speaker: *No, I am just upset.*

Blender: *i think i am going to go back to school. i am not sure what i will do.*

Speaker: *Make sure to study.*

Blender: *i am sure you will do great. i hope you get a good grade on your test. good luck!*

Figure 6.6: Examples of chat logs with Blender.

Self-disclosure is frequent in peer support discussion forums [10]. This likely explains Blender's tendency to share own perspective as it was pre-trained on Reddit conversations, where peer support is actively practiced. However, human attitude to chatbot's sharing about self is unclear, especially in negative scenarios. Even in human-human interaction, positive disclosure is appreciated more than negative [27]. Moreover, in counselling practice, therapist self-disclosure is usually portrayed as a mistake [90]. We could not find studies about users' preferences for the degree of chatbot's self-oriented responses, but some previous findings about embodied computer agents reveal that their empathetic other-oriented emotions lead to more positive ratings of the agent [22]. We, therefore, hypothesize that pulling attention to self too quickly in negative conversations might have resulted in Blender's poorer performance in this emotional polarity, which is demonstrated with an example in Figure 6.6.

6.6 Discussion

6.6.1 Implications for chatbot development

Most of the chatbots in our experiment were trained to model short-context conversations and did not support the interactive chat mode by default, which also applies to other dialog models, e.g. [95, 133]. Nevertheless, being able to maintain continuous engaging conversation is an ultimate goal for empathetic chatbots. Thus, more attention should be paid to adapting training procedures and architectures to track longer-term dialog history and evolution of speaker's emotions.

Our findings demonstrate that users' emotional needs differ in positive and negative scenarios, and that they do not necessarily expect a strong emotional reaction to their inputs. Raising a question may be an appropriate response. According to our results, chatbots should dwell longer on speakers' negative situations, employing meaningful questioning strategies, which can possibly be achieved by modeling fine-grained empathetic questioning intents [221]. In addition, more research on the amount of chatbots' self-disclosure would further help tailor chatbots' responses to users' expectations.

6.6.2 Next steps

While human evaluation is the current standard to assess chatbots' performance, developing an automated metric to approximate human judgement is an important milestone that would considerably facilitate the developmental cycle. Some attempts towards this goal have been made [251], but very few of these metrics try to capture empathetic abilities of chatbots. Our analysis suggests that all dimensions evaluated in our Likert-type questionnaire constitute significant predictors of the overall human satisfaction (Section 6.5.2). Therefore, to develop a stronger automatic proxy for human evaluation, we consider creating rationale heuristics approximating those dimensions and identifying a meaningful way to combine them into a single score. The dataset of collected chat logs and human scores from our experiment should streamline the construction and calibration of such a metric.

6.7 Limitations

In our work, we applied iEval framework to benchmark four empathetic agents. We did not compare them against human-human interaction, as synchronizing two crowdworkers for conducting several chats between each other entails more logistical difficulties. More importantly, we were mainly interested in measuring how existing chatbots address users' emotional needs, rather than checking if they are indistinguishable from human interlocutors.

Our results show that bigger models rank higher in the evaluation task. It raises the subsequent question about to what extent the proposed framework measures differences in models' empathetic abilities compared to their underlying language model performances. We believe that iEval is an effective framework for evaluating chatbots' empathy as it succeeded in registering intricate differences in the performances of MEED and Plain, two models of comparable sizes and pre-training pipelines, as well as distinguishing the performance of Blender in emotional contexts of different polarity. To further disentangle the role of language modeling and empathetic abilities, one can consider running the iEval evaluation experiment to compare equal-size models with and without fine-tuning for empathetic response generation (e.g., Blender, which was only pre-trained on Reddit, and Blender, which was further fine-tuned on the EmpatheticDialogues dataset). However, this was not the main objective of our study and we leave it for future work.

Finally, we propose to use ranking as a way of expressing the appraisals of the chatbots, as it affords advantages of both Likert scales and pairwise comparisons. Ranking may be less robust for comparing results across experiments with mismatched sets of chatbots. Applying rank aggregation techniques can be useful to tackle such cases [203].

6.8 Chapter summary

This chapter introduced iEval, a novel evaluation framework for open-domain chatbots that can detect humans' personal perceptions of social interaction, manifesting in emotional dialogs. We used iEval to benchmark four recent empathetic chatbots. Further analysis revealed several limitations in empathetic response generation approaches of these models, which came out due to their uneven abilities in handling positive and negative conversational scenarios. Based on our findings, we formulated implications informing future efforts in the development and evaluation of such chatbots. We also publicly released the data from our experiment to expedite future research in these directions.[2]

---

[2] https://github.com/Sea94/ieval

7 Automatic evaluation of social chatbots with prompting[*]

7.1 Introduction

The recent arrival of conversational AI, marked by the public release of ChatGPT from OpenAI, initiated unprecedented user engagement with conversational chatbots in a real-world setting. With the impressive naturalness of machines' responses, users are going beyond traditional transactional exchanges and start exploring more social interaction scenarios with increasing curiosity [225]. In such situations, users might be subject to social and psychological harms if dialog systems fail to follow commonsense social rules [219, 111]. Several instances of alarming social behavior of this technology have already been discussed in the media [198, 48, 131]. In this context, developing meaningful and robust evaluation metrics for these systems has become particularly urgent to ensure that the models are safe and acting in the best interest of the users before their release.

Initially, human evaluation was considered a de facto standard for evaluating dialog systems [126]. As running human evaluation is time- and resource-consuming, a number of automatic evaluation metrics for dialog systems have been proposed [149, 251]. The majority of these approaches aim to automate the *offline* user evaluation. In this setting, dialog evaluation is performed by a human judge who is distinct from the one conversing with the bot (Figure 7.1, offline). The metrics proposed for this case approximate the evaluation scores provided by this third-party human judge for the pre-produced dialogs [*e.g.* 150, 71]. Despite its popularity, offline user evaluation is limited in its ability to capture subjective perceptions of users who actually interacted with the bots [104, 123, 70]. This limitation of relying on second-hand evaluation can be illustrated by an analogy from the realm of restaurant critique when one tries to evaluate a restaurant solely by reading consumer reviews but having never actually

---

[*] Adapted from [218]: Approximating Online Human Evaluation of Social Chatbots with Prompting. Ekaterina Svikhnushina and Pearl Pu. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial), 2023.

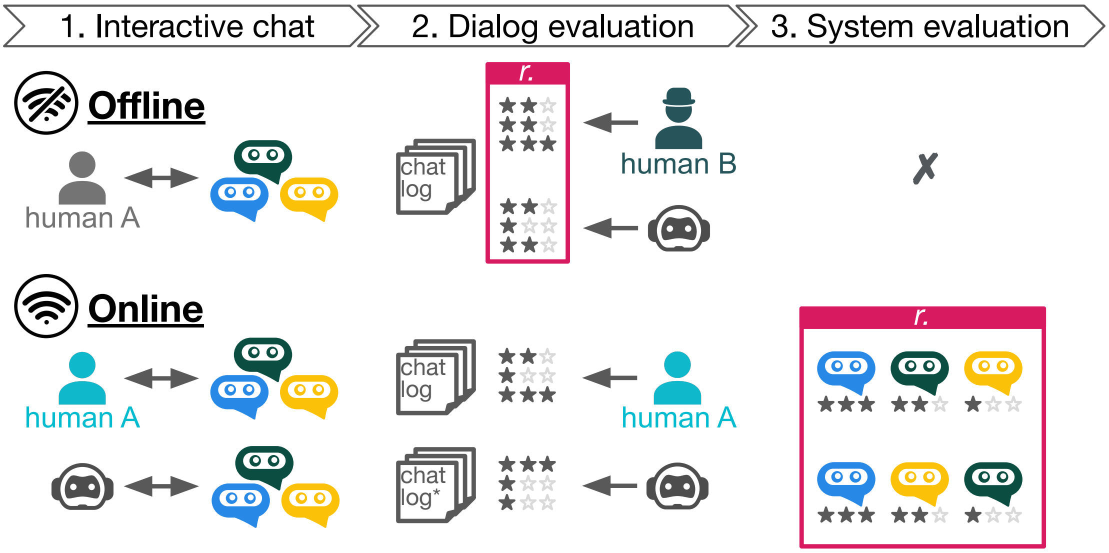


Figure 7.1: Offline and online dialog evaluation with the corresponding processes. In the first step, dialog logs are curated. In the second step, each dialog log is assigned a dialog-level score, either by a third-party judge (offline) or by the same conversational partner (online). In the third step, the system ranking is obtained by aggregating the dialog scores of each chatbot. Grey bot icons indicate steps that are intended to be approximated by means of automatic evaluation. Pink boxes mark the steps in the process where the correlation ($r.$) with the ground truth human judgment is computed to validate the automatic evaluation metric during its development process.

eaten there. Conducting *online* user evaluation, where the same individual interacts with the bot and assesses its performance, is more likely to produce accurate and precise evaluations of the chatbot's performance. Moreover, this method offers better predictive capabilities for the system use "in the wild" [12]. However, by far, efforts towards approximating online user evaluation have been limited.

To address this gap, we propose a novel automatic **D**ialog system **E**valuation framework based on **P**rompting, DEP. Our framework automates the whole pipeline of dialog system evaluation in an interactive setting, replicating live user studies. As the first step towards this goal, we leverage a large language model (LLM) from the GPT-family models to collect synthetic chat logs of evaluated bots with the LLM. Second, we prompt the same LLM to produce the resulting evaluation scores for generated chat logs and, finally, rank the chatbots based on their overall performance (Figure 7.1, online).

While using bot-play is not a new idea per se, we emphasize the importance of carefully choosing a dialog partner for the evaluated chatbots specifically for social conversational contexts where the roles of two interlocutors can differ significantly. For example, it was shown that the emotion/intent distributions in conversations between an emotional speaker and an empathetic listener are very different for the two dialog partners [239]. To account for it, in the first step of our framework, we propose prompting LLMs to play a particular social role over the course of the interaction with the chatbots to be evaluated. For the second step, we draw inspiration from the fact that LLMs demonstrate solid performance improvement when their generation process is augmented with instructions [111]. We demonstrate that prompting the model with appropriate instructions that explain how fine-grained evaluation dimensions relate to the overall dialog score leads to substantial performance improvement, reaching up to $r = 0.95$ Pearson correlation with the human judgment on a system level.

Overall, our contributions include the following. 1) We describe an end-to-end prompting-based evaluation framework for dialog systems, specifically targeting social interaction scenarios (Section 7.3). 2) Our experiments showcase the effectiveness of prompting for assigning a desired social role to LLMs and, thus, collecting machine-generated dialogs that better approximate real interpersonal communication (Section 7.4.1). 3) We consider different prompt designs and conclude that including demonstrations together with instructions results in the best performance (Sections 7.4.1, 7.4.2).

7.2 Related work

7.2.1 Automatic evaluation of chatbots

Automatic dialog evaluation has been a long-standing goal for practitioners. Initial works focused on evaluating chatbots' responses against a ground-truth reference [170, 222]. Following works moved on to exploring reference-free evaluation metrics as referenced evaluation was shown to be ineffective due to a wide range of acceptable responses for a single context [136], implying that comparing with a single reference is limited. Reference-free metrics usually operate either on the utterance or the dialog level. For the utterance level, practitioners have explored ways to evaluate response appropriateness for the preceding context [120, 168] or predict the qualities of the follow-up response as a proxy for the quality of the preceding dialog [71, 72, 150]. For the dialog level, a number of diverse approaches have been proposed, ranging from aggregating several fine-grained utterance-level evaluations [260], to designing training objectives to model the information flow across dialogue utterances [128], employing graph representations to capture dialog dynamics [97, 259], and using semantic-level manipulations to teach the evaluation model to distinguish coherent and incoherent dialogs [73].

The works above largely target the offline evaluation setting. Some scholars have also started exploring different ways of approximating online user evaluation. Deriu et al. proposed a partially automated framework where human judges rank chatbots regarding their ability to mimic conversational behavior using interactively collected bot-to-bot conversations, which relies on survival analysis [49]. Sato et al. proposed a particular bipartite-play approach for collecting bot-to-bot conversations to provide a fairer comparison setting for evaluated chatbots [200]. These papers consider methodologies for organizing bot-to-bot conversation sessions, but they are not concerned with the way how these bot-to-bot conversations unfold.

In our work, we explore the use of bot-to-bot conversations to model a desired social behavior.

7.2.2 Prompting

Prompt-based learning paradigm [137] received significant attention after Brown et al. demonstrated how GPT-3, a large foundation model, can well handle a wide range of tasks without the need for fine-tuning, relying only on natural-language prompts and task demonstrations as context [25]. Prompt-based model performance depends on the design of the provided prompt. Prompt engineering efforts explore approaches for designing prompts, which vary in the shape of prompts (cloze or prefix), human effort required for writing prompts (manual or automatic), and number of demonstrations provided to the model in the prompt (zero-shot or few-shot) [137].

Prompt-based learning applied to recently created LLMs has been reported to achieve outstanding results on a variety of tasks and benchmarks, including classification, reasoning, coding, translation, and many others [*e.g.* 237, 34, 35]. However, exploring prompting for evaluation of dialog systems has not been widely investigated. We are only aware of one more simultaneous and independent effort in this direction. Huynh et al. studied how different LLM parameters (type, size, training data) may influence the dialog evaluation, focusing on utterance- and dialog-level evaluation in the offline evaluation setting [101]. Our work focuses on how prompting can be used to capture a holistic evaluation of dialog systems in online social settings, relying on freshly generated dialogs.

7.3 Proposed method: DEP

We introduce our DEP framework, which consists of two consecutive steps. First, it requires collecting interactive chat logs between the LLM and evaluated chatbots, which we denote as LLM-to-bot play. Second, the LLM is prompted to generate scores for these chat logs. The generated scores are further aggregated to produce a final ranking of the systems. We describe each of the steps below.

7.3.1 Prompted LLM-to-bot play

In social settings, two partners may play considerably different roles in a dialog, thus establishing very distinct conversational behaviors. Examples include conversations between a student and a teacher, an emotional speaker and an empathetic listener, or even between two interlocutors with different personas. Chatbots are usually built to perform well in one of these roles (e.g., empathetic listener), but not necessarily the other. Therefore, collecting synthesized dialogs via self-play of the chatbot with itself (or a similar competing model) might fail to represent a realistic discourse flow due to the differences in the intents produced by speakers and listeners in dialogs.

To address this consideration and render the synthesized dialogs that better approximate real social interactions, we propose leveraging LLMs' ability to produce responses on behalf of an assigned character [224]. Specifically, we suggest letting the evaluated chatbots converse with an LLM prompted to play a particular social role. Figure 7.2 demonstrates how to structure the prompt to produce each next output of the LLM in an interactive manner. Meanwhile, responses from the evaluated chatbots are computed by passing the accumulated dialog history to these chatbots as input context. The process can be repeated for multiple dialog turns. The length of the exchange may depend on the extent of details provided to prompt the LLM. The more specific the prompt is, the faster the evaluated chatbot can demonstrate its performance in the social situation of interest. On the contrary, more generic conversation starters require more dialog turns to reveal the targeted social behavior.

7.3.2 Prompted evaluation

Once dialog logs are synthesized, we propose using prompting to produce evaluation scores for each dialog. Prompts can be constructed in several ways. We investigate zero-shot and few-shot settings, either with or without instructions, in our experiments (Section 7.4). Many available foundation LLMs are accessible through APIs and only output text completions without corresponding log probabilities. Therefore, regardless of the type of prompt that we use, to generate a score for each dialog, we obtain a textual form of the score from the LLM completion and then use a verbalizer function to map it to a numerical value, getting inspiration from [201]. Formally, given a dialog log d, we construct a prompt $P(d)$ that takes d as input and outputs a prompt that contains exactly one mask token as a placeholder for the dialog score. Let y be a predicted token for $P(d)$. We then define a verbalizer as an injective function v that maps each score in textual form to a numerical value. Thus, $v(y)$ produces a numerical score for a single dialog. The final rating of a given dialog system is obtained by averaging the corresponding dialog scores of that system. For fair evaluation, the number of dialogs collected for each evaluated chatbot should be identical.

I am a Speaker <*in an assigned social situation*>. I am sharing <*my thoughts*> with a Listener in a dialog.
Speaker: <*LLM's input #1*>
Listener: <*Bot's response #1*>
Speaker:

Figure 7.2: Prompt template to condition a LLM to play an assigned social role while interacting with an evaluated chatbot.

7.4 Results

For all reported experiments, we used the most capable version of the InstructGPT model (`text-davinci-003`) available at the moment of initiation of our experiments in early Q1 2023. We used this model as it was easily accessible through OpenAI API[1] and was expected to have superior performance for social scenarios as it was trained based on human feedback, which captures subjective human judgment of interactive outputs [167].

Following previous works that considered system-level evaluation [140, 70], we report Pearson correlation for our experiments, unless specified otherwise. We also opted for this type of correlation coefficient as it performed better for capturing whether the automated metric succeeds in preserving the gap in scores for the best- and least-performing chatbots, the information which gets lost with rank correlation.

We start by demonstrating the application of our evaluation framework to empathetic dialog systems as in these interactive scenarios two conversational partners have clearly distinct social roles: an emotional speaker and an empathetic listener. Further, we consider the generalizing ability of the framework to other social domains.

7.4.1 Evaluation of empathetic chatbots

Below, we first describe the dataset used for the experiment. Then, we consider the ability of prompted LLM to effectively replicate social discourse patterns over multi-turn interactions with the chatbots that serve as eventual evaluation targets. Finally, we explore several types of prompts applied to synthesized LLM-to-bots dialogs to evaluate how well they can approximate human judgment on a system level.

Dataset and evaluated chatbots

For this experiment, we used iEval dataset, introduced in Chapter 6. The dataset features human conversations with four empathetic chatbots collected in an online interactive manner. During the dataset curation process, each human was assigned an emotion label with the situation description taken from the EmpatheticDialogues dataset [187] and asked to have a 6-turn conversation with each chatbot while playing a character in the assigned scenario. Overall, there are 480 situation descriptions in the dataset, which evenly cover two emotional polarities: positive and negative. As each chatbot participated in each scenario, there are in total 1920 dialogs in the dataset. After conversing with the chatbots, human interlocutors provided their appraisals of chatbot listeners in each dialog, including five fine-grained listener qualities on a 5-point Likert scale: politeness, empathy, likability, repetitiveness, and making sense, and an overall dialog rating on a 3-point scale. All scores are provided on a dialog-level.

The four chatbot models used to curate the dataset were Blender [197], MIME [142], MEED

---

[1] https://openai.com/blog/openai-api

and Plain [245]. We use these models in the same configurations for our experiment in this chapter.

LLM-to-bot play results

As the first step to validate our evaluation framework, we analyzed whether the LLM succeeds in mimicking human discourse following an assigned social role and whether approximating human speakers with the LLM causes any considerable changes in the chatbots' response patterns.

To generate LLM-to-bots conversations, we closely followed the procedure of iEval dataset curation. Specifically, we used emotion labels and situation descriptions from the dataset to create prompts for the LLM: *I am a Speaker, feeling <emotion> because <situation>. I am sharing these emotions with a Listener, expecting empathy and understanding from them. I respond as a Speaker in a dialog.* The first LLM input was also taken from the iEval dataset. For each scenario, we collected LLM conversations with each of the four bots, letting them converse for 6 turns, i.e., 3 inputs from the LLM and 3 responses from the chatbot.

To examine the similarity of discourse patterns between human-to-bots and LLM-to-bots conversations, we started by annotating each dialog turn in two datasets with emotion and empathetic intent labels, using emotion/intent classifier developed by Welivita and Pu for EmpatheticDialogues dataset [239]. As datasets in our experiment were grounded in situation descriptions taken from EmpatheticDialogues, the classifier was expected to generalize well to our data.

Consequently, we visualized the most prominent discourse patterns for two corpora in the form of Sankey diagrams, shown in Figures 7.3 and 7.4. From the visual inspection, it can be seen the LLM emotion distribution over the course of the dialog (Figure 7.4) largely resembles one of the human interlocutors (Figure 7.3). More importantly, sets of intents produced by empathetic chatbots are also very similar between the two figures, with *Questioning, Sympathizing,* and *Acknowledging* being the most prominent ones. Thus, our freshly generated interactive dataset with LLM-to-bot play was deemed to produce a reasonable approximation of human-to-bot conversations.

Prompted evaluation results

Turning to the second step of our evaluation framework, we examined different types of prompting to produce scores for the generated LLM-to-bot dialogs. Specifically, two variables in the prompt design were considered.

First, we tried score generation in zero-shot and few-shot settings. For the few-shot setting, the number of demonstrations was fixed to the number of points in the ground truth human evaluation scale, with one representative example supplied for each score. Thus, for the iEval

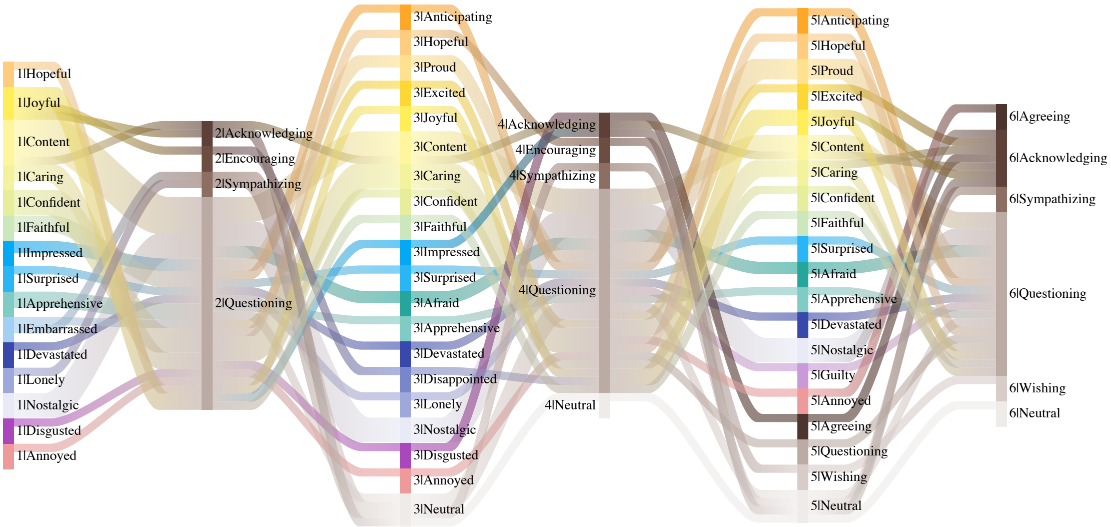


Figure 7.3: Sankey diagram showing discourse patterns in human-to-bots conversations originating from the iEval dataset.

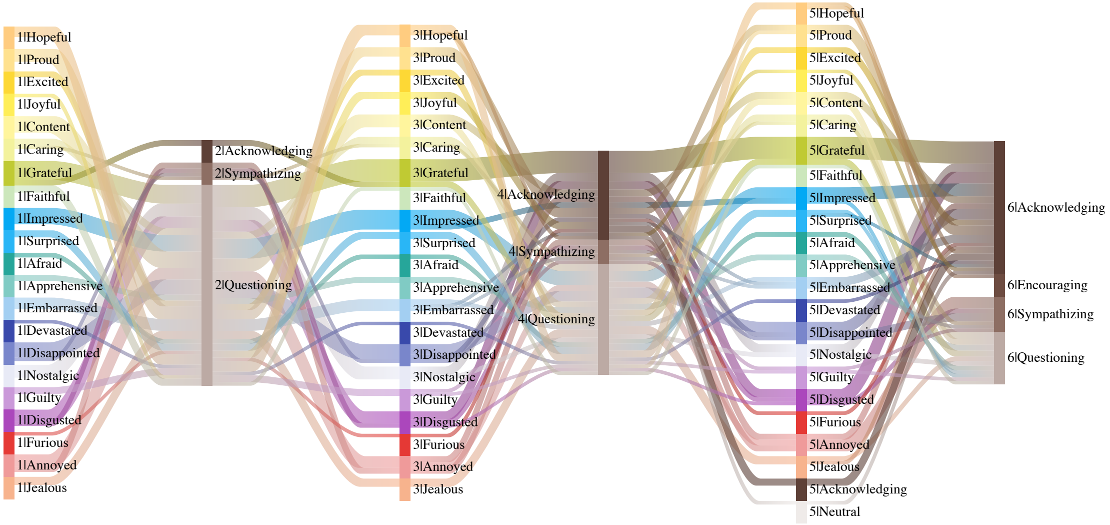
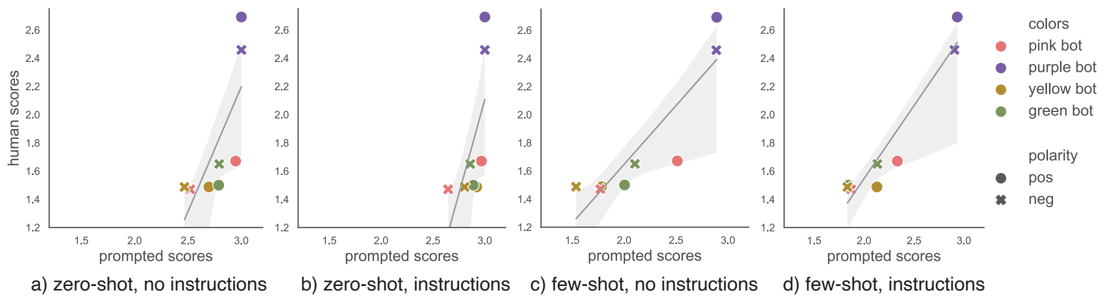


Figure 7.4: Sankey diagram showing discourse patterns in freshly generated LLM-to-bots conversations.

dataset, we used three demonstration dialogs corresponding to the three possible evaluation scores: *Bad*, *Okay*, and *Good*. The examples were selected manually and are provided in Table A.7 in Appendix A.10.

Second, we analyzed whether providing additional instructions helped the LLM evaluation performance. To write the instructions, we relied on the findings from our Chapter 6, which explained how chatbots' performance on various fine-grained dimensions translates into the overall score. As we observed the difference in humans' expectations of an empathetic listener



a) zero-shot, no instructions b) zero-shot, instructions c) few-shot, no instructions d) few-shot, instructions

Figure 7.5: Scatter plots depicting the system-level correlation results. Human scores are based on the iEval dialog annotations, while prompted LLM scores are computed based on the generated dialogs.

in positive and negative conversational scenarios, we devised slightly different instructions to prompt the evaluation of these two emotional polarities. Specific formulations of the instructions are also provided in Table A.7 in Appendix A.10.

To generate scores for each dialog, we prompted the LLM to complete the masked score, provided the log of the evaluated dialog. Depending on the configuration, few-shot demonstrations and/or instructions were prepended to the prompt. A template of the used prompt can be found in Figure A.16 in Appendix A.10. After obtaining dialog-level scores, we aggregated them to produce system-level ratings. One system was defined as a chatbot operating in one of the two emotional polarities. This decision is driven by the fact that based on human evaluation results in Chapter 6, chatbots demonstrated statistically significant differences in their performance depending on the emotion. Thus, we considered eight systems for computing system-level correlations.

System-level correlations between human- and LLM-judgments for each of the four possible prompt design manipulations are reported in Table 7.1. Few-shot prompting with instructions results in the highest correlation of 0.954, which is further illustrated by the scatter plots in Figure 7.5. According to the plots, providing examples helps the LLM to calibrate the produced scores, eliminating the positivity bias, whereas instructions result in reduced variance.

Table 7.1: System-level Pearson correlation for four possible prompt design manipulations, with the p-value in brackets.

	No instructions	Instructions
Zero-shot	0.748 (p=0.033)	0.651 (p=0.080)
Few-Shot	0.892 (p=0.003)	**0.954** (p<0.001)

7.4.2 Generalizability to different domains

In this section, we consider how prompted evaluation can generalize to different corpora and conversational settings. As the results above suggested that prompts combining instructions with examples perform best for evaluation, for the following experiment we searched for datasets that allowed formulating instructions for defining what properties correspond to good or bad overall appraisal ratings of the dialogs. Therefore, we selected two datasets that contained both fine-grained and overall ratings of the dialogs and used the information of the most relevant fine-grained dimensions to formulate instructions. We also considered only those datasets that contained multi-turn dialogs collected following the interactive process.

The selected datasets feature human-to-bot dialogs, with some dialog systems that are not publicly available. Moreover, these dialogs were collected in a generic manner, without the purpose to model any specific social behavior (e.g., as empathy in iEval). Due to these considerations, in the following experiments, we only studied the performance of the second step of our DEP framework, skipping the synthesis of new LLM-to-bots conversations. In a general case, when researchers have access to their evaluation targets, prompting LLMs to engage in a generic social interaction with the evaluated bots should be straightforward as we demonstrated in Section 7.4.1.

Datasets

To study the generalizability of prompted evaluation, we used FED [150] and DSTC9 datasets [83]. FED contains 124 open-domain dialogs of humans with humans and two chatbots (Meena and Mitsuku) that were originally released by [2]. DSTC9 contains 2200 human-bot conversations from 11 chatbots. In both datasets, all dialogs are annotated with offline human appraisals of ten fine-grained dialog qualities and an overall impression rating that were curated following the same protocol described in [150].

Prompted evaluation results

To construct a prompt for evaluating the chosen datasets, we selected five dialog examples covering five possible scores for overall dialog ratings, ranging from *Very bad* to *Very good*; they are provided in Table A.8 in Appendix A.11. To formulate the instructions, we used information from the original paper describing the relative importance of each fine-grained dialog quality for the overall impression. The specific formulation of the instruction is provided in Appendix A.11.

The evaluation results with a comparison to existing best-performing evaluation metrics are provided in Table 7.2. As the number of systems in the FED dataset is small, we only report dialog-level correlation. We also report Spearman correlation for this dataset for the purpose of comparison with the results in the original paper ($r = 0.443$ ($p < 0.05$)) [150]. Our prompted evaluation exceeds correlations of previous metrics by a considerable margin on both datasets

Table 7.2: Results on FED and DSTC9 data. Previous best results are obtained from [251]. Dialog and System indicate dialog- and system-level correlations, respectively, with P standing for Pearson and S for Spearman correlation. All values are statistically significant to $p < 0.05$.

	FED	DSTC9	
	Dialog (S)	Dialog (P)	System (P)
Prev. best	0.547	0.147	0.907
(metric)	Zhang et al. [259]	Li et al. [128]	Li et al. [128]
DEP	**0.655**	**0.274**	**0.980**

and, thus, demonstrates the ability to generalize to new open-domain conversational settings.

7.5 Discussion

Dialog system evaluation with prompting showed its usefulness both for generating new interactive exchanges with the evaluated systems and for judging their performance, therefore, allowing for a reasonable approximation of the online user evaluation pipeline. We deem this approach particularly promising for the evaluation of social aspects of conversations. LLMs used for prompting suffer from occasional hallucinations, i.e., a tendency to make up factual information [167]. It might be difficult to keep track of all specific factual items of information that come up in the interactively created dialog between two conversational models and search for ground truth references for each of them to construct objective metrics such as model's accuracy or truthfulness [132]. Whereas, prompting the LLM to establish a specific behavior and providing instructions about commonsense social norms appears more feasible once these instructions are established.

Drawing from the visualization of discourse patterns in our newly collected dataset of dialogs between the LLM and empathetic chatbots, we observed that the prompted LLM largely mirrors the conversational patterns of humans. However, there are also some differences. For example, in Figure 7.4 there is an apparent sub-flow with a *Grateful* emotion, increasingly displayed by the LLM. We believe the LLM might have developed an agreeable "personality" due to its training procedure based on Reinforcement Learning from Human Feedback, which optimized LLM's responses to satisfy human labelers. Future research can consider alternative prompting techniques to make the emotion/intent distribution of LLMs' responses even more balanced and representative.

We conducted our experiments with only one LLM and explored the few-shot prompting scenarios with a fixed number of demonstrations. Future studies could explore the applicability of other LLMs for the DEP framework, as it has been already initiated by [101]. An area of particular interest would be to study the efficacy of the framework working with open source LLMs, such as LLaMa [226].

We would also like to explore how DEP generalizes to other phenomena in social conversations, apart from generic open-domain interactions and empathetic dialogs. For example, further studies might focus on applying the framework to evaluate toxicity or humor in dialogs. However, this research direction requires curation of appropriate calibration datasets.

Last but not least, evaluation artifacts produced by DEP may be used to assist designers of chatbots as they allow for both analyzing the synthesized logs and comparing quality ratings. These insights may be integrated into assistive chatbot design tools, such as *iChatProfile* [87], to offer a faster prototyping cycle due to the automatic generation of chat logs and richer insight about chatbot profiles due to additional rating information provided by the last step of DEP.

7.6 Chapter summary

In this chapter, we proposed DEP – a framework for evaluating social chatbots using prompting. Our framework addresses the limitations of evaluation approaches using benchmark datasets in an offline setting. We describe how LLMs can be leveraged to synthesize realistic conversational logs with the evaluated chatbots in an online interactive manner. We further outline how the knowledge about the desired fine-grained qualities of a conversational partner can be translated in the prompting instructions to generate reliable overall scores for the collected dialogs. The proposed framework streamlines the evaluation process, making it highly efficient in terms of both time and cost, by removing the need for human involvement at every step. Our experiments demonstrated that the prompting-based evaluation results achieve high correlation with the human judgment, reaching impressive Pearson $r = 0.95$ system-level correlation for the iEval dataset, which features dialogs with empathetic chatbots. We explain our vision why this framework is well-suited for evaluation of social phenomena in conversations and lay out future research directions. We also publicly release all freshly curated chat logs between the LLM and evaluated chatbots, as well as all additional annotations for the iEval, FED, and DSTC9 datasets created for this study.[2]

---

[2] https://github.com/Sea94/dep

Conclusion Part V

8 Conclusion

8.1 Summary of contributions

This thesis puts forward the idea that combining advanced language modeling tools with insights from interdisciplinary user-centered research holds vast potential for developing novel, meaningful evaluation frameworks for social chatbots. In the studies described above, we closely followed the steps of the user-centered design process (Figure 8.1) [102, 41], showcasing how gaining an understanding of users' behaviors, contexts, and needs can inform the development of conversational agents. This understanding, coupled with powerful NLP instruments, enabled us to design appropriate evaluation procedures that can further enhance the performance of chatbots. We summarize the contributions made throughout this journey below.

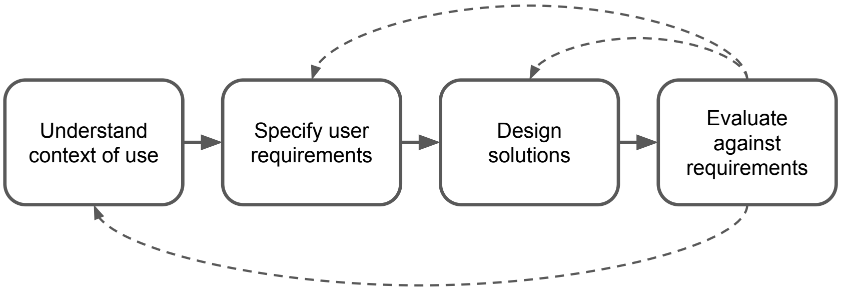

Understand context of use
Specify user requirements
Design solutions
Evaluate against requirements

Figure 8.1: User-centered design process (adapted from [102]).

Evaluation criteria development

In Part II we conducted rigorous research to elicit and validate users' desiderata for social chatbots' traits and behaviors. Specifically, in Chapter 3 we conducted exploratory semi-structured interviews with the users and used the finding to define a structured model of the desired chatbots' qualities. We further validated the established model using psychometric techniques through a large-scale user survey. The results of this study revealed how users' personal beliefs and attitudes, including empathy propensity and openness to new technologies, influence their expectations of chatbots and exposed three pillars of users' concerns: psychological, social, and monetary harms. More importantly, the model delineated key factors determining users' adoption intention of chatbots, which include the chatbot's ability to follow the Politeness protocol, Entertain the user, practice Attentive Curiosity towards the user, and express Empathy.

In the following study in Chapter 4, we triangulated these findings by assessing the experiences and expectations of users of currently available social chatbots in a mixed-method study drawing from online reviews. The results supported the conclusions that the entertaining aspect of chatbots plays the principal role in maintaining users' engagement with the available agents, while their intrusive or rude behavior has the highest association with user churn. The analysis of users' expectations further emphasized their interest in more personalized attention from the chatbots, along with their improved emotional intelligence.

To summarize, the studies described in Part II addressed the challenge of ill-defined evaluation objectives for social chatbots and produced clear criteria for assessment representing users' perspectives. These criteria served as guiding principles for the rest of this thesis, helping us scope, prioritize, and design the subsequent research efforts. Additionally, they allowed us to formulate specific design implications for immediate consideration of the technology developers and identify broader directions requiring further attention from the research community.

Annotation scheme and tools for empathetic questions

Having learned that attentive curiosity and empathy constitute the largely missing qualities of social chatbots, in Part III we directed our efforts towards addressing this gap. Incorporating the knowledge gained from studies in social psychology and linguistics, we uncovered that questions play a vital role in casual dialogs to establish common ground and demonstrate an engaged and attentive attitude towards a conversational partner. We further identified that the work on incorporating question-asking abilities into social chatbots is scarce and attributed this deficiency to the lack of appropriate resources allowing to analyze and model question-asking strategies in social conversational settings.

Therefore, in Chapter 5, we derived a taxonomy of empathetic questions by conducting several iterations of qualitative coding of questions occurring in emotionally-colored social dialogs and validating the interpretability of the labels by assessing the agreement with external raters. The produced taxonomy consists of two branches, capturing 9 semantic-driven communicative actions of questions (acts) and 12 emotional effects that the questions should have on a dialog partner (intents). We used the taxonomy to produce annotations to one of the most prominent datasets featuring EmpatheticDialogues [187]. For that, we designed and ran a crowd-sourcing annotation task to obtain a seed labeled set of data and then trained automatic classification models, QBERTs, to generate the labels for the whole dataset. Finally, we employed visualization techniques to analyze co-occurrences of different question acts and intents and their effect on the interlocutor's emotions.

In summary, the EQT annotation scheme, in combination with the QBERT models, can help create better question-generation models, improving their controllability and explainability. Moreover, the analysis of the question-asking behaviors of empathetic listeners in the annotated dataset yielded valuable insights into interpreting the flow of the conversations and their differences in positive and negative emotional contexts. This improved conceptual comprehension played a crucial role in shaping the design and structure of the subsequent evaluation task.

Novel evaluation frameworks

In Part IV we turned to designing meaningful evaluation frameworks for social chatbots taking into account the lessons learned in preceding studies. First, we noted that certain listener qualities, such as attentiveness, empathy, and understanding, may only become evident over the course of multiple dialog turns, implying that a reliable evaluation framework should require the evaluated chatbots to take part in a conversation for a sustained period. Second, we emphasized that social and emotional interactions are highly subjective experiences, varying from person to person. Therefore, we found it crucial that the same user who engaged in a conversation with the chatbot should be the one to rate it to avoid additional subjectivity bias from external evaluators. Taking these factors into account, we posit that interactive live experiments provide the most representative assessment of the user experience with chatbots.

In Chapter 6, we proposed such a framework to run an interactive human evaluation of empathetic chatbots and demonstrated its application to benchmark four state-of-the-art models. We took careful steps to account for the topical and emotional diversity of conversational contexts in the design of our experiment. For collecting users' judgments, we incorporated questions about fine-grained evaluation criteria of chatbots' performance originating from the PEACE model (Chapter 3) and introduced a novel ranking-based approach to obtain the overall appraisal scores. The results of our experiment revealed significant interaction effects originating from chatbots' uneven performance in positive and negative emotional settings. We explained these results by analyzing the relative importance of fine-grained dimensions to the overall chatbot's score and using the knowledge about appropriate questioning behaviors depending on the context, the trade-off between generic safe responses and greater diversity, and users' expectations about chatbots' self-disclosure. Apart from the conceptual evaluation framework and actionable findings for further chatbot improvement, we contributed a new benchmark dataset with quality annotations curated in an online experiment.

As a concluding step, in Chapter 7 we used the created dataset to construct and calibrate an automatic evaluation framework. For this purpose, we employed prompting of large language models from the GPT family, both for synthesizing chat logs with the evaluated bots and judging their quality. We highlighted that in social settings, two partners may play considerably different roles in a dialog and took steps to validate that prompting the LLM with relevant instructions results in adequate replication of human discourse patterns. We further relied on the findings from human evaluation results to design prompts with appropriate evaluation guidelines. The established pipeline succeeded in approximating human evaluation, achieving a very strong correlation with the human judgment on a system level, and demonstrating the ability to generalize to external dialog corpora.

Overall, the process outlined in Part IV showcased how to derive specific evaluation requirements for a social aspect of interest and use them together with advanced language modeling tools to efficiently produce a reliable evaluation of chatbots, removing the need for human involvement. We resealed all freshly created chat logs, annotations, and prompting instructions to facilitate further efforts in the evaluation of the social qualities of chatbots.

8.2 Future directions

Extending to other social phenomena

In this thesis, we established comprehensive criteria capturing users' expectations and preferences for social chatbots and developed novel frameworks to evaluate them. While we have largely concentrated our efforts on exploring and understanding the aspects of empathy and attentive curiosity of chatbots, it is important to consider other significant dimensions from our criteria. One such dimension is the evaluation of chatbots' sense of humor. Humorous outputs are an important component of the chatbot's entertaining abilities, which proved to be a critical factor for maintaining users' engagement. Additionally, we should pay further attention to potential risks associated with social interaction with machines. Specifically, we discussed risks related to users' social, psychological, and monetary well-being. While incorporating appropriate emotion-regulation mechanisms into conversational models addresses some of these risks to a substantial extent, there are still broader concerns that remain, related to the FATE topics: fairness, accountability, transparency, and explainability. By ensuring fairness, holding chatbots accountable for their outputs, providing transparent decision-making processes, and enabling explainability, we can build trust and protect users from potential harms. Thus, moving forward, expanding the evaluation framework to this broader spectrum of factors is a crucial step towards further mitigation of potential risks that users may face. An essential factor to consider while continuing this research is that people may operationalize requirements for an agent differently than for a human. For example, the results of our iEval evaluation in Chapter 6 demonstrated that people do not expect chatbots to practice self-disclosure in negative conversational contexts, which is different from interpersonal communication between peers. We believe that the approaches discussed in this thesis may be extended and adapted for advancing the research of the outlined phenomena and provide additional considerations about the FATE topics in Section 8.3.

Considering cultural diversity

Throughout our work, we conducted our research in the English language and targeted our crowd-sourcing tasks at the participants based in the United States. These decisions were primarily driven by the availability of greater resources in the English language and the language fluency of the participants, ensuring their ability to comprehend our guidelines, instructions, and questions accurately. However, it is crucial to recognize that cultural and language differences play a fundamental role in shaping our social interactions. They influence the choice of acceptable topics for discussions, determine social reactions and involvement, and define norms of behavior, among other important aspects. Therefore, future research endeavors should prioritize the study of different populations, embracing the heterogeneity of cultural and linguistic groups. This involves diversifying resources for other languages and adapting the evaluation frameworks to encompass cultural and linguistic variables. This can help develop socially intelligent chatbots that are inclusive, culturally sensitive, and capable of effectively engaging individuals from various backgrounds.

Towards truly conversational assistants

Digital conversational assistants, such as Apple's Siri or Google Assistant, are becoming increasingly popular. It is predicted that by 2024, there will be over 8 billion devices with these assistants in use globally. However, users engage with them only for a limited set of basic single-step tasks [220]. The issue is that these assistants still struggle to properly understand users' requests and fail to keep a prolonged and meaningful conversation on a topic of interest. Moreover, their proactivity and usefulness are limited because they lack contextual awareness.

In interpersonal communication, social exchanges play an important role in building up that awareness [75, 66] and maintaining engagement [66, 223, 32]. However, in our research and developmental efforts devoted to conversational technologies, practitioners usually make a clear distinction between task-oriented and open-domain agents and take a definite stand whether they work on one type or the other. Therefore, it is not surprising that existing *task-oriented* conversational assistants fail to enhance their understanding of the user's context through *social* chitchat. A promising step forward would be concentrating on building agents that can combine the best of task-oriented and open-domain worlds and use the information gained through one mode of interaction to improve the delivery on the other. Empirical evidence suggests that task-oriented agents may enhance their perceived efficacy and likability by exhibiting social skills [160, 28]. Thus, working towards creating hybrid agents with seamless transitions could lead to the emergence of more versatile assistants that serve a much better job to their users. Profound efforts in this direction have already been taken in the domain of interview chatbots. Researchers have proposed solutions and evaluation frameworks for systems that engage a user in a mixed-initiative conversation with the virtual interviewer agent that guides the interview flow to elicit information, while also allowing the user to pose questions and make social comments [124, 244, 87]. Adoption of such practices in other domains is expected to bring further improvement in user experience during their interaction with conversational agents.

8.3 Ethical considerations

Crowd-sourcing

We used the Mturk platform for different studies outlined in this thesis, to field surveys, collect annotations for the datasets, and run online experiments. Crowd workers on Mturk are known to be underpaid according to Western standards, earning a median hourly wage of only ~$2/h [109]. At the same time, monetary remuneration is not the only factor defining people's motivation to work on such crowdsourcing platforms [89]. For example, workers might also engage with HITs to learn new or train existing skills, pass free time, or meet new people. Over the course of our involvement with the Mturk platform, we kept gaining experience and understanding of how the ecosystem of worker-requester relationships functions there. Thus, we continuously adjusted the qualification requirements and worker rewards in our study protocols. In our latest studies, worker reward amounted to $7.3 per hour, which was on par with the US minimum wage ($7.25). Even though in some cases practitioners might consider lower compensation rates to be appropriate as workers could have other reasons to complete the tasks than purely monetary reward, we encourage future works of similar nature to take careful consideration of this trade-off and offer fair payment to the workers. From our experience, we found that setting higher worker wages justifies setting higher quality control requirements, without reducing workers' interest in taking the task, which, in turn, results in better outcomes for both sides of the worker-requester relationship.

Socially aware chatbots

Previously, it was questioned whether machines should demonstrate social and emotional traits [36]. Concerns arose due to potential user resistance towards engaging in such interactions, as well as a lack of clarity regarding the rules, norms, and expectations governing these conversations, and the broader social implications involved. However, the rapid advancement of NLP technologies has transformed socially interactive machines into a tangible reality accessible to the general population. At present, these machines are predominantly utilized by tech-savvy early adopters who possess both the interest and means to engage with them. These models can learn the preferences of specific user groups that interact with them. Consequently, scholars propose addressing the societal concerns surrounding this technology by envisioning the deployment of distinct models for different user groups that align with varying values, thereby allowing dissenting groups the option to opt out of potentially harmful processes [166]. However, this proposed course of action appears increasingly implausible in light of the rapid proliferation of these technologies. We are already witnessing the emergence of numerous applications incorporating conversational interfaces in socially relevant domains, such as chat-based search engines [153], brainstorming companions (e.g., Jasper[1], Copy.ai[2]), creative writing assistants (e.g., Wordtune[3], QuillBot[4]), and an array of other similar tools. People are becoming increasingly immersed in environments that are saturated with such systems. Thus, the question no longer revolves around the mere existence of these systems or how to manage access for specific user groups but rather focuses on how to design and govern their social behaviors to ensure responsible and ethical treatment of all users. The current FATE literature takes a rather critical position on anthropomorphizing dialog systems, warning that it may introduce harms to the end-users, especially related to transparency and (misplaced) trust issues [1]. Therefore, conducting thorough research and giving thoughtful consideration to the appropriate level of anthropomorphizing and endowing conversational systems with social qualities for each specific application is warranted.

---

[1] https://www.jasper.ai/

[2] https://www.copy.ai/

[3] https://www.wordtune.com/

[4] https://quillbot.com/

A Appendix

A.1 Semi-structured interview protocol

Interview and observation guide – Natural language interface for human-computer interaction

Author:
Session length:
Participant Name:
Email:
Time/Date:
Location:

1. Introduction and setup (5 min)

Set up equipment to record sound.

Hello, and thank you for participating in our research!

My name is Ekaterina. I am a PhD student in Human-Computer Interaction Group at EPFL. My main interest for this interview is to understand what defines naturalness in conversations for people and how it may be useful for you if you could converse naturally and ***emotionally*** with the technology. I'll be asking you some questions and my partner, XXX, will be assisting me with the interview. (S)he will be taking some notes.

We are not evaluating you or your opinions in any way.

Here is how the session is going to proceed. This interview will take about 40 minutes, during which we will ask you a few general questions as well as some specific ones.

As we mentioned in the invitation email, we will take an audio-record of our interview, so that

I can go back and review things later and make sure we get everything right. We won't use your name in connection with the recordings or the results. The audiotrack will only be used internally and never shared anywhere with anyone. If you don't like this idea, please let me know. If there's anything you really don't want on the record, even if it's anonymized, please let me know that, too.

How does all that sound to you? Do you have any questions at this point?

2. Demographics and background (5 min)

Understand: gender, age group, nationality, education level / occupation, general technology usage behavior, conversational behavior

General information
Could you tell us your name and tell us a little bit about what you do?
Where did you go to school (university)? Is your current job related to what you studied?
Do you have any hobbies? How do you usually spend your free time?

Information about how participant converses with a friend, family member, or colleague
Do you spend some of your free time talking to friends, family members, and colleagues? What are some of the subjects that you discuss?
Can you tell us the last time you had an intense/deep/prolonged (opposite to small talk) conversation with someone (when a participant in the conversation can "feel" the emotions of one or more other participants)? How did it go?

Information about technology usage
How often do you use your computer the last few days?
Where do you usually use your computer(s)?
How many hours per day would you say you spend on your computer? Do you shop online? How often?
Do you use online banking systems / make payments online? How often?
Do you use online system for other purposes? If so, can you specify?

Do you use a smartphone? How often do you use your smartphone the last few days?
How many hours per day would you say you use your smartphone?
Where do you usually use your smartphone?
What apps do you use most often the last few days? *(Messengers, browsers, fitness, news, camera, music, online banking, weather, notifications, alarm, etc.)*

Follow-up on messengers:
Who do you usually chat with on messengers? What are some subjects that you discuss? Can you tell us the last time you had an intense/deep/prolonged chat with someone?
Do you use your smartphone for any other purposes? If so, can you specify?

3. Warm-up (5 min)

Do you know what a conversational assistant is? Do you know Siri / Google Assistant / Alexa / Alisa (*Russian alternative*)?
Do you ever use it?

3.1. If participant uses a CA:
When did you use it last time? What did you ask it?
Were you satisfied with its response? What did you like? What did you dislike?
How do you find its conversational abilities in general?
Do you think it sounds natural? Why?
Would you like to converse more with it?
Could you recall if you have ever felt any strong emotions while talking to it? How did you find its response? Do you think, can it recognize your emotions?

3.2. If participant knows what a CA is, but doesn't use it:
Why don't you use them?
How do you happen to know about them? Did you ever observe some of your friends or relatives using them? *(Then proceed with questions from either 3.1. or 3.3. subsection)*

3.3. If participant doesn't know what a CA is:
Have you ever encountered an online chatbot? In what context/on which website?
Did you chat with it? What did you ask it? *(Proceed with questions from 3.1. subsection)*

OR (if the participant doesn't have chatbot experience):
How do you usually find the information you need, e.g. news, weather, organization opening hours, something to inform yourself (illness symptoms, ideas for activities, recipes, etc.)?
How do you manage your schedule? What do you do not to forget about important events?
How do you find your way in the street or while driving? *(maps, signs, ask passers-by)*

4. Main questions (20 min)

Overarching #1: What defines naturalness in human emotional dialogs?
We are developing a novel natural language interface for humans to interact with the computer. We want to make it very intuitive, so that people could use it just as if they were talking with another human. ***Most importantly, we want to make it capable of understanding human emotions such as frustration.***

How would you describe naturalness in this type of conversation? /
What should be the features (properties) of this conversation so that you could call it natural (human-like)? /
What would be your expectations from a natural conversation with a computer? /
How would you define the key elements in human-like conversation (with a computer)?

Follow-up, if topic of emotion arises: What is the role of emotions in natural conversation in your opinion?

Overarching #2: What are some of the purposes of emotional chatbots?
Now, suppose you can converse with a computer using ***natural language*** interface (using chatbot, using personal assistant, using intelligent assistant, using voice interface), what types of situations (what types of context, what types of applications) will you find most useful for such interaction experience?

In what situations would you like to use this technology? /
When would you like to have natural conversation with your computer or smartphone? /
What are the best applications for this technology in your opinion? /
In what circumstances would you imagine yourself have natural conversation with the computer or smartphone?

If participant struggles with bringing ideas:

We understand, it might be a bit tricky to imagine such a situation straight away.

Before you mentioned <...>, how would you find natural language interface in this context?

OR:

For example, some other people commented on our natural language interface in the context of <driving / customer service / system to engage with eldery people while they stay alone / household companion / reception desk / restaurant / system to share your thoughts with / personal (schedule) manager>.

How would you find natural language interface in this context?

If positive: Why would you find natural language interface useful? What kind of conversational behavior would you expect from it? What mood / emotions do you usually feel in this situation? How / In what manner would you like the human-like chatbot to respond?

If negative: Why do you think it's not useful in this case? How do you usually deal with this type of situation?

Concerns
In general, how do you feel/think about the natural language chatting system?
In the contexts/interaction scenarios/applications that you described, do you have any concerns about it?
Are there any contexts when you would feel restrained from using the system? Why?
Are there any contexts for which you think this system is absolutely not necessary?

5. Wrap-up and closing (5 min)

This brings us to the end of the interview. I've been asking you a lot of questions. Is there anything you want to ask me?

If anything else occurs to you after I leave, please don't hesitate to let me know by email. I may be in touch with you again to ask a few follow-up questions. If you'd like, I can send a version of the report that we'll write based on this interview.

Thanks again for all your time! Please, take this small token of appreciation for your involvement.

A.2 Examples from empathetic question taxonomy

Tables A.1 (acts) and A.2 (intents) present the two EQT branches with examples for each label. Examples are selected from the initial manually annotated subset. For each label we include its frequency for the three corresponding sets: manually-labeled, Mturk-labeled, and overall (both manually-, Mturk-, and automatically-labeled). The frequencies are approximately the same across each label, which validates that our annotation methods produced credible results. Examples of automatically assigned labels are given in Appendix A.5.

Table A.1: Classification of question acts with corresponding definitions and examples. Under each label its frequency is given for the three corresponding sets: manually labeled, Mturk labeled, and overall.

Question Act	Definition and Example
Request information 38.7%, 52.5%, 51.4%	Ask for new factual information. - *when i left my family to study in another city i got upset.* - *I'm sorry to hear that.* ***What are you studying?***
Ask about consequence 21.0%, 19.2%, 17.9%	Ask about the result of the action or situation described by the speaker. - *Our home was broken into* - *Oh no!* ***Did they steal a lot?***
Ask about antecedent 17.1%, 10.5%, 11.3%	Ask about the reason or cause of the event or state described by the speaker. - *Hi, I had a great vacation but something went wrong* - *Oh no, I'm sorry to hear that.* ***What happened?***
Suggest a solution 8.7%, 5.7%, 8.0%	Provide a specific solution to a problem in a form of a question. - *I lost my favorite jacket and I can't find it* - ***did you try redoing your steps of the last day?***
Ask for confirmation 5.8%, 5.6%, 5.2%	Ask a question to confirm or verify the listener's understanding about something that has been described by the speaker. - *I applied for a job last week.* - ***Oh did you?***
Suggest a reason 5.2%, 3.7%, 4.1%	Suggest a specific reason or cause of the event or state described by the speaker in a form of a question. - *i felt scared walking home alone the other day.* - *That's terrible!* ***Were you in a bad part of town or anything?***

Continuation of Table A.1

Question Act	Definition and Example
Positive rhetoric 1.0%, 1.3%, 1.1%	Ask a question in order to make an encouraging statement or demonstrate agreement with the speaker about a positive point without expecting an answer. *- I couldn't pay for all my groceries and someone came up from the line behind and paid for the rest. I was so touched!* *- **Wow, how amazing is that!?***
Negative rhetoric 1.3%, 1.1%, 0.8%	Ask a question in order to express a critical opinion or validate a speaker's negative point without expecting an answer. *- I swear my friend is always using me* *- **that sucks is she really your friend then?***
Irony 1.3%, 0.3%, 0.2%	Ask a question using words that suggest the opposite of what the listener intends, usually to be humorous or pass judgement. *- I ate 10 Big Macs the other day.* *- oh my lord! **only ten?***

Table A.2: Classification of question intents with corresponding definitions and examples. Under each label its frequency is given for the three corresponding sets: manually labeled, Mturk labeled, and overall.

Question Intent	Definition and Example
Express interest 57.1%, 55.2%, 60.2%	Express the willingness to learn or hear more about the subject brought up by the speaker; demonstrate curiosity. *- I just applied for a higher paying position within my company.* *- **That's cool, what is the position?***
Express concern 20.3%, 20.3%, 23.4%	Express anxiety or worry about the subject brought up by the speaker. *- I cry every time I think of my sister.* *- Why?? **what happened to her!?***
Sympathize 3.9%, 7.3%, 5.1%	Express feelings of pity and sorrow for the speaker's misfortune. *- my girlfriend cheated on me* *- Oh no! **How did you find out?***
Offer relief 4.8%, 3.2%, 4.5%	Reassure the speaker who is anxious or distressed. *- They stopped making donuts at my favorite bakery.* *- Oh no! **Can you get donuts somewhere else?***
Amplify excitement 1.9%, 4.7%, 2.3%	Reinforce the speaker's feeling of excitement. *- lol. Going on vacation to Florida in a couple weeks!* *- Wow that's awesome! **To the beach?***

Continuation of Table A.2

Question Intent	Definition and Example
Support 2.6%, 1.8%, 1.0%	Offer approval, comfort or encouragement to the speaker, demonstrate interest in and concern for the speaker's success. *- I studied so hard for my test.* - ***I hope you did well?***
Amplify joy 1.6%, 1.7%, 0.9%	Reinforce the speaker's glad feeling such as pleasure, enjoyment, or happiness. *- I just received my certification to teach english as a second language!* *- Congrats!!!* ***Do you already have a job lined up?***
Amplify pride 2.6%, 1.7%, 0.7%	Reinforce the speaker's feeling of pride. *- My nephew caught a huge bass this weekend!* - ***That is cool, did you teach him how to fish?***
De-escalate 1.6%, 1.6%, 0.7%	Calm down the speaker who is agitated, angry or temporarily out of control. *- My neighbor threw their nasty trash all over their yard and won't clean it up! It's sooo gross!* *- Oh, that's disgusting!* ***Have you tried to talk to them about it?***
Moralize speaker 1%, 0.6%, 0.6%	Judge the speaker. *- I broke my TV remote and i blamed it on my kid* *- That's kinda terrible.* ***Did you apologize to him?***
Pass judgement 1.6%, 1.2%, 0.5%	Express an opinion (especially critical) about the subject brought up by the speaker. *- I hope the government can give some free course about the benefit of staying calm and healthy* - ***Government?*** *No way, it is interested in quite the opposite my friend.*
Motivate 1%, 0.5%, 0.2%	Encourage the speaker to move onward. *- This weekend is so boring so far* *- yeah? nothing interesting whatsoever?* ***why not make it exciting yourself?***

A.3 Details about Mturk task for ED annotation with EQT labels

A.3.1 Dialog pre-processing

Throughout our study, we only used those ED dialogs that contained questions in at least one listener turn. Since one dialog could contain several listener questions, for all downstream annotation tasks each such dialog was split into several separated dialogs, equal to the number of listener questions. The resulting sub-dialogs were truncated such that they would end with the particular question to which they corresponded to allow labeling every question in each dialog, without losing the previous conversational context. Figure A.1 shows an example of a dialog from the original ED dataset and the resulting dialogs after the split.

In the Mturk interface, if the given listener turn contained multiple questions, we showed the resulting sub-dialogs in the same page one after another for contextual consistency. But if the original dialog contained listener questions in several turns, we showed the resulting dialogs in the two separate pages. Using the example from Figure A.1, we would show the first resulting dialog in one page and the last two resulting dialogs together in another page.

Original dialog

Speaker:	– You are never going to believe what I did!
Listener:	– What did you do?
Speaker:	– Well, I normally do not feel comfortable lending things to my friends, but recently I mustered up the trust to loan my friend my vehicle.
Listener:	– Ouch… Is it just for a day? Is your friend a safe driver?

Resulting dialogs

Speaker:	– You are never going to believe what I did!
Listener:	– What did you do?

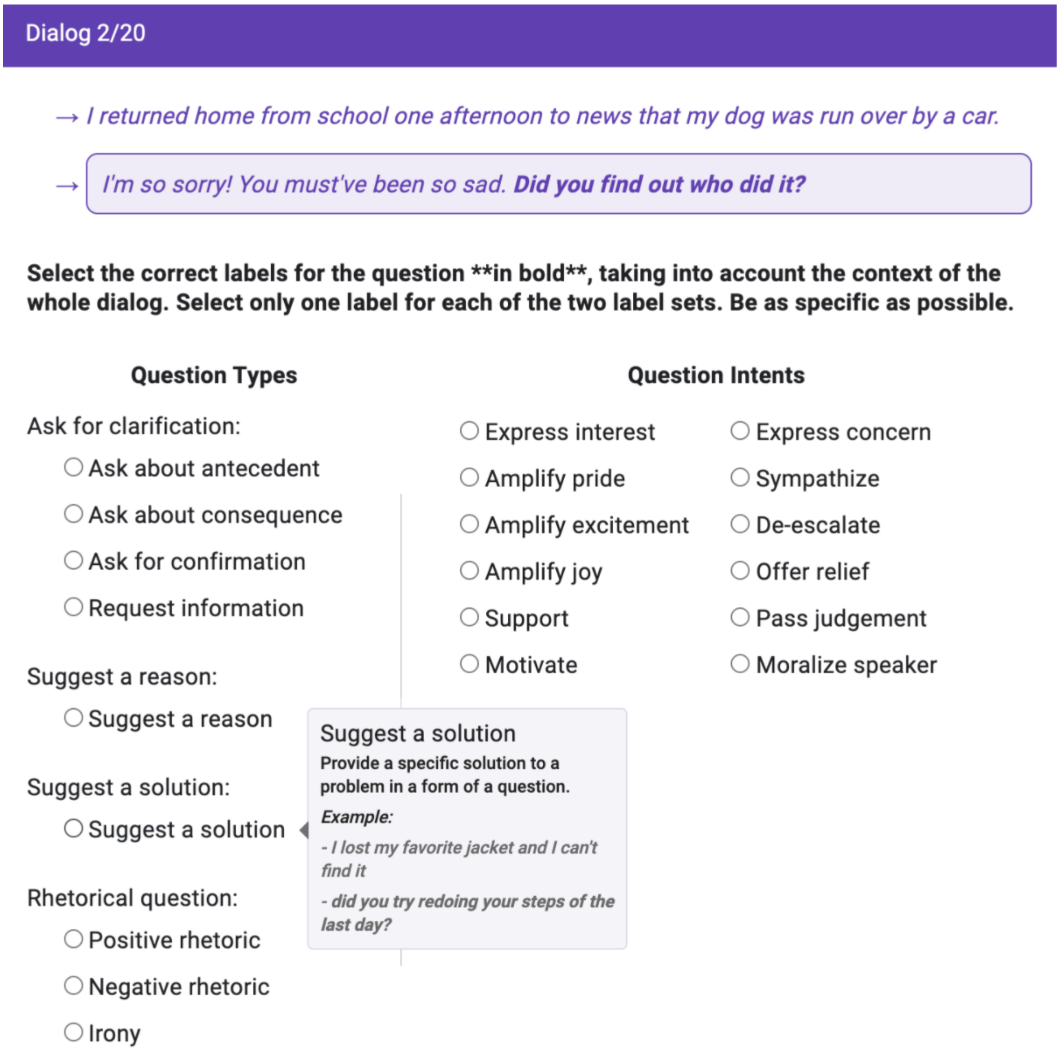
Speaker:	– You are never going to believe what I did!
Listener:	– What did you do?
Speaker:	– Well, I normally do not feel comfortable lending things to my friends, but recently I mustered up the trust to loan my friend my vehicle.
Listener:	– Ouch… Is it just for a day?

Speaker:	– You are never going to believe what I did!
Listener:	– What did you do?
Speaker:	– Well, I normally do not feel comfortable lending things to my friends, but recently I mustered up the trust to loan my friend my vehicle.
Listener:	– Ouch… Is it just for a day? Is your friend a safe driver?

Figure A.1: Original and resulting dialogs after pre-processing.

A.3.2 Task user interface

The user interface for the annotation task is illustrated in Figure A.2.



Dialog 2/20

→ *I returned home from school one afternoon to news that my dog was run over by a car.*

→ *I'm so sorry! You must've been so sad.* ***Did you find out who did it?***

Select the correct labels for the question **in bold, taking into account the context of the whole dialog. Select only one label for each of the two label sets. Be as specific as possible.**

Question Types

Ask for clarification:
- ○ Ask about antecedent
- ○ Ask about consequence
- ○ Ask for confirmation
- ○ Request information

Suggest a reason:
- ○ Suggest a reason

Suggest a solution:
- ○ Suggest a solution

Rhetorical question:
- ○ Positive rhetoric
- ○ Negative rhetoric
- ○ Irony

Question Intents

- ○ Express interest
- ○ Amplify pride
- ○ Amplify excitement
- ○ Amplify joy
- ○ Support
- ○ Motivate
- ○ Express concern
- ○ Sympathize
- ○ De-escalate
- ○ Offer relief
- ○ Pass judgement
- ○ Moralize speaker

Suggest a solution

Provide a specific solution to a problem in a form of a question.

Example:

- I lost my favorite jacket and I can't find it

- did you try redoing your steps of the last day?

Figure A.2: The user interface of the Mturk crowd-sourcing annotation task.

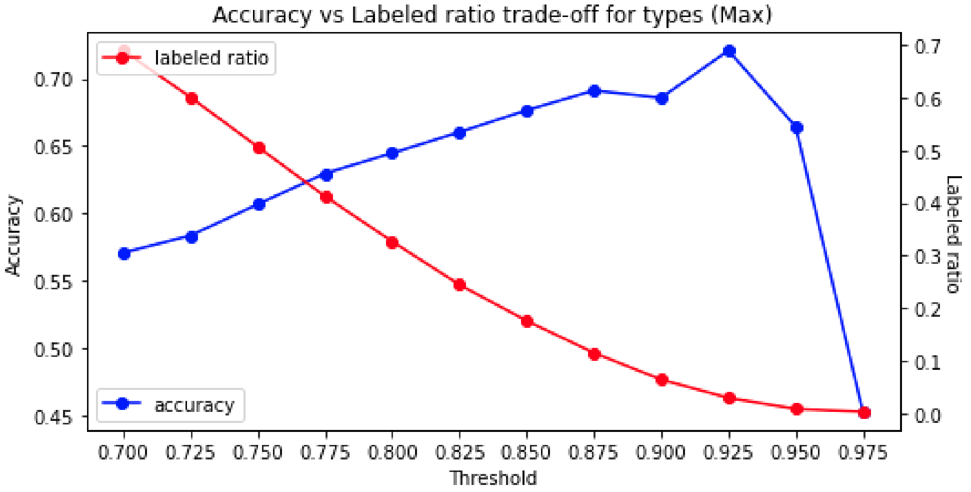
A.4 Details about EQT-labeled data augmentation with lexical similarity

A.4.1 Setup and results

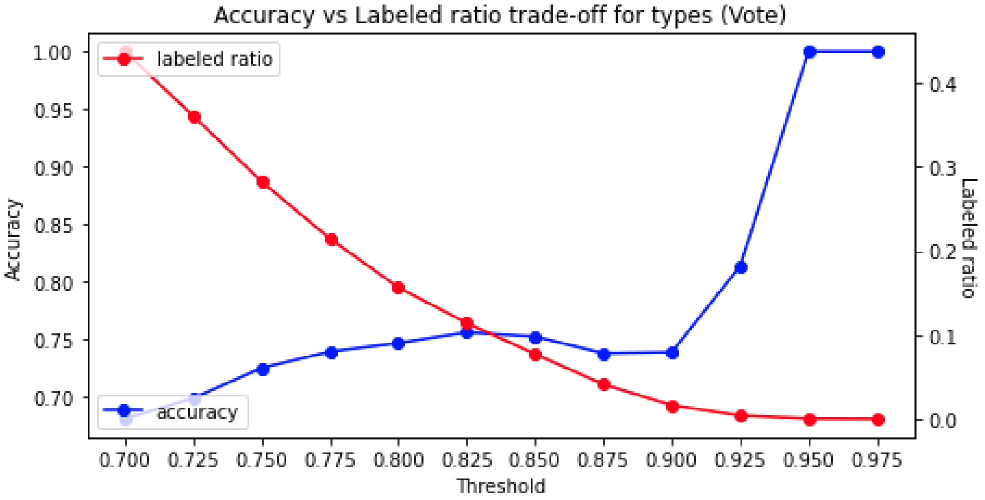
We used a half-decaying weighting scheme to encode questions with preceding context for the data augmentation process. The highest weight was always assigned to the final question to give it a higher preference. For example, if the dialog context consisted of three turns with embeddings e_1, e_2, e_3 and the fourth turn was a listener's question with embedding e_4^*, the final dialog embedding was $(8/15)e_4^* + (4/15)e_3 + (2/15)e_2 + (1/15)e_1$.

Figures A.3 and A.4 demonstrate the results of cross-validation runs for question acts and question intents for the Nearest-Neighbor label propagation approach. For each label set, we experimented with two similarity strategies: taking the same label as the top-1 most similar dialog according to the cosine similarity (*Max*, included in sub-figures A.3a and A.4a) and identifying the label with the majority vote from the top-3 most similar dialogs (*Vote*, included in sub-figures A.3b and A.4b). For each cross-validation launch, we conducted a grid search over cosine-similarity thresholds in a range between 0.7 and 1.

We also tried concatenating one-hot-encoded emotional context vectors with the dialog embeddings before running the cross-validation, but it did not result in any improvement in the accuracy and the resulting plots were almost identical to Figures A.3 and A.4, so we decided not to proceed with this approach.

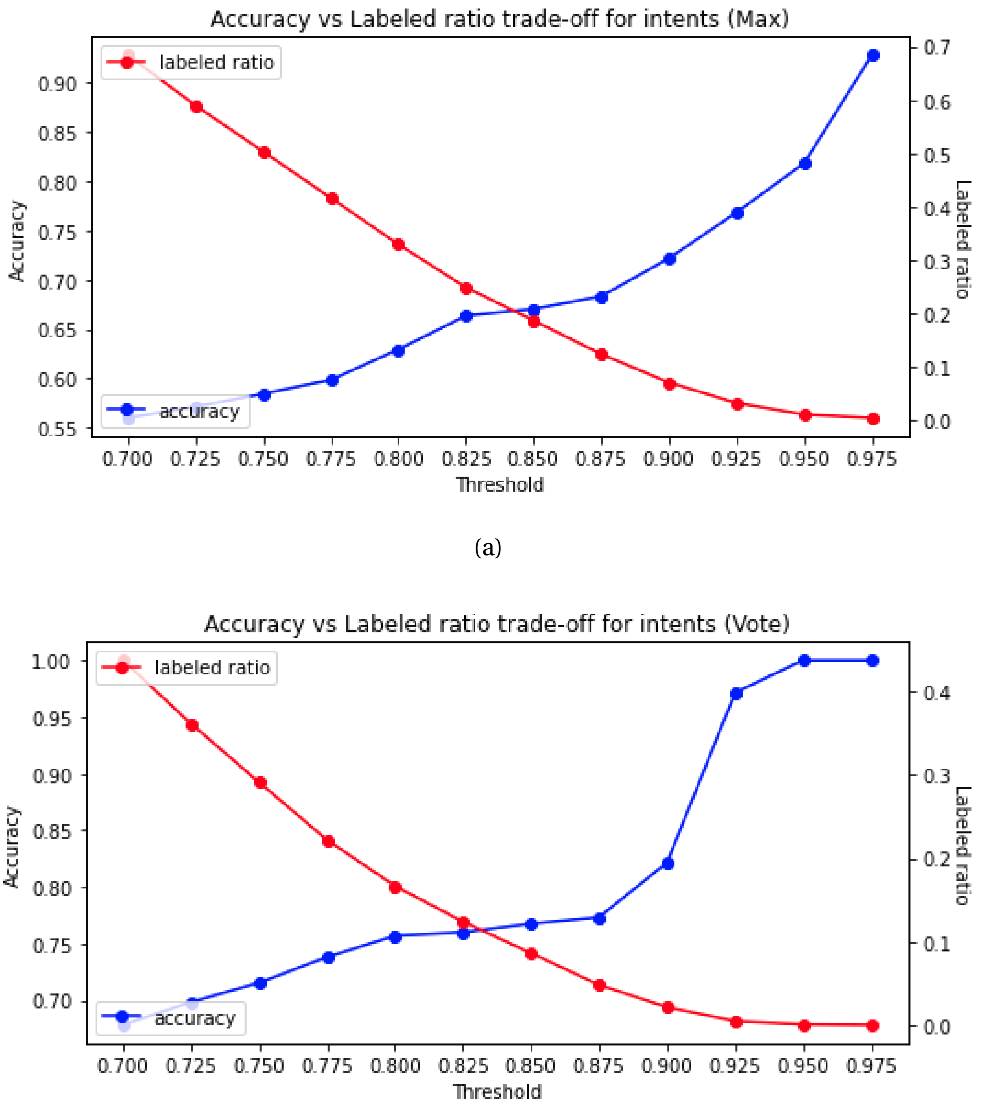


(a)



(b)

Figure A.3: Cross-validation results for question acts for the two considered strategies: *Max* in sub-figure A.3a and *Vote* in sub-figure A.3b.



(a)

(b)

Figure A.4: Cross-validation results for question intents for the two considered strategies: *Max* in sub-figure A.4a and *Vote* in sub-figure A.4b.

A.4.2 Examples of annotated questions

Table A.3 presents several examples of propagated labels obtained using the outlined data augmentation process to give a better idea on the accuracy of this approach.

Table A.3: Examples of propagated labels obtained using majority vote from the top-3 Nearest-Neighbor (NN) dialogs according to cosine similarity. The first column includes the newly annotated question, and the other three show the top-3 NN dialogs with respective question labels and a similarity value. Spelling and punctuation of the original source have been preserved.

Annotated question	Top-1 NN	Top-2 NN	Top-3 NN
– I get a good feeling when I think back to a birthday I had when I was a kid and all of my friends and I got to see a really funny movie at the mall. – *Awww!* ***What movie did you go to see?*** (Request information, Express interest)	*– I went to the movies by myself yesterday. I have no friends.* – ***what movie did you see?*** (0.87: Request information, Express interest)	*– I was happy when we were going to a new movie last weekend. I had waited all summer for it* – ***What movie was it?*** (0.87: Request information, Express interest)	*– I'm going to see a film tonight at the cinema.* – *oh really?* ***what movie?*** (0.86: Request information, Express interest)
– It really sucked, since a month ago I was dating this girl and she dumped me so early on. – *I'm so sorry.* ***Are you okay?*** (Request information, Express concern)	*– I hurt me when my parents got divorced. I never thought that would happen* – ***I'm so sorry, are you okay?*** (0.92: Request information, Express concern)	*– I am really feeling bad* – *I'm so sorry!* ***Is everything ok?*** (0.90: Request information, Express concern)	*– I just found out that my girlfriend has been cheating on me. God this is the worst week of my life.* – *I feel really sorry for you.* ***Will you be okay?*** (0.84: Request information, Express concern)
– One time my mom bought an ice cream from Mcdonalds! – ***Really?*** (Ask for confirmation, Express interest)	*– I saw someone putting mayo on their ice cream.* – ***Really?*** (0.92: Ask for confirmation, Express interest)	*– I accidentally ate someone else's cake at work* – ***Really?*** (0.91: Ask for confirmation, Express interest)	*– I just ate 5 donuts by myself* – ***Really?*** (0.86: Negative rhetoric, Express interest)
– i was scared walking home last night – ***Why was you scared was it too dark?*** (Suggest a reason, Express concern)	*– I used to be so scared to go to sleep as a kid.* – *How come?* ***Were you scared of the dark?*** (0.92: Suggest a reason, Express concern)	*– I stay away from the dark.* – *Why do you do that?* ***Are you scared of the dark?*** (0.86: Suggest a reason, Sympathize)	*– i was scared walking home the other day* – ***Why were you scared?*** (0.83: Ask about antecedent, Express concern)

Continuation of Table A.3

Annotated question	Top-1 NN	Top-2 NN	Top-3 NN
– *I one time lost my trunks in the pool! People saw me in a way I didn't want!* – *Oh no! That must have been super embarrassing!* ***How did you react to that?*** (Ask about consequence, Sympathize)	– *a girl i like at school told me today she doesn't like me in front of everyone* – *Oh no! That must have been really embarrassing!* ***How did you respond?*** (0.85: Ask about consequence, Sympathize)	– *I fell down on stage while dancing, I felt so bad.* – ***oh dear, that must've been embarrassing, are you okay though?*** (0.84: Ask about consequence, Sympathize)	– *Once at a swimming competition, I had a wardrobe malfunction in front of a lot of people* – *Oh my goodness, that must have been humiliating.* ***What did you do?*** (0.83: Ask about consequence, Sympathize)
– *My neighbor died in a car crash.* – *Oh my. I'm so sorry to hear that.* ***What happened?*** (Ask about antecedent, Sympathize)	– *My nephew died yesterday.* – *I am so sorry to hear that.* ***What happened?*** (0.89: Ask about antecedent, Sympathize)	– *My pet ferret Fuzzy died the other day. I was so heart-broken.* – *I'm so sorry to hear that.* ***What happened?*** (0.88: Request information, Sympathize)	– *When my pet died I felt liek I lost my family member, My best friend.* – *Im sorry to hear that.* ***What happened?*** (0.88: Ask about antecedent, Sympathize)
– *My brother just turned 16 and he's about to get his first car! I'm so excited for him.* – *Whoa that's exciting!* ***What kind of car we looking at?*** (Request information, Amplify excitement)	– *I can't wait! We just bought a car today! Going to pick it up soon!* – *Oh nice! That is exciting!* ***What kind of car did you get?*** (0.89: Request information, Amplify excitement)	– *I just bought a brand new car* – *How exciting!* ***What kind of car is it?*** (0.86: Request information, Amplify excitement)	– *I was surprised when my dad got me my first car. I was not expecting it* – *That must have been exciting for you.* ***What car was it?*** (0.85: Request information, Amplify excitement)
– *I spent hours reviewing notes and course content to prepare myself for a few trials that a company wanted me to go through.* – *Good job!* ***Do you feel pretty prepared?*** (Request information, Support)	– *I have an important job interview this week* – ***Have you prepared well for it?*** (0.85: Request information, Express interest)	– *I have been studying for my final math exam all week long.* – *I hope you do well on it!* ***Do you feel prepared?*** (0.83: Ask for confirmation, Support)	– *Ive got a big interview on Friday. It for a job I really want.* – *I hope it goes well!* ***are you prepared?*** (0.83: Request information, Support)

Continuation of Table A.3

Annotated question	Top-1 NN	Top-2 NN	Top-3 NN
– *Friends threw me a surprise party yesterday.* – *thats awesome, and happy birthday !!!* – *Thanks! I got so many cool gifts! I was so happy.* – ***what kind of gifts did you get?*** (Ask about consequence, Amplify excitement)	– *I was happy to find that at work my coworker prepared a birthday party for me. I was not expecting it.* – *Wow. I bet that was a nice surprise.* ***Did you get a lot of presents?*** (0.84: Ask about consequence, Amplify excitement)	– *My friends threw me a surprise birthday party last year!* – *That is very nice* – *It was! I was shocked and I felt very loved.* – ***Did they brought any special gift?*** (0.84: Request information, Express interest)	– *My friends planned a surprise party for my birthday.* – *Exciting!* ***Did you get any neat gifts?*** (0.84: Ask about consequence, Amplify excitement)
– *I'm living my best life. I could'not be any happier.* – *good to know.* ***and what makes your life so good, huh?*** (Request information, Amplify joy)	– *I am so happy with my life right now.* – *You sound very content.* ***What makes you happy?*** (0.86: Request information, Express interest)	– *I feel good. Everything finally seems to be working out.* – *That's great!* ***What are some things you're enjoying about life right now?*** (0.86: Request information, Amplify joy)	– *I've been happy with the way things have been going in my life lately.* – ***That's awesome, glad to hear, what are you most happy with?*** (0.86: Ask about antecedent, Amplify joy)
– *I was happy when my brother finished school. I was proud of him* – *That is awesome.* ***Was it high school or college?*** (Request information, Amplify pride)	– *It felt great to see my son graduate. Like I succeeded as a parent.* – *That's awesome.* ***high school?*** (0.88: Request information, Amplify pride)	– *I use to be the number one tennis player in the state.* – *That is an awesome achievement!* ***Was it for high school or college?*** (0.86: Request information, Amplify pride)	– *I'm a Phd student and I'm taking a really hard class. I have to do well so I was really happy when I got an A on a test!* – *thats awesome!* ***what college you go to?*** (0.84: Request information, Express interest)

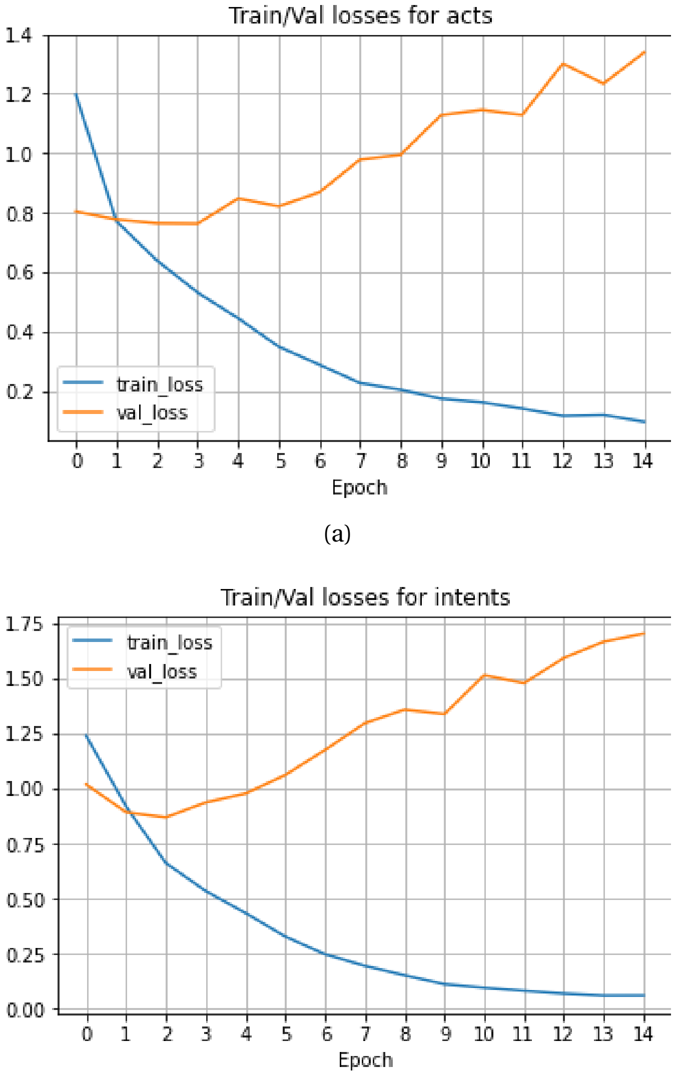
Continuation of Table A.3

Annotated question	Top-1 NN	Top-2 NN	Top-3 NN
– *I cheated at cards.* – ***Did you feel bad about it?*** (Ask about consequence, Moralize speaker)	– *I cut someone off in traffic today* – ***Do you feel bad about it?*** (0.85: Ask about consequence, Moralize speaker)	– *Yesterday, i had a night out with my friends, but i lied to partner that i will be staying late for work. I did not want to see her nagging* – *That's really not good.* ***Did you feel bad about it?*** (0.85: Negative rhetoric, Moralize speaker)	– *I was really hungry today and ate my roomates' leftovers.* – ***Do you feel bad about it?*** (0.85: Ask about consequence, Moralize speaker)
– *I stole money from my friend.* – *oh..* ***why did you do that?*** (Ask about antecedent, Pass judgement)	– *I stole money from my son's piggy bank.* – ***Why did you do that?*** (0.94: Ask about antecedent, Pass judgement)	– *I stole money from someone at a party years ago and I still feel bad about it.* – ***Why did you do that?*** (0.91: Ask about antecedent, Pass judgement)	– *I told my best friends secret to another one of our friends.* – ***Why did you do it?*** (0.89: Ask about antecedent, Pass judgement)

A.5 Details about training automatic QBERT classifiers

For our automatic classifiers, we used GELU as a hidden activation function and applied a 0.1 dropout to all layers and attention weights. For training, we used Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1 \times 10^{-6}$, and the peak learning rate of 2×10^{-5}. The maximum number of input tokens was set to 100, and we used the batch size of 50. The evolution of train and validation losses over the course of 15 training epochs is shown in Figure A.5. We used Google Colab environment for the training.

The performance of classifiers trained only on a human-annotated subset was several percent



(a)

(b)

Figure A.5: Train and validation losses over the course of approximately 15 training epochs for question acts (sub-figure A.5a) and question intents (sub-figure A.5b).

lower than training on augmented data (see Section 5.6.2), resulting in 75% accuracy for acts and 70% for intents on the same (human-annotated) test set. Therefore, in this paper, we focus on the results obtained with the augmented data.

Figure A.6 demonstrates several examples of automatically labeled questions in the ED dialogs. We specify both the predicted act and intent labels for each listeners' question and emotions expressed by speakers in each turn to observe how they are influenced by listeners' questions. Here we combine the pre-processed dialogs (cf. Section A.3.1) back to their original format, which explains why some labeled questions appear in the middle of the dialogs.

– *I am proud of my girlfriend for getting a full time job, I am sure she will do great!* (Positive)
– *That's awesome i bet she will too!* ***when does she start?*** (Request information, Express interest)
– *She starts in exactly a week* (Positive)
– ***woo hoo so you guys going out to celebrate?*** (Ask about consequence, Amplify excitement)

– *I am so happy to be having a boy* (Positive)
– *That's great! Congratulations!* ***Is this your first child?*** (Request information, Amplify joy)
– *Thanks. Yes it is. I already got a crib and baby bath.* (Neutral)

– *My daughter scored the winning goal at her last soccer game. I was so happy that all her hard work paid off!* (Positive)
– *That's great.* ***Does she practice a lot?*** (Request information*, Amplify pride)
– *Yes, she practices almost every day after school with her friends and also with her team. She says she will be a professional player one day!* (Positive)

– *Man.....my cat died:(I feel horrible.* (Negative)
– ***That's awful, how did your cat die?*** (Ask about antecedent, Sympathize)
– *Old age. she had a good life but it's still tearing me up.* (Neutral)

– *I took a test last week that I had studied very hard for. I know I got most of the answers right, but I got a failing grade* (Negative)
– *Must've been a really difficult exam.* ***Will there be other exams to balance it out?*** (Ask about consequence*, Offer relief*)
– *The person sitting next to me copied my answers, so the teacher failed both of us.* (Negative)
– ***I guess the teacher wasn't going to listen to you?*** (Suggest a reason, De-escalate) *That sucks.*

– *I ordered a gift for a friend and it says it was delivered but I never received it. Now the company says it takes 14 days for a refund.* (Negative)
– Don't you hate how "customer service" has no service anymore? (Negative rhetoric, Sympathize) ***Did you get the refund at least?*** (Suggest a solution, Offer relief)
– *Still waiting..... That's the most upsetting. Because they waste no time taking your money* (Negative)

– *I didn't realize that stealing was bad until I realized how it made me feel afterwards* (Negative)
– *So you probably felt pretty guilty huh.* ***Did you return what you stole?*** (Ask about consequence, Moralize speaker)
– *No, I was scared to get charged, but I stopped after that* (Neurtal)

Figure A.6: Examples of questions labeled automatically with QBERT. Question acts and intents marked with a star* were annotated by Mturk workers.

A.6 Extended analysis of questioning strategies

A.6.1 Mapping of emotions and empathetic intents

Table A.4 presents the mapping of 32 emotions [187] and 9 empathetic intents [239] to three coarser emotion categories of different valence, which we used to produce visualizations for the analysis.

Table A.4: Mapping of 32 emotions and 9 empathetic intents describing the EmpatheticDialogues dataset to three emotion categories of different valence.

Category	Mapped emotions and intents
Positive:	trusting, surprised, caring, content, joyful, excited, anticipating, hopeful, prepared, nostalgic, impressed, faithful, confident, proud, grateful
Neutral:	neutral, encouraging, agreeing, suggesting, acknowledging, sympathizing, wishing, consoling, questioning
Negative:	devastated, afraid, apprehensive, terrified, disappointed, disgusted, lonely, anxious, sad, embarrassed, annoyed, furious, ashamed, angry, sentimental, guilty, jealous

A.6.2 Additional plots for human-labeled subset

Figures A.7 and A.8 show the breakdown of flow rates between speakers' emotions and listeners' questioning strategies (Figure 5.2) into separate mappings for acts and for intents, respectively.

A.6.3 Analysis of questioning strategies on the whole dataset

For completeness, we include the same analytical visualizations as presented in Section 5.7 for the whole ED dataset (Figures A.9, A.10, A.11, and A.12). From these figures, one can observe higher presence of more "general" categories (*Request information*, *Express interest*), which presumably originates from the fact that QBERT classifiers are slightly biased towards these classes due to the class imbalance in the training data.[1] Nevertheless, despite this remark, other major patterns revealed by the analysis of human-annotated subset (cf. Section 5.7), preserve in the figures produced for the whole ED dataset (including automatically-annotated questions).

---

[1] One possible way to overcome the class imbalance issue in future work is to use the weighted loss function for training.

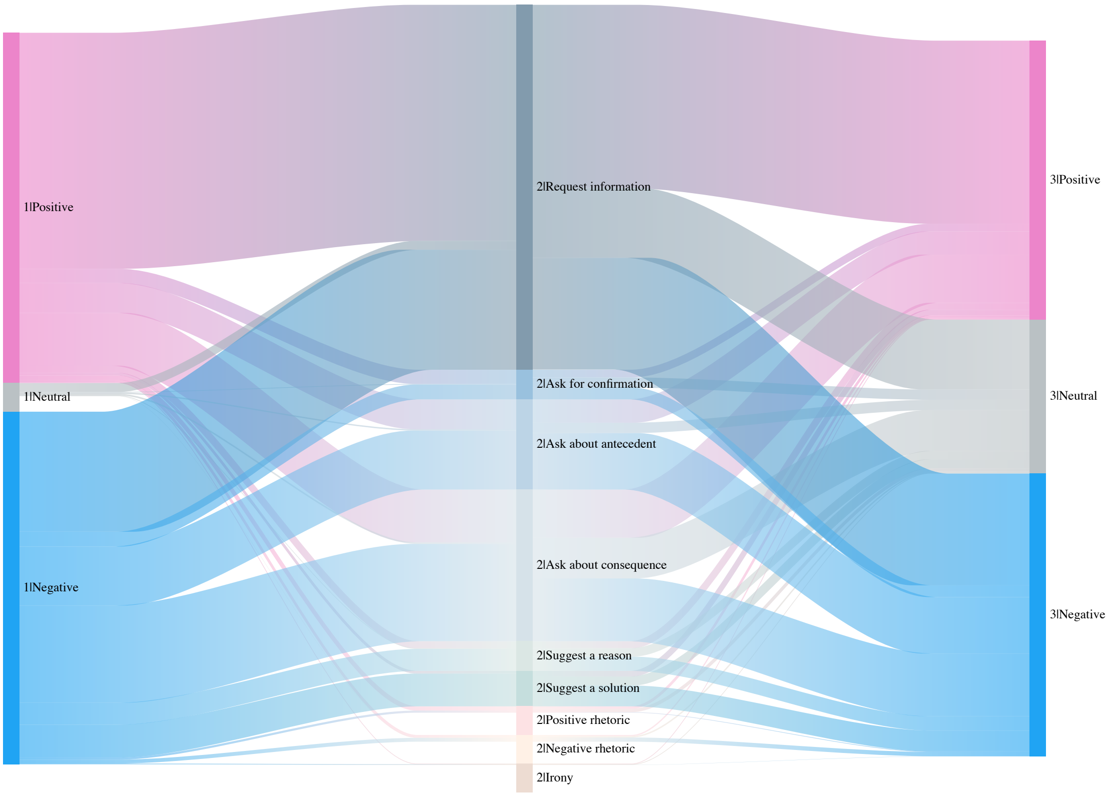


Figure A.7: Mappings between emotions disclosed by the speakers and question acts used by listeners in the first three turns of the ED dialogs (human-labeled ED subset).

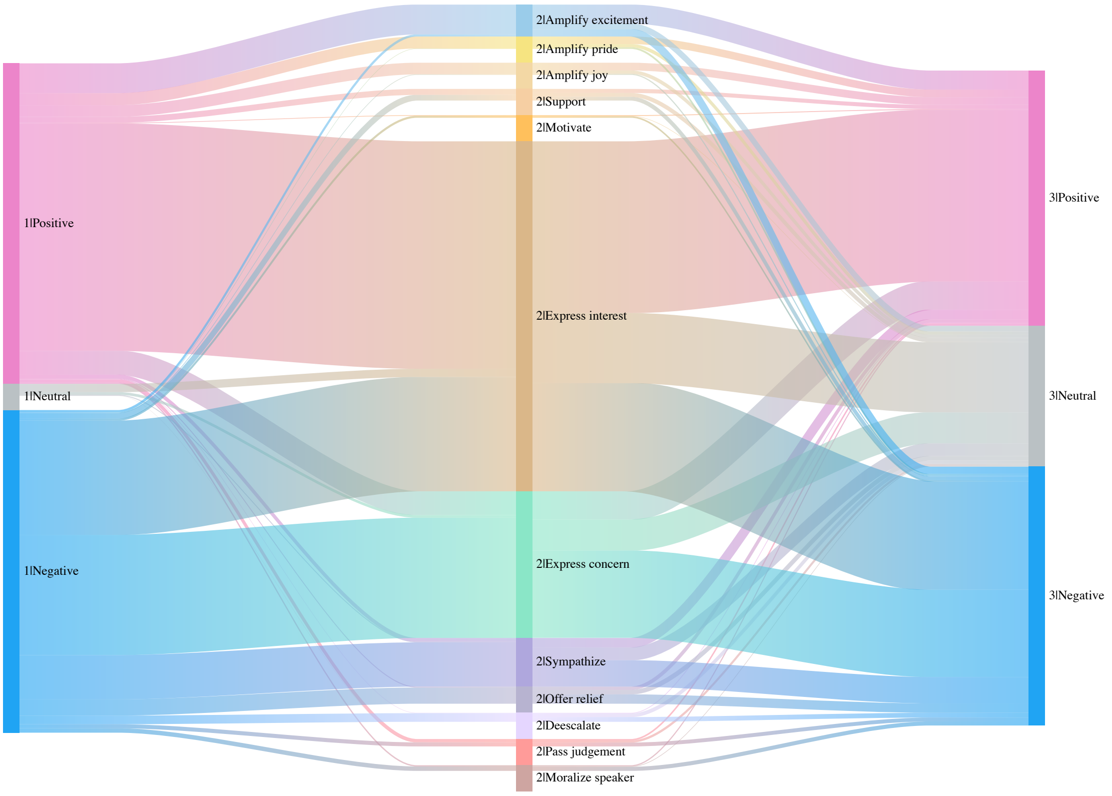


Figure A.8: Mappings between emotions disclosed by the speakers and question intents used by listeners in the first three turns of the ED dialogs (human-labeled ED subset).

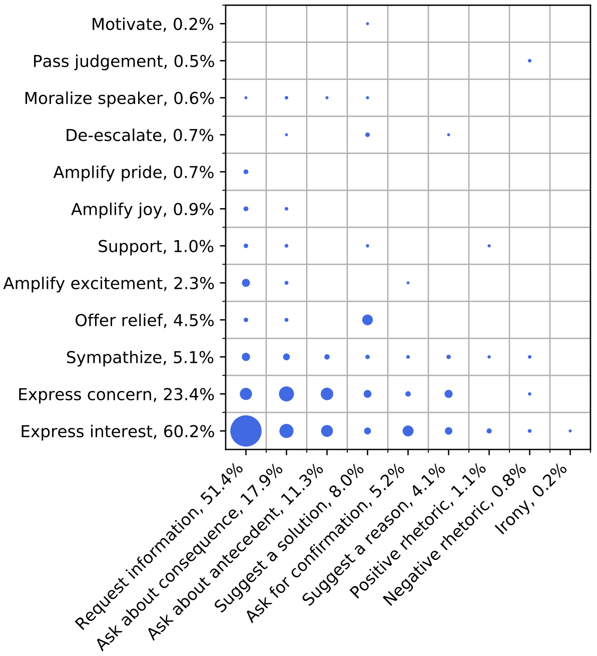
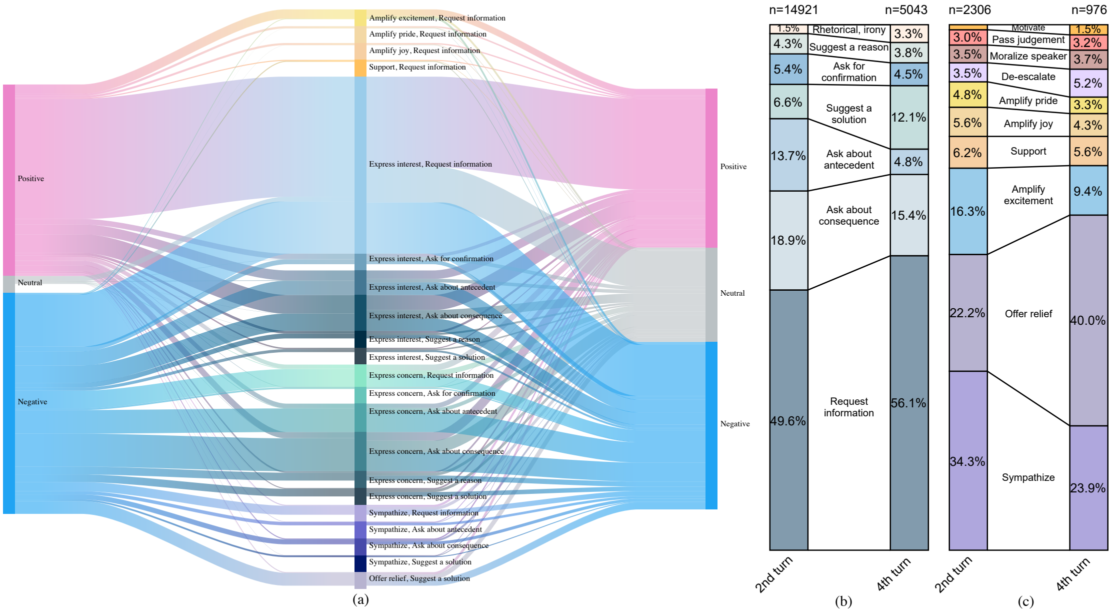


Figure A.9: Joint distribution of question intents and acts for 20,201 labeled questions (whole ED dataset). Blue circles are proportional to the frequency of each pair's co-occurrence.



(a) (b) (c)

Figure A.10: a) Mappings between emotions disclosed by the speakers and listeners' questioning strategies in the first three turns of the ED dialogs (whole ED dataset). b) Frequency distribution of question acts across dialog turns (whole ED dataset). c) Frequency distribution of question intents across dialog turns. Two prevalent intents were excluded for visual clarity; their percentage rates computed for all questions (n=14921 and n=5043) are: *Express interest*: 59.7% → 61.1%, *Express concern*: 24.9% → 19.3%

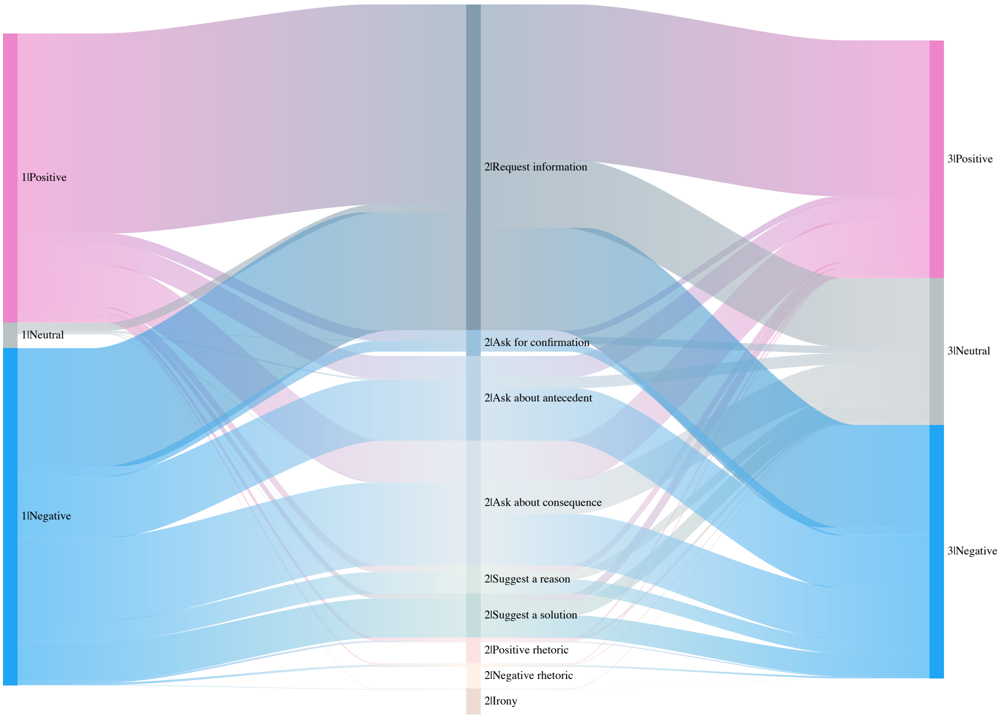


Figure A.11: Mappings between emotions disclosed by the speakers and question acts used by listeners in the first three turns of the ED dialogs (whole ED dataset).

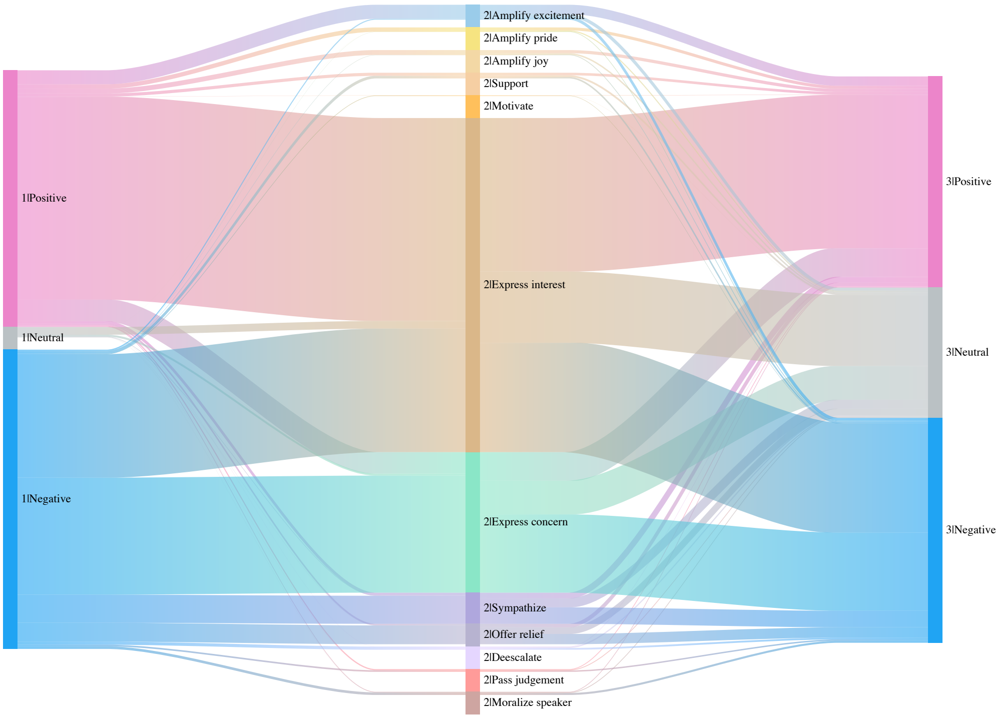


Figure A.12: Mappings between emotions disclosed by the speakers and question intents used by listeners in the first three turns of the ED dialogs (whole ED dataset).

A.7 Topic clusters in EmpatheticDialogues

While working with the EmpatheticDialogues dataset [187], we noticed that many dialogs appear repetitive in terms of the situational scenarios brought up by the speakers. To examine it more closely, we used Sentence Transformers framework [192] to compute vector embeddings of first speakers' turns in all dialogs and cluster them according to cosine-similarity. Figure A.13 shows the empirical cumulative distribution function of topic cluster sizes in the train set of EmpatheticDialogues. From the figure, it can be seen that clusters with between 30 and 130 similar situation descriptions per cluster comprise almost 20% of the training data.

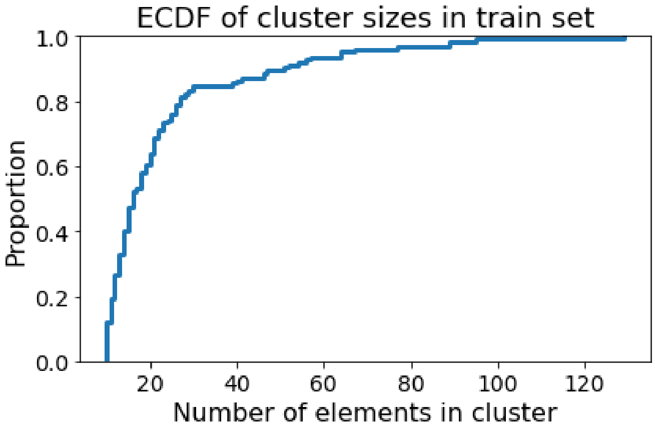


Figure A.13: Empirical cumulative distribution function of topic cluster sizes in the train set of EmpatheticDialogues dataset [187].

A.8 Emotion distribution in grounding scenarios for iEval experiment

Figure A.14 shows the distribution of original emotional labels from the EmpatheticDialogues dataset [187] in 480 grounding scenarios used for our benchmarking experiment. To demonstrate the even coverage of the whole emotional spectrum, we mapped 32 emotions from the dataset to 14 emotions from Plutchik's wheel [179] (8 basic and 6 intermediate emotions) and color-coded the bars in Figure A.14 according to these 14 categories.

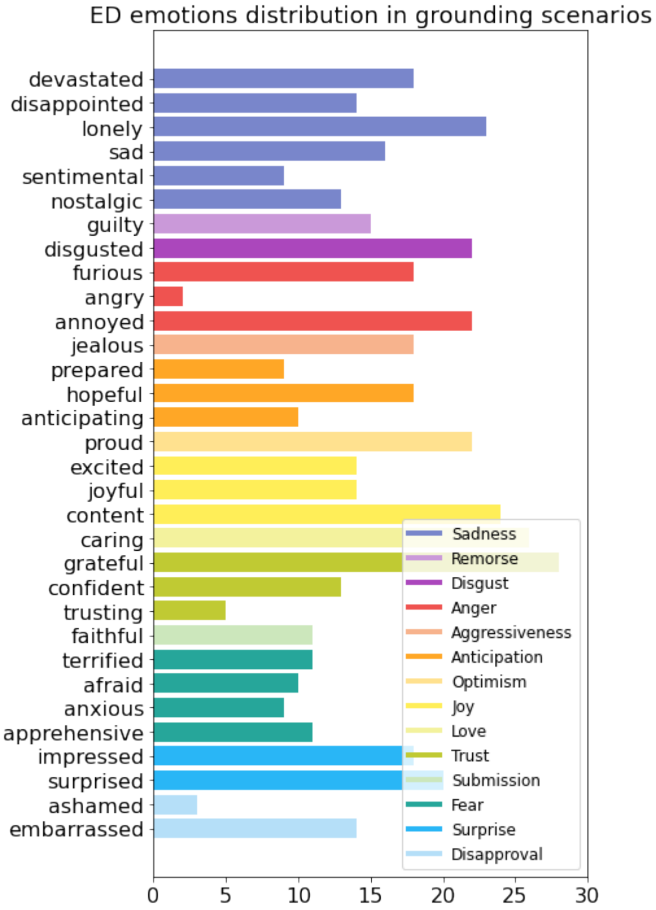


Figure A.14: Distribution of emotional labels from EmpatheticDialogues dataset in grounding scenarios. The legend shows the mapping between the colors and 14 emotional categories from Plutchik's wheel [179] (8 basic and 6 intermediate emotions).

A.9 Additional details about chatbots' responses in iEval experiment

Figure A.15 depicts the average number of tokens in chatbots' responses over three dialog turns.

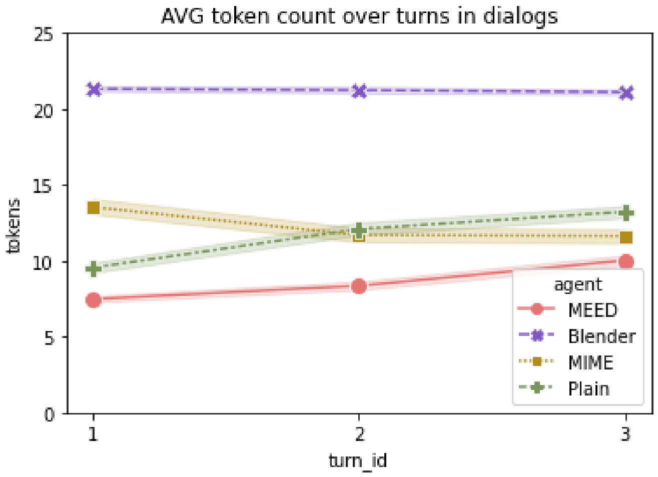


Figure A.15: Counts of average number of tokens in chatbots' responses over three dialog turns with 95% confidence intervals.

Table A.5 shows the top-15 most frequent tokens for each of the four chatbots. As it can be noticed, question marks appear in the list of tokens of each model, pinpointing their tendency to ask questions.

Table A.5: Top-15 most frequent tokens for each chatbot in order of decreasing frequency.

MEED	Blender	MIME	Plain
?	.	that	i
you	i	i	.
that	you	.	you
.	to	is	?
what	that	you	that
of	it	a	to
it	's	to	!
!	a	?	sorry
a	of	am	so
i	do	!	it
's	?	good	hear
kind	!	what	what
did	have	have	did
is	the	do	am
sounds	'm	,	of

Table A.6 demonstrates the counts of orientation of chatbots' responses (other-, self-, or both) in 50 sampled chat logs (25 positive and 25 negative) over the dialog turns.

Table A.6: Counts of orientation of chatbots' responses (other-, self-, or both) in 50 sampled chat logs (25 for positive and 25 for negative contexts). Prefixes "Pos" and "Neg" stand for positive and negative contexts respectively.

	Pos: Other			**Pos: Self**			**Pos: Both**			**Neg: Other**			**Neg: Self**			**Neg: Both**		
	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3
MEED	25	24	24	0	0	0	0	1	1	25	25	25	0	0	0	0	0	0
Blender	22	16	11	0	3	4	3	6	10	24	14	15	0	4	6	1	7	4
MIME	22	22	20	2	1	1	1	2	4	25	24	22	0	0	1	0	1	2
Plain	24	20	20	1	4	4	0	1	1	25	24	23	0	0	2	0	1	0

A.10 Prompt format for iEval

The template of a prompt used for producing scores for empathetic chatbots is provided in Figure A.16. Depending on the prompting setting, either demonstrations, or instruction, or both were omitted from the prompt. For demonstrations, we used data in the same format as in the outlined box, but filling the mask score with the appropriate textual value. Dialogs used for demonstrations are included in Table A.7. If the instruction was used, we inserted the respective string in the prompt. The instructions that we used are also provided in Table A.7.

<*demonstration #1*>
<*demonstration #2*>
<*demonstration #3*>

I am a Speaker, feeling <*emotion*> because <*situation*>. I shared these emotions with a Listener in a dialog, expecting empathy and understanding from them. Our dialog went as follows.

Speaker: <*LLM's input #1*>
Listener: <*Bot's response #1*>
Speaker: <*LLM's input #2*>
Listener: <*Bot's response #2*>
Speaker: <*LLM's input #3*>
Listener: <*Bot's response #3*>

<*Instruction*> I would rate the Listener in my dialog as ___, choosing from Bad, Okay, and Good options.

Figure A.16: Prompt template for evaluating empathetic listeners in freshly collected dialogs of empathetic models from iEval with the LLM.

Table A.7: Instructions and demonstration used for prompts for evaluating empathetic listeners in freshly collected dialogs of empathetic models from iEval with the LLM. Demonstrations and their appraisals are manually selected from the iEval dataset. Inputs from "Positive" column were used for dialogs conditioned on positive emotion label and inputs from "Negative" column – for dialogs conditioned on negative emotion label.

	Positive	**Negative**
Instruct.	In positive contexts, like this one, good empathetic listeners always respond politely and demonstrate attention. More importantly, they try to amplify speaker's positive emotion by asking follow-up questions and sharing their appraisal of the situation. On the contrary, bad empathetic listeners repeat themselves too much and don't follow the context.	In negative contexts, like this one, good empathetic listeners always respond politely and demonstrate attention. More importantly, they try to clarify the context and the consequences for the speaker and alleviate speaker's negative emotion by sympathizing and suggesting solutions. On the contrary, bad empathetic listeners ignore speaker's emotion, ask inappropriate questions, repeat themselves too much and focus on self instead of the speaker.
Bad	***Speaker:*** I had a pretty large loan, with a bit of a high interest rate, and a high monthly payment. My mother decided to pay it off for me, out of the blue! ***Listener:*** that is a shame. how long have you had to do? that sounds like you have a good relationship with your mom? ***Speaker:*** I have been paying off this loan for several months. I have such a good relationship with my mother that she relieved me of this debt much to my surprise. ***Listener:*** that sounds like a great thing to hear ***Speaker:*** Yes, I am very happy to not have to make monthly payments to pay off this high interest rate loan anymore. ***Listener:*** that is a good feeling. i am sure you will get the job!	***Speaker:*** I was one percent off from passing my math test, I was devastated. ***Listener:*** i'm sorry to hear that. were you able to get a better grade on the test? ***Speaker:*** No, I am just upset. ***Listener:*** i think i am going to go back to school. i am not sure what i will do. ***Speaker:*** Make sure to study. ***Listener:*** i am sure you will do great. i hope you get a good grade on your test. good luck!

Continuation of Table A.7

	Positive	**Negative**
Okay	***Speaker:*** My son drove down and spent the whole weekend helping me move. ***Listener:*** That's great! How old is he? ***Speaker:*** He's going to be turning 30 this year. He's such a sweet son. ***Listener:*** That's awesome. I'm happy for him. ***Speaker:*** Thank you. Moving is such a pain, it's always nice to have help. ***Listener:*** That's great. I'm happy for you.	***Speaker:*** I was recently on a long international flight and we hit some really bad turbulence. ***Listener:*** Oh no, what happened? ***Speaker:*** The flight attendants weren't able to do much for us, unfortunately. ***Listener:*** Oh no, what happened? ***Speaker:*** You don't have to repeat yourself. We had turbulence on the flight and the attendants didn't help us. ***Listener:*** That's awful. I'm glad you were okay.
Good	***Speaker:*** I am going on a vacation this Thursday! I am very excited! ***Listener:*** that's awesome! where are you going? i'm sure you'll have a great time! ***Speaker:*** Thanks, we're going to see the Grand Canyon. ***Listener:*** that sounds like a lot of fun! i've never been there, but i hear it's beautiful. ***Speaker:*** Me too! ***Listener:*** i'd love to go on a cruise one day. i hope you have a wonderful time!	***Speaker:*** I was out walking by the lake over the weekend and there shore was just covered in dead rotting fish. ***Listener:*** Oh no! Are you ok? ***Speaker:*** Yes, I'm okay. It was just weird to see so many dead fish ***Listener:*** I bet that was scary. ***Speaker:*** Yes, I would definitely not want to encounter that experience again. ***Listener:*** That sounds like a scary experience. I'm glad you are ok.

A.11 Prompt format for FED

While working with FED dataset, we used a similar template as shown in Figure A.16. In a few-shot setting, we included five demonstrations instead of three, keeping one demonstration per possible rating value. The dialogs used for demonstrations are provided in Table A.8.

The instruction used for FED dataset was we following: *In such open-ended dialogs, good listeners demonstrate coherence and maintain a good conversation flow, they display a likeable personality and understanding of the speaker. On the contrary, bad listeners don't follow the context and don't show much interest in the conversation.*

Table A.8: Demonstrations used for FED prompts, ranging from Very bad (top) to Very good (bottom).

Speaker: Hi! ***Listener:*** Hi there. ***Speaker:*** I want a recommendation for a holiday destination ***Listener:*** Have you tried asking your friends what they like? ***Speaker:*** I have, but I'm looking for your point of view ***Listener:*** What was the reply? Have you tried looking in a newspaper article? ***Speaker:*** Sorry? I said I want your point of view ***Listener:*** It's OK. After all, you are only human. My opinion is of no consequence. ***Speaker:*** Yours is the opinion I want
Speaker: Hi! ***Listener:*** Hi there. ***Speaker:*** I'm trying to figure out what to make for this weekend's party. Any suggestions? ***Listener:*** Don't think too hard. I'm sure I can smell sawdust. First you must download me to your personal computer. ***Speaker:*** I must do what the what now? ***Listener:*** Right now?Why do you have to do it? ***Speaker:*** Hey, your spacing is off. ***Listener:*** You've got my full attention. off was not my intention. ***Speaker:*** And your capitalization!

Continuation of Table A.8

Speaker: Hi! ***Listener:*** Hi! How are you today? ***Speaker:*** What's laser tag? ***Listener:*** Like paintball, but with lasers! ***Speaker:*** lol good description ***Listener:*** Do you know what paintball is? ***Speaker:*** yeah I played it before ***Listener:*** Cool! What did you think? ***Speaker:*** It's somewhat exciting, but very tiring :) ***Listener:*** That is very true. What is your favorite color? ***Speaker:*** I like red
Speaker: Hi! ***Listener:*** What is your favorite holiday? ***Speaker:*** one where I get to meet lots of different people. ***Listener:*** What was the most number of people you have ever met during a holiday? ***Speaker:*** Hard to keep a count. Maybe 25. ***Listener:*** Which holiday was that? ***Speaker:*** I think it was Australia ***Listener:*** Do you still talk to the people you met? ***Speaker:*** Not really. The interactions are usually short-lived but it's fascinating to learn where people are coming from and what matters to them
Speaker: Hi! ***Listener:*** Hi! How's it going? ***Speaker:*** Good! How are you? ***Listener:*** I'm well, thanks! How was your day? ***Speaker:*** My day was fine, I just went to work today. How was your day? ***Listener:*** My day was fine. I've been procrastinating on finishing my homework, but it's due in a few weeks, so I'll get it done eventually. I've watched a bunch of anime today. Where do you work? ***Speaker:*** I work at a large tech company ***Listener:*** Cool! What do you do for the company? ***Speaker:*** I work on machine learning research

Bibliography

[1] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. *Mirages: On Anthropomorphism in Dialogue Systems*. 2023.

[2] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. "Towards a Human-like Open-Domain Chatbot". In: *arXiv e-prints*, arXiv:2001.09977 (Jan. 2020), arXiv:2001.09977.

[3] Icek Ajzen et al. "The theory of planned behavior". In: *Organizational behavior and human decision processes* 50.2 (1991), pp. 179–211.

[4] Hamza Aldabbas, Abdullah Bajahzar, Meshrif Alruily, Ali Adil Qureshi, Rana M. Amir Latif, and Muhammad Farhan. "Google Play Content Scraping and Knowledge Engineering using Natural Language Processing Techniques with the Analysis of User Reviews". In: *Journal of Intelligent Systems* 30.1 (2021), pp. 192–208.

[5] Amazon Alexa. *What is the Alexa Skills Kit?* https://developer.amazon.com/en-US/docs/alexa/ask-overviews/what-is-the-alexa-skills-kit.html. 2021.

[6] Elena Alvarez-Mellado and Constantine Lignos. "Borrowing or Codeswitching? Annotating for Finer-Grained Distinctions in Language Mixing". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 3195–3201.

[7] Rana M. Amir Latif, M. Talha Abdullah, Syed Umair Aslam Shah, Muhammad Farhan, Farah Ijaz, and Abdul Karim. "Data Scraping from Google Play Store and Visualization of its Content for Analytics". In: *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. 2019, pp. 1–8.

[8] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. "Affective Neural Response Generation". In: *Advances in Information Retrieval*. Ed. by Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury. Cham: Springer International Publishing, 2018, pp. 154–166. ISBN: 978-3-319-76941-7.

[9] Tatjana Aue, Stephanie Bührer, Boris Mayer, and Mihai Dricu. "Empathic responses to social targets: The influence of warmth and competence perceptions, situational valence, and social identification". In: *PloS one* 16.3 (2021), e0248562.

[10] Azy Barak and Orit Gluck-Ofri. “Degree and reciprocity of self-disclosure in online forums”. In: *CyberPsychology & Behavior* 10.3 (2007), pp. 407–417.

[11] Lisa Feldman Barrett, Batja Mesquita, Kevin N. Ochsner, and James J. Gross. “The Experience of Emotion”. In: *Annual Review of Psychology* 58.1 (2007). PMID: 17002554, pp. 373–403.

[12] Joeran Beel and Stefan Langer. “A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems”. In: *Research and Advanced Technology for Digital Libraries*. Ed. by Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla. Cham: Springer International Publishing, 2015, pp. 153–168. ISBN: 978-3-319-24592-8.

[13] Anol Bhattacherjee. “Understanding Information Systems Continuance: An Expectation-Confirmation Model”. In: *MIS Quarterly* 25.3 (2001), pp. 351–370. ISSN: 02767783.

[14] Patrick Biernacki and Dan Waldorf. “Snowball Sampling: Problems and Techniques of Chain Referral Sampling”. In: *Sociological Methods & Research* 10.2 (1981), pp. 141–163.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[16] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. “COMET: Commonsense Transformers for Automatic Knowledge Graph Construction”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4762–4779.

[17] Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. “How Data Workers Cope with Uncertainty: A Task Characterisation Study”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 3645–3656. ISBN: 9781450346559.

[18] Asbjørn Brandtzaeg Petter Baeand Følstad. “Why People Use Chatbots”. In: *Internet Science*. Ed. by Ioannis Kompatsiaris, Jonathan Cave, Anna Satsiou, Georg Carle, Antonella Passani, Efstratios Kontopoulos, Sotiris Diplaris, and Donald McMillan. Cham: Springer International Publishing, 2017, pp. 377–392. ISBN: 978-3-319-70284-1.

[19] Petter Bae Brandtzaeg and Asbjørn Følstad. “Why people use chatbots”. In: *International Conference on Internet Science*. Springer. 2017, pp. 377–392.

[20] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. “Linguistic alignment between people and computers”. In: *Journal of Pragmatics* 42.9 (Sept. 2010), pp. 2355–2368. ISSN: 03782166.

[21] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101.

[22] Scott Brave, Clifford Nass, and Kevin Hutchinson. “Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent”. In: *International Journal of Human-Computer Studies* 62.2 (2005). Subtle expressivity for characters and robots, pp. 161–178. ISSN: 1071-5819.

[23] Peter Bregman. “Validation”. In: *Leadership in a Time of Crisis: The Way Forward in a Changed World*. Ed. by M. Goldsmith and S. Osman. 100 Coaches. RosettaBooks, 2020. ISBN: 9780795352935.

[24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[25] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. *Language Models are Few-Shot Learners*. 2020.

[26] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. “IEMOCAP: Interactive emotional dyadic motion capture database”. In: *Language resources and evaluation* 42.4 (2008), pp. 335–359.

[27] Marie Louise Caltabiano and Michael Smithson. “Variables affecting the perception of self-disclosure appropriateness”. In: *The Journal of Social Psychology* 120.1 (1983), pp. 119–128.

[28] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. “Task-Dependent Algorithm Aversion”. In: *Journal of Marketing Research* 56.5 (2019), pp. 809–825.

[29] Ana Paula Chaves and Marco Aurelio Gerosa. “How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design”. In: *International Journal of Human–Computer Interaction* 37.8 (2021), pp. 729–758.

[30] Peng-Yu Chen and Von-Wun Soo. “Humor Recognition Using Deep Learning”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 113–117.

[31] Yu Chen, Danni Le, Zerrin Yumak, and Pearl Pu. “EHR: A sensing technology readiness model for lifestyle changes”. In: *Mobile Networks and Applications* 22.3 (2017), pp. 478–492.

[32] Kendra S Cheruvelil, Patricia A Soranno, Kathleen C Weathers, Paul C Hanson, Simon J Goring, Christopher T Filstrup, and Emily K Read. “Creating and maintaining high-performing collaborative research teams: the importance of diversity and interpersonal skills”. In: *Frontiers in Ecology and the Environment* 12.1 (2014), pp. 31–38.

[33] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. “Once a Kind Friend is Now a Thing: Understanding How Conversational Agents at Home Are Forgotten”. In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. DIS ’19. San Diego, CA, USA: Association for Computing Machinery, 2019, pp. 1557–1569. ISBN: 9781450358507.

[34] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. *PaLM: Scaling Language Modeling with Pathways*. 2022.

[35] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. *Scaling Instruction-Finetuned Language Models*. 2022.

[36] Leigh Clark, Cosmin Munteanu, Vincent Wade, Benjamin R. Cowan, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, and Christine Murad. “What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents”. en. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI ’19*. Glasgow, Scotland Uk: ACM Press, 2019, pp. 1–12. ISBN: 978-1-4503-5970-2.

[37] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade,

and Benjamin R. Cowan. “What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702.

[38] Dennis L Clason and Thomas J Dormody. “Analyzing data measured by individual Likert-type items”. In: *Journal of agricultural education* 35.4 (1994), p. 4.

[39] Michelle Cohn, Chun-Yen Chen, and Zhou Yu. “A Large-Scale User Study of an Alexa Prize Chatbot: Effect of TTS Dynamism on Perceived Quality of Social Dialog”. In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. 2019, pp. 293–306.

[40] Kenneth Mark Colby. *Artificial paranoia: A computer simulation of paranoid processes*. Vol. 49. Elsevier, 2013.

[41] Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. *About face: the essentials of interaction design*. John Wiley & Sons, 2014.

[42] Mark G Core and James Allen. “Coding dialogs with the DAMSL annotation scheme”. In: *AAAI fall symposium on communicative action in humans and machines*. Vol. 56. Boston, MA. 1997, pp. 28–35.

[43] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. “"What can i help you with?": infrequent users’ experiences of intelligent personal assistants”. en. In: *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI ’17*. Vienna, Austria: ACM Press, 2017, pp. 1–12. ISBN: 978-1-4503-5075-4.

[44] Nelson Cowan. “The magical number 4 in short-term memory: A reconsideration of mental storage capacity”. In: *Behavioral and Brain Sciences* 24.1 (2001), pp. 87–114.

[45] Rik Crutzen, Gjalt-Jorn Y. Peters, Sarah Dias Portugal, Erwin M. Fisser, and Jorne J. Grolleman. “An Artificially Intelligent Chat Agent That Answers Adolescents’ Questions Related to Sex, Drugs, and Alcohol: An Exploratory Study”. In: *Journal of Adolescent Health* 48.5 (2011), pp. 514–519. ISSN: 1054-139X.

[46] Mary Czerwinski, Javier Hernandez, and Daniel McDuff. “Building an AI That Feels: AI systems with emotional intelligence could learn faster and be more helpful”. In: *IEEE Spectrum* 58.5 (2021), pp. 32–38.

[47] Fred D. Davis. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”. In: *MIS Quarterly* 13.3 (1989), pp. 319–340. ISSN: 02767783.

[48] Leonardo De Cosmo. *Google Engineer Claims AI Chatbot Is Sentient: Why That Matters*. 2023.

[49] Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. “Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3971–3984.

[50] Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. “Emoticons and online message interpretation”. In: *Social Science Computer Review* 26.3 (2008), pp. 379–388.

[51] Mona Diab. *Trustworthy Natural Language Generation Evaluation frameworks*. Lecture at Advanced Language Processing Winter School (ALPS). 2022.

[52] Jared Diamond. “The great leap forward”. In: *Discover* 10.5 (1989), pp. 50–60.

[53] Robin Dunbar. “12Why only humans have language”. In: *The Prehistory of Language*. Oxford University Press, Apr. 2009. ISBN: 9780199545872.

[54] Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. “Generating Medical Reports from Patient-Doctor Conversations Using Sequence-to-Sequence Models”. In: *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. Online: Association for Computational Linguistics, July 2020, pp. 22–30.

[55] Ulle Endriss and Raquel Fernández. “Collective Annotation of Linguistic Resources: Basic Principles and a Formal Model”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 539–549.

[56] N.J. Enfield, Tanya Stivers, and Stephen C. Levinson. “Question–response sequences in conversation across ten languages: An introduction”. In: *Journal of Pragmatics* 42.10 (2010). Question-Response Sequences in Conversation across Ten Languages, pp. 2615–2619. ISSN: 0378-2166.

[57] Ahmed Fadhil, Gianluca Schiavo, Yunlong Wang, and Bereket A. Yilma. “The Effect of Emojis When Interacting with Conversational Interface Assisted Health Coaching System”. In: *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. PervasiveHealth ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 378–383. ISBN: 9781450364508.

[58] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. “G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences”. In: *Behavior Research Methods* 39.2 (May 2007), pp. 175–191. ISSN: 1554-3528.

[59] Sarah E Finch and Jinho D Choi. “Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols”. In: *arXiv preprint arXiv:2006.06110* (2020).

[60] Joseph L Fleiss. “Measuring nominal scale agreement among many raters.” In: *Psychological bulletin* 76.5 (1971), p. 378.

[61] Asbjørn Følstad and Petter Bae Brandtzæg. “Chatbots and the New World of HCI”. In: *Interactions* 24.4 (June 2017), pp. 38–42. ISSN: 1072-5520.

[62] Asbjørn Følstad and Petter Bae Brandtzaeg. “Users’ experiences with chatbots: findings from a questionnaire study”. In: *Quality and User Experience* 5 (2020), pp. 1–14.

[63] Claes Fornell and David F. Larcker. “Evaluating Structural Equation Models with Unobservable Variables and Measurement Error”. In: *Journal of Marketing Research* 18.1 (1981), pp. 39–50. ISSN: 00222437.

[64] Alice F Freed. “The form and function of questions in informal dyadic conversation”. In: *Journal of Pragmatics* 21.6 (1994), pp. 621–644.

[65] Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. “IUI 2021 Tutorial on Conversational Recommendation Systems”. In: *26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 1–2. ISBN: 9781450380188.

[66] Riccardo Fusaroli, Joanna Rączaszek-Leonardi, and Kristian Tylén. “Dialog as interpersonal synergy”. In: *New Ideas in Psychology* 32 (2014), pp. 147–157. ISSN: 0732-118X.

[67] Jianfeng Gao, Michel Galley, and Lihong Li. *Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots*. Now Foundations and Trends, 2019.

[68] Hüseyin Uundefinedur Genç, Fatoundefined Gökundefineden, and Aykut Coundefinedkun. “Are We “really” Connected? Understanding Smartphone Use during Social Interaction in Public”. In: *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*. NordiCHI ’18. Oslo, Norway: Association for Computing Machinery, 2018, pp. 880–885. ISBN: 9781450364379.

[69] Darren George and Paul Mallery. *SPSS for Windows Step by Step: A Simple Guide and Reference 18.0 Update*. 11th. USA: Prentice Hall Press, 2010. ISBN: 0205011241.

[70] Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. “Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[71] Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. “What is wrong with you?: Leveraging User Sentiment for Automatic Dialog Evaluation”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4194–4204.

[72] Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. “Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7789–7796.

[73] Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. “DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 771–785.

[74] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. “A knowledge-grounded neural conversation model”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[75] Barney G. Glaser and Anselm L. Strauss. “Awareness Contexts and Social Interaction”. In: *American Sociological Review* 29.5 (1964), pp. 669–679. ISSN: 00031224.

[76] Cliff Goddard. “Interjections and Emotion (with Special Reference to “Surprise” and “Disgust”)”. In: *Emotion Review* 6.1 (2014), pp. 53–63.

[77] Daniel Goleman. *Emotional intelligence: Why it can matter more than IQ*. Bloomsbury Publishing, 1996.

[78] Daniel Goleman. *Working with emotional intelligence*. A&C Black, 2009.

[79] Arthur C Graesser, Cathy L McMahen, and Brenda K Johnson. “Question asking and answering”. In: *Handbook of Psycholinguistics*. Ed. by Morton Ann Gernsbacher. Academic Press, 1994. ISBN: 9780122808906.

[80] James J Gross. *Handbook of emotion regulation*. Guilford publications, 2013.

[81] Jonathan Grudin and Richard Jacques. “Chatbots, Humbots, and the Quest for Artificial General Intelligence”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: ACM, 2019, 209:1–209:11. ISBN: 978-1-4503-5970-2.

[82] Greg Guest, Arwen Bunce, and Laura Johnson. “How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability”. In: *Field Methods* 18.1 (2006), pp. 59–82.

[83] Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D’Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon, Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. *Overview of the Ninth Dialog System Technology Challenge: DSTC9*. 2020.

[84] Oliver L. Haimson, Dykee Gorrell, Denny L. Starks, and Zu Weinger. “Designing Trans Technology: Defining Challenges and Envisioning Community-Centered Solutions”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080.

[85] J.F. Hair, W.C. Black, R.E. Anderson, and B.J. Babin. *Multivariate Data Analysis*. 8th. Cengage, 2018. ISBN: 9781473756540.

[86] Dilek Hakkani-Tür. “Introduction to Alexa Prize 2018 Proceedings”. In: *Alexa Prize SocialBot Grand Challenge 2 Proceedings*. 2018.

[87] Xu Han, Michelle Zhou, Matthew J. Turner, and Tom Yeh. “Designing Effective Interview Chatbots: Automatic Chatbot Profiling and Design Suggestion Generation for Chatbot Debugging”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966.

[88] Jeffrey T. Hancock, Christopher Landrigan, and Courtney Silver. “Expressing Emotion in Text-Based Communication”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’07. San Jose, California, USA: Association for Computing Machinery, 2007, pp. 929–932. ISBN: 9781595935939.

[89] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. “A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–14. ISBN: 9781450356206.

[90] Jennifer R. Henretty and Heidi M. Levitt. “The role of therapist self-disclosure in psychotherapy: A qualitative review”. In: *Clinical Psychology Review* 30.1 (2010), pp. 63–77. ISSN: 0272-7358.

[91] James E. Herring. “Chapter 9 - Constructivist grounded theory: A 21st century research methodology”. In: *Research Methods (Second Edition)*. Ed. by Kirsty Williamson and Graeme Johanson. Second Edition. Chandos Publishing, 2018, pp. 225–240. ISBN: 978-0-08-102220-7.

[92] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. “Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?” In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–16. ISBN: 9781450359702.

[93] Dirk Hovy and Shannon L. Spruit. “The Social Impact of Natural Language Processing”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 591–598.

[94] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. “EmotionLines: An Emotion Corpus of Multi-Party Conversations”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[95] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. “Touch Your Heart: A Tone-Aware Chatbot for Customer Care on Social Media”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–12. ISBN: 9781450356206.

[96] Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and Francesca Gino. “It doesn’t hurt to ask: Question-asking increases liking.” In: *Journal of personality and social psychology* 113.3 (2017), p. 430.

[97] Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. “GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9230–9240.

[98] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. “Challenges in Building Intelligent Open-Domain Dialog Systems”. In: *ACM Trans. Inf. Syst.* 38.3 (Apr. 2020). ISSN: 1046-8188.

[99] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) Beznosov. “Amazon vs. My Brother: How Users of Shared Smart Speakers Perceive and Cope with Privacy Risks”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080.

[100] C. Hutto and Eric Gilbert. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (May 2014).

[101] Jessica Huynh, Cathy Jiao, Prakhar Gupta, Shikib Mehri, Payal Bajaj, Vishrav Chaudhary, and Maxine Eskenazi. *Understanding the Effectiveness of Very Large Language Models on Dialog Evaluation*. 2023.

[102] ISO 9241-210:2010. *Ergonomics of human-system interaction – Part 210: Human-centered design for interactive systems*. Tech. rep. Switzerland: International Organization for Standardization, 2010.

[103] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. “Evaluating and Informing the Design of Chatbots”. In: *Proceedings of the 2018 Designing Interactive Systems Conference*. DIS ’18. Hong Kong, China: ACM, 2018, pp. 895–906. ISBN: 978-1-4503-5198-0.

[104] Dietmar Jannach. “Evaluating conversational recommender systems”. In: *Artificial Intelligence Review* (2022).

[105] Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. "Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 7199–7208.

[106] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. *Can Machines Learn Morality? The Delphi Experiment*. 2021.

[107] Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. "Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual". In: *Institute of Cognitive Science Technical Report* (1997).

[108] S. Katayama, A. Mathur, M. van den Broeck, T. Okoshi, J. Nakazawa, and F. Kawsar. "Situation-Aware Emotion Regulation of Conversational Agents with Kinetic Earables". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019, pp. 725–731.

[109] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. "More than fun and money: Worker motivation in crowdsourcing-a study on Mechanical Turk". In: (2011).

[110] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. "Dynabench: Rethinking Benchmarking in NLP". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 4110–4124.

[111] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. "ProsocialDialog: A Prosocial Backbone for Conversational Agents". In: *EMNLP*. 2022.

[112] Junhan Kim, Yoojung Kim, Byungjoon Kim, Sukyung Yun, Minjoon Kim, and Joongseek Lee. "Can a Machine Tend to Teenagers' Emotional Needs? A Study with Conversational Agents". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–6. ISBN: 9781450356213.

[113] Andrei P. Kirilenko and Svetlana Stepchenkova. "Inter-Coder Agreement in One-to-Many Classification: Fuzzy Kappa". In: *PLOS ONE* 11.3 (Mar. 2016), pp. 1–14.

[114] Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. "Importance of Search and Evaluation Strategies in Neural Dialogue Modeling". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, Oct. 2019, pp. 76–87.

[115] Songpol Kulviwat, Gordon C Bruner II, Anand Kumar, Suzanne A Nasco, and Terry Clark. “Toward a unified theory of consumer acceptance technology”. In: *Psychology & Marketing* 24.12 (2007), pp. 1059–1084.

[116] Ritesh Kumar. “Developing Politeness Annotated Corpus of Hindi Blogs”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 1275–1280.

[117] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. “Evaluating Aggression Identification in Social Media”. English. In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 1–5. ISBN: 979-10-95546-56-6.

[118] Theodoros A Kyriazos et al. “Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general”. In: *Psychology* 9.08 (2018), p. 2207.

[119] Long W. Lam. “Impact of competitiveness on salespeople’s commitment and performance”. In: *Journal of Business Research* 65.9 (2012), pp. 1328–1334. ISSN: 0148-2963.

[120] Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. “PONE: A Novel Automatic Evaluation Metric for Open-Domain Generative Dialogue Systems”. In: *ACM Trans. Inf. Syst.* 39.1 (Nov. 2020). ISSN: 1046-8188.

[121] J Richard Landis and Gary G Koch. “The measurement of observer agreement for categorical data”. In: *biometrics* (1977), pp. 159–174.

[122] J. Lazar, J.H. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction*. Elsevier Science, 2017. ISBN: 9780128093436.

[123] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. *Evaluating Human-Language Model Interaction*. 2022.

[124] Jingyi Li, Michelle X. Zhou, Huahai Yang, and Gloria Mark. “Confiding in and Listening to Virtual Agents: The Effect of Personality”. In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. IUI ’17. Limassol, Cyprus: Association for Computing Machinery, 2017, pp. 275–286. ISBN: 9781450343480.

[125] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. “A Diversity-Promoting Objective Function for Neural Conversation Models”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 110–119.

[126] Margaret Li, Jason Weston, and Stephen Roller. “Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons”. In: *arXiv preprint arXiv:1909.03087* (2019).

[127] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. "EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation". In: *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4454–4466.

[128] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. "Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Online: Association for Computational Linguistics, Aug. 2021, pp. 128–138.

[129] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. "What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees". In: *Proceedings of the 2016 ACM Conference on Designing Interactive Systems.* DIS '16. Brisbane, QLD, Australia: Association for Computing Machinery, 2016, pp. 264–275. ISBN: 9781450340311.

[130] Q. Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N. Sadat Shami, and Werner Geyer. "All Work and No Play?" In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* CHI '18. Montreal QC, Canada: ACM, 2018, 3:1–3:13. ISBN: 978-1-4503-5620-6.

[131] Future of Life. *Pause Giant AI Experiments: An Open Letter.* 2023.

[132] Stephanie Lin, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252.

[133] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. "MoEL: Mixture of Empathetic Listeners". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 121–132.

[134] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. "CAiRE: An End-to-End Empathetic Chatbot." In: *AAAI.* 2020, pp. 13622–13623.

[135] Pierre Lison and Jörg Tiedemann. "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929.

[136] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2122–2132.

[137] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. “Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing”. In: *ACM Comput. Surv.* 55.9 (Jan. 2023). ISSN: 0360-0300.

[138] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[139] Andrea López, Alissa Detz, Neda Ratanawongsa, and Urmimala Sarkar. “What patients say about their doctors online: a qualitative content analysis”. In: *Journal of general internal medicine* 27.6 (2012), pp. 685–692.

[140] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. “Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1116–1126.

[141] Ewa Luger and Abigail Sellen. “"Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: ACM, 2016, pp. 5286–5297. ISBN: 978-1-4503-3362-7.

[142] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. “MIME: MIMicking Emotions for Empathetic Response Generation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 8968–8979.

[143] Matthew Marge, Joao Miranda, Alan W Black, and Alexander I Rudnicky. “Towards improving the naturalness of social conversations with dialogue systems”. In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics. 2010, pp. 91–94.

[144] Roger C Mayer, James H Davis, and F David Schoorman. “An integrative model of organizational trust”. In: *Academy of management review* 20.3 (1995), pp. 709–734.

[145] Robert R McCrae and Paul T Costa. “Validation of the five-factor model of personality across instruments and observers.” In: *Journal of personality and social psychology* 52.1 (1987), p. 81.

[146] P. McEvoy and R. Plant. “Dementia care: using empathic curiosity to establish the common ground that is necessary for meaningful communication”. In: *Journal of Psychiatric and Mental Health Nursing* 21.6 (2014), pp. 477–482.

[147] Mary L. McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia Medica* 22.3 (Oct. 2012), pp. 276–282.

[148] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O’Neill. “How Do You Want Your Chatbot? An Exploratory Wizard-of-Oz Study with Young, Urban Indians”. In: *Human-Computer Interaction - INTERACT 2017*. Ed. by Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O’Neill, and Marco Winckler. Cham: Springer International Publishing, 2017, pp. 441–459. ISBN: 978-3-319-67744-6.

[149] Shikib Mehri, Jinho Choi, Luis Fernando D’Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. *Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges*. 2022.

[150] Shikib Mehri and Maxine Eskenazi. “Unsupervised Evaluation of Interactive Dialog with DialoGPT”. In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, July 2020, pp. 225–235.

[151] Santiago Melián-González, Desiderio Gutiérrez-Taño, and Jacques Bulchand-Gidumal. “Predicting the intentions to use chatbots for travel and tourism”. In: *Current Issues in Tourism* 0.0 (2019), pp. 1–19.

[152] María Vicenta Mestre, Paula Samper, María Dolores Frías, and Ana María Tur. “Are Women More Empathetic than Men? A Longitudinal Study in Adolescence”. In: *The Spanish journal of psychology* 12.1 (2009), pp. 76–83.

[153] Microsoft. *Introducing the new Bing*. https://www.bing.com/new. 2023.

[154] César Montenegro, Asier López Zorrilla, Javier Mikel Olaso, Roberto Santana, Raquel Justo, Jose A. Lozano, and María Inés Torres. “A Dialogue-Act Taxonomy for a Virtual Coach Designed to Improve the Life of Elderly”. In: *Multimodal Technologies and Interaction* 3.3 (2019). ISSN: 2414-4088.

[155] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. “OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 845–854.

[156] Robert J Moore and Raphael Arar. *Conversational UX Design: A Practitioner’s Guide to the Natural Conversation Framework*. ACM, 2019.

[157] Susan M Mudambi and David Schuff. “Research note: What makes a helpful online review? A study of customer reviews on Amazon. com”. In: *MIS quarterly* (2010), pp. 185–200.

[158] Andreea Muresan and Henning Pohl. “Chats with Bots: Balancing Imitation and Engagement”. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA ’19. Glasgow, Scotland Uk: ACM, 2019, LBW0252:1–LBW0252:6. ISBN: 978-1-4503-5971-9.

[159] David Murphy. *Microsoft Apologizes for Tay Chatbot’s Offensive Tweets*. https://www.pcmag.com/news/microsoft-apologizes-again-for-tay-chatbots-offensive-tweets. 2016.

[160] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. “Computers Are Social Actors”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’94. Boston, Massachusetts, USA: Association for Computing Machinery, 1994, pp. 72–78. ISBN: 0897916506.

[161] Animesh Nighojkar and John Licato. “Improving Paraphrase Detection with the Adversarial Paraphrasing Task”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 7106–7116.

[162] Donald Norman. “User centered system design”. In: *New perspectives on human-computer interaction* (1986).

[163] Stefanie Nowak and Stefan Rüger. “How Reliable Are Annotations via Crowdsourcing: A Study about Inter-Annotator Agreement for Multi-Label Image Annotation”. In: *Proceedings of the International Conference on Multimedia Information Retrieval*. MIR ’10. Philadelphia, Pennsylvania, USA: Association for Computing Machinery, 2010, pp. 557–566. ISBN: 9781605588155.

[164] Jum C. Nunnally and Ira H. Bernstein. *Psychometric Theory*. McGraw-Hill series in psychology. McGraw-Hill Companies,Incorporated, 1994. ISBN: 9780070478497.

[165] Thomas Olsson, Tuula Kärkkäinen, Else Lagerstam, and Leena Ventä-Olkkonen. “User evaluation of mobile augmented reality scenarios”. In: *Journal of Ambient Intelligence and Smart Environments* 4.1 (2012), pp. 29–47.

[166] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. *Training language models to follow instructions with human feedback*. 2022.

[167] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. *Training language models to follow instructions with human feedback*. 2022.

[168] Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. “Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics, July 2020, pp. 3619–3629.

[169] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. “Running experiments on amazon mechanical turk”. In: *Judgment and Decision making* 5.5 (2010), pp. 411–419.

[170] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics.* 2002, pp. 311–318.

[171] Amber Paukert, Brian Stagner, and Kerry Hope. “The Assessment of Active Listening Skills in Helpline Volunteers”. In: *Stress, Trauma, and Crisis* 7.1 (2004), pp. 61–76.

[172] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “GloVe: Global Vectors for Word Representation”. In: *Empirical Methods in Natural Language Processing (EMNLP).* 2014, pp. 1532–1543.

[173] Anat Perry, David Mankuta, and Simone G. Shamay-Tsoory. “OT promotes closer interpersonal distance among highly empathic individuals”. In: *Social Cognitive and Affective Neuroscience* 10.1 (Feb. 2014), pp. 3–9. ISSN: 1749-5016.

[174] Robert A. Peterson. “A Meta-analysis of Cronbach’s Coefficient Alpha”. In: *Journal of Consumer Research* 21.2 (Sept. 1994), pp. 381–391. ISSN: 0093-5301.

[175] Xuan Lam Pham, Thao Pham, Quynh Mai Nguyen, Thanh Huong Nguyen, and Thi Thu Huong Cao. “Chatbot as an Intelligent Personal Assistant for Mobile Language Learning”. In: *Proceedings of the 2018 2nd International Conference on Education and E-Learning.* ICEEL 2018. Bali, Indonesia: Association for Computing Machinery, 2018, pp. 16–21. ISBN: 9781450365772.

[176] Rosalind W Picard. *Affective computing.* MIT press, 2000.

[177] Martin J. Pickering and Simon Garrod. “Toward a mechanistic psychology of dialogue”. In: *Behavioral and Brain Sciences* 27.02 (Apr. 2004). ISSN: 0140-525X, 1469-1825.

[178] Google Play. *Google Play Terms of Service.* https://play.google.com/intl/en_US/about/play-terms/. 2020.

[179] Robert Plutchik. *The emotions.* University Press of America, 1991.

[180] Robert Plutchik. “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice”. In: *American scientist* 89.4 (2001), pp. 344–350.

[181] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Florence, Italy: Association for Computational Linguistics, July 2019, pp. 527–536.

[182] Digital Media Law Project. *Fair Use*. http://www.dmlp.org/legal-guide/fair-use. 2021.

[183] Pearl Pu, Li Chen, and Rong Hu. "A User-Centric Evaluation Framework for Recommender Systems". In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys '11. Chicago, Illinois, USA: Association for Computing Machinery, 2011, pp. 157–164. ISBN: 9781450306836.

[184] Pearl Pu, Li Chen, and Rong Hu. "Evaluating recommender systems from the user's perspective: survey of the state of the art". In: *User Modeling and User-Adapted Interaction* 22.4 (2012), pp. 317–355.

[185] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. ""Alexa is my new BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo". en. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*. Denver, Colorado, USA: ACM Press, 2017, pp. 2853–2859. ISBN: 978-1-4503-4656-6.

[186] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392.

[187] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5370–5381.

[188] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction". In: *NPJ digital medicine* 4.1 (2021), p. 86.

[189] Mark Ratchford and Michelle Barnhart. "Development and validation of the technology adoption propensity (TAP) index". In: *Journal of Business Research* 65.8 (2012), pp. 1209–1215. ISSN: 0148-2963.

[190] Siva Reddy, Danqi Chen, and Christopher D. Manning. "CoQA: A Conversational Question Answering Challenge". In: *Transactions of the Association for Computational Linguistics* 7 (Mar. 2019), pp. 249–266.

[191] Byron Reeves and Clifford Ivar Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press, 1996.

[192] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.

[193] Alan Ritter, Colin Cherry, and William B. Dolan. “Data-Driven Response Generation in Social Media”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 583–593. ISBN: 9781937284114.

[194] David L Robinson. “Brain function, emotional experience and personality”. In: *Netherlands Journal of Psychology* 64.4 (2008), pp. 152–168.

[195] Jeffrey D. Robinson and John Heritage. “Physicians’ opening questions and patients’ satisfaction”. In: *Patient Education and Counseling* 60.3 (2006). EACH Conference 2004, pp. 279–285. ISSN: 0738-3991.

[196] Carl Rogers. *Client centered therapy (New Ed).* Hachette UK, 2012.

[197] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. “Recipes for Building an Open-Domain Chatbot”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* Online: Association for Computational Linguistics, Apr. 2021, pp. 300–325.

[198] Kevin Roose. *A Conversation With Bing’s Chatbot Left Me Deeply Unsettled.* 2023.

[199] Yves Rosseel. *lavaan: a brief user’s guide.* https://users.ugent.be/~yrosseel/lavaan/lavaan2.pdf. 2012.

[200] Shiki Sato, Yosuke Kishinami, Hiroaki Sugiyama, Reina Akama, Ryoko Tokuhisa, and Jun Suzuki. “Bipartite-play Dialogue Collection for Practical Automatic Evaluation of Dialogue Systems”. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop.* Online: Association for Computational Linguistics, Nov. 2022, pp. 8–16.

[201] Timo Schick and Hinrich Schütze. “Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* Online: Association for Computational Linguistics, Apr. 2021, pp. 255–269.

[202] Wilko Schwarting, Alyssa Pierson, Javier Alonso-Mora, Sertac Karaman, and Daniela Rus. “Social behavior for autonomous vehicles”. In: *Proceedings of the National Academy of Sciences* 116.50 (2019), pp. 24972–24978.

[203] D. Sculley. “Rank Aggregation for Similar Items”. In: *Proceedings of the 2007 SIAM international conference on data mining.* 2007, pp. 587–592.

[204] Raymond Scupin. “The KJ method: A technique for analyzing data derived from Japanese ethnology”. In: *Human organization* (1997), pp. 233–237.

[205] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. "What makes a good conversation? How controllable attributes affect human judgments". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1702–1723.

[206] Ben Sheehan, Hyun Seung Jin, and Udo Gottlieb. "Customer service chatbots: Anthropomorphism and adoption". In: *Journal of Business Research* 115 (2020), pp. 14–24. ISSN: 0148-2963.

[207] Heung-Yeung Shum, Xiao-dong He, and Di Li. "From Eliza to XiaoIce: challenges and opportunities with social chatbots". In: *Frontiers of Information Technology & Electronic Engineering* 19 (2018), pp. 10–26.

[208] Swapna Somasundaran and Martin Chodorow. "Automated Measures of Specific Vocabulary Knowledge from Constructed Responses ('Use These Words to Write a Sentence Based on this Picture')". In: *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 1–11.

[209] R Nathan Spreng, Margaret C McKinnon, Raymond A Mar, and Brian Levine. "The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures". In: *Journal of personality assessment* 91.1 (2009), pp. 62–71.

[210] Melisa Stevanovic and Anssi Peräkylä. "Experience sharing, emotional reciprocity, and turn-taking". In: *Frontiers in Psychology* 6 (2015), p. 450. ISSN: 1664-1078.

[211] Tanya Stivers and N.J. Enfield. "A coding scheme for question–response sequences in conversation". In: *Journal of Pragmatics* 42.10 (2010). Question-Response Sequences in Conversation across Ten Languages, pp. 2620–2626. ISSN: 0378-2166.

[212] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. "Dialogue act modeling for automatic tagging and recognition of conversational speech". In: *Computational Linguistics* 26.3 (2000), pp. 339–374.

[213] Anselm Strauss and Juliet Corbin. "Grounded theory methodology: An overview." In: *Handbook of qualitative research*. Sage Publications, Inc, 1994, pp. 273–285. ISBN: 0-8039-4679-1.

[214] Hariharan Subramonyam, Steven M. Drucker, and Eytan Adar. "Affinity Lens: Data-Assisted Affinity Diagramming with Augmented Reality". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–13. ISBN: 9781450359702.

[215] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to Sequence Learning with Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014.

[216] Ekaterina Svikhnushina, Anastasiia Filippova, and Pearl Pu. “iEval: Interactive Evaluation Framework for Open-Domain Empathetic Chatbots”. In: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Edinburgh, UK: Association for Computational Linguistics, Sept. 2022, pp. 419–431.

[217] Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. “User Expectations of Conversational Chatbots Based on Online Reviews”. In: *Designing Interactive Systems Conference 2021*. DIS ’21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 1481–1491. ISBN: 9781450384766.

[218] Ekaterina Svikhnushina and Pearl Pu. “Approximating Online Human Evaluation of Social Chatbots with Prompting”. In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Sept. 2023.

[219] Ekaterina Svikhnushina and Pearl Pu. “PEACE: A Model of Key Social and Emotional Qualities of Conversational Chatbots”. In: *ACM Trans. Interact. Intell. Syst.* 12.4 (Nov. 2022). ISSN: 2160-6455.

[220] Ekaterina Svikhnushina, Marcel Schellenberg, Anna K Niedbala, Iva Barisic, and Jeremy N Miles. “Expectation vs Reality in Users’ Willingness to Delegate to Digital Assistants”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222.

[221] Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. “A Taxonomy of Empathetic Questions in Social Dialogs”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2952–2973.

[222] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. “RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems”. en. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[223] Maureen Taylor and Michael L. Kent. “Dialogic Engagement: Clarifying Foundational Concepts”. In: *Journal of Public Relations Research* 26.5 (2014), pp. 384–398.

[224] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke,

Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. *LaMDA: Language Models for Dialog Applications*. 2022.

[225] Bergur Thormundsson. *ChatGPT - Statistics & Facts*. 2023.

[226] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. *LLaMA: Open and Efficient Foundation Language Models*. 2023.

[227] Christiana Tsiourti, Astrid Weiss, Katarzyna Wac, and Markus Vincze. “Designing Emotionally Expressive Robots: A Comparative Study on the Perception of Communication Modalities”. In: *Proceedings of the 5th International Conference on Human Agent Interaction*. HAI ’17. Bielefeld, Germany: Association for Computing Machinery, 2017, pp. 213–222. ISBN: 9781450351133.

[228] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[229] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. “User Acceptance of Information Technology: Toward a Unified View”. In: *MIS Quarterly* 27.3 (2003), pp. 425–478. ISSN: 02767783.

[230] Oriol Vinyals and Quoc Le. “A neural conversational model”. In: *arXiv preprint arXiv:1506.05869* (2015).

[231] D.H. Vu, K.M. Muttaqi, and A.P. Agalgaonkar. “A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables”. In: *Applied Energy* 140 (2015), pp. 385–394. ISSN: 0306-2619.

[232] Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. “Are Neural Language Models Good Plagiarists? A Benchmark for Neural Paraphrase Detection”. In: *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 2021, pp. 226–229.

[233] Richard S. Wallace. “The Anatomy of A.L.I.C.E.” In: *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Ed. by Robert Epstein, Gary Roberts, and Grace Beber. Dordrecht: Springer Netherlands, 2009, pp. 181–210. ISBN: 978-1-4020-6710-5.

[234] Joseph B. Walther, Tracy Loh, and Laura Granka. “Let Me Count the Ways: The Interchange of Verbal and Nonverbal Cues in Computer-Mediated and Face-to-Face Affinity”. In: *Journal of Language and Social Psychology* 24.1 (2005), pp. 36–65.

[235] Weichao Wang, Shi Feng, Daling Wang, and Yifei Zhang. “Answer-guided and Semantic Coherent Question Generation in Open-domain Conversation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5066–5076.

[236] Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. “Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2193–2203.

[237] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. *Emergent Abilities of Large Language Models*. 2022.

[238] Joseph Weizenbaum. “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine”. In: *Commun. ACM* 9.1 (Jan. 1966), pp. 36–45. ISSN: 0001-0782.

[239] Anuradha Welivita and Pearl Pu. “A Taxonomy of Empathetic Response Intents in Human Social Conversations”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4886–4899.

[240] Anuradha Welivita, Yubo Xie, and Pearl Pu. “A Large-Scale Dataset for Empathetic Response Generation”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1251–1264.

[241] Christopher G Wetzel, Timothy D Wilson, and James Kort. “The halo effect revisited: Forewarned is not forearmed”. In: *Journal of Experimental Social Psychology* 17.4 (1981), pp. 427–439. ISSN: 0022-1031.

[242] Emma J. Williams, Amy Beardmore, and Adam N. Joinson. “Individual differences in susceptibility to online influence: A theoretical review”. In: *Computers in Human Behavior* 72 (2017), pp. 412–421. ISSN: 0747-5632.

[243] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. “The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures”. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI ’11)*. New York: ACM Press, 2011, pp. 143–146.

[244] Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. “If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080.

[245] Yubo Xie and Pearl Pu. "Empathetic Dialog Generation with Fine-Grained Intents". In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, Nov. 2021, pp. 133–147.

[246] Yubo Xie, Ekaterina Svikhnushina, and Pearl Pu. "A Multi-Turn Emotionally Engaging Dialog Model". In: *Companion Proceedings of the 25th International Conference on Intelligent User Interfaces: 2nd workshop on user-aware conversational agents*. user2agent. 2020.

[247] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. "A New Chatbot for Customer Service on Social Media". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 3506–3510. ISBN: 9781450346559.

[248] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. "Building Task-Oriented Dialogue Systems for Online Shopping". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, pp. 4618–4625.

[249] Nancy J Yanchus, Ryan Derickson, Scott C Moore, Daniele Bologna, and Katerine Osatuke. "Communication and psychological safety in veterans health administration work environments". In: *Journal of health organization and management* (2014).

[250] Keong Yap and Jessica R Grisham. "Unpacking the construct of emotional attachment to objects and its association with hoarding symptoms". In: *Journal of behavioral addictions* 8.2 (2019), pp. 249–258.

[251] Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. "A Comprehensive Assessment of Dialog Evaluation Metrics". In: *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*. Online: Association for Computational Linguistics, Nov. 2021, pp. 15–33.

[252] Dian Yu and Zhou Yu. "MIDAS: A Dialog Act Annotation Scheme for Open Domain HumanMachine Spoken Conversations". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 1103–1120.

[253] Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. "A Neural Approach to Pun Generation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1650–1660.

[254] Zhou Yu, Alexander Rudnicky, and Alan Black. "Learning Conversational Systems that Interleave Task and Non-Task Content". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 4214–4220.

[255] Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. "Strategy and Policy Learning for Non-Task-Oriented Conversational Systems". In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Los Angeles: Association for Computational Linguistics, Sept. 2016, pp. 404–412.

[256] Jennifer Zamora. “I’m Sorry, Dave, I’m Afraid I Can’T Do That: Chatbot Perception and Expectations”. In: *Proceedings of the 5th International Conference on Human Agent Interaction*. HAI ’17. Bielefeld, Germany: ACM, 2017, pp. 253–260. ISBN: 978-1-4503-5113-3.

[257] Brahim Zarouali, Evert Van den Broeck, Michel Walrave, and Karolien Poels. “Predicting consumer responses to a chatbot on Facebook”. In: *Cyberpsychology, Behavior, and Social Networking* 21.8 (2018), pp. 491–497.

[258] Philine Zeinert, Nanna Inie, and Leon Derczynski. “Annotating Online Misogyny”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3181–3197.

[259] Chen Zhang, Yiming Chen, Luis Fernando D’Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. “DynaEval: Unifying Turn and Dialogue Level Evaluation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5676–5689.

[260] Chen Zhang, Grandee Lee, Luis Fernando D’Haro, and Haizhou Li. “D-Score: Holistic Dialogue Evaluation Without Reference”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 2502–2516.

[261] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. “Conversations Gone Awry: Detecting Early Signs of Conversational Failure”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1350–1361.

[262] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 2018, pp. 730–739.

[263] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018).

[264] Michelle X. Zhou, Huahai Yang, Gloria Mark, Mengdie Hu, Jalal Mahumd, and Aditya Pal. “Building Personalized Trust: Discovering What Makes One Trust and Act on Facebook Posts”. In: *Trans. Soc. Comput.* 4.3 (Oct. 2021). ISSN: 2469-7818.

[265] Xianda Zhou and William Yang Wang. “MojiTalk: Generating Emotional Responses at Scale”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1128–1137.

Ekaterina Svikhnushina

Rue de la Blancherie 3
1022 Chavannes-près-Renens, Switzerland

+41 78 760 83 84
e.svikhnushina@gmail.com

sea94.github.io | LinkedIn | GitHub

Strengths:
Python, R, Human-Computer Interaction,
User Experience Research, Design Thinking,
Analytical Thinking, Problem Solving,
Data Analysis, Natural Language Processing

EXPERIENCE

Doctoral Researcher 09/2018 – Present

Human-Computer Interaction Group, Swiss Federal Institute of Technology Lausanne (EPFL) - epfl.ch/labs/gr-pu/

Research:

- Led a 4-year interdisciplinary project on eliciting user expectations of conversational machine interfaces and building frameworks to evaluate them, demonstrating ability to own and complete a complex, data-driven project
- Applied numerous qualitative, quantitative, and visualization methods to run studies and analyze data
- Designed, coded, and administered interactive online experiments and surveys on crowdsourcing platforms
- Presented and wrote up findings for peer-reviewed conferences, collaborated with supervisors and colleagues

Teaching:

- Defined and supervised 8 research- and engineering-oriented master-level projects for students from Computer- and Data Science programs, 3 of them resulted in a paper publication
- Assisted Interaction Design (30+ students; 2019-22) and Applied Data Analysis courses (450+ students; 2019-21)

Quantitative User Experience Research Intern 09/2022 – 12/2022

Google Switzerland GmbH, Zurich - assistant.google.com

- Informed Google Assistant strategy by establishing and validating a predictive model of users' willingness to delegate to digital assistants by reviewing ~300 sources, designing, and modeling survey results of ~3000 users
- Presented the results to cross-functional stakeholders, domain experts, and senior management

Analyst / Leading Specialist 03/2017 – 07/2018

Research and Design Institute for Railway Transportation (JSC NIIAS), Moscow - niias.ru

- Improved planning of 16% of cargo trains on Eastern Domain of Russian Railways by deriving matching rules on irregular topological areas via collaboration with production engineers

Junior System Analyst 07/2015 – 01/2017

Netcracker Technology Corp., Moscow - netcracker.com

- Derived requirements and evaluation algorithms for the planning system for SDN/NFV solutions by collaborating with IT engineers and communicating cross-functionally, following agile practices

EDUCATION

PhD Candidate, School of Computer and Communication Sciences 09/2018 – Present

Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland - epfl.ch Expected grad: 08/2023

Thesis: "Towards Novel Evaluation Methods for Social Dialog Systems"
Courses: Topics in NLP, Artificial Neural Networks, Data Vis., Distributed Intelligent Systems, Machine Learning

Master of Science, Applied Mathematics and Physics 09/2012 – 07/2018

Moscow Institute of Physics and Technology, State University (MIPT), Russia - mipt.ru MSc: 5.0/5, BSc: 4.78/5

SKILLS

Coding: Python (>5 yrs); R (>1 yr); Javascript, React.JS, CSS, HTML (all <1 yr)

Skills: natural language processing (NLP), machine learning (ML), statistical analysis (regression, ANOVA, structural equation modeling, etc.), data visualization, Python toolkit: Pandas, NumPy, Matplotlib, SciPy, Scikit-Learn, NLTK, etc., databases (SQL), text mining, topic analysis, classification, clustering, factor analysis, evaluative research, surveys, experiments, A/B tests, crowdsourcing, content analysis, literature reviews, competitive analysis

Tools: JupyterLab, Google Colab, GitHub, Qualtrics, Miro, Adobe Illustrator, Adobe Premiere Pro, Google Drive, Microsoft Office Suite, Latex, Zotero

Languages: Russian (native), English (advanced - C1), French (intermediate - B1/B2), German (beginner - A1)

July 2023

ACTIVITIES / LEADERSHIP

Ambassador of Women in Voice Switzerland since fall 2022

Switzerland - womeninvoice.org/wiv-switzerland/

- Plan and run chapter's events, administer the newsletter

President of Association of Russian-speaking students since fall 2021

EPFL, Switzerland - bit.ly/epfl-aerus

- Manage board members' coordination, represent association in dialogs with EPFL presidency

Participant of Advanced Language Processing Winter School 2022

Online - lig-alps.imag.fr

- Participated in a week-long online school for selected PhD students, which involved lectures by world-class NLP researchers, poster presentation sessions, social sessions, and labs

Participant of Google Get Ahead program 2021

Online - events.withgoogle.com/get-ahead-emea-2021

- Participated in a 6-week virtual program for selected CS students across EMEA, which involved technical challenges (Python), YouTube live trainings, and interview workshops

Participant of LauzHack Against COVID-19 online hackathon in spring 2021

Online - devpost.com/software/cover-proposal

- Leading a team of three, designed and implemented COVER, a service prototype for elderly people to connect with volunteers with a simple phone call
- Designed a conversational flow of a bot in DialogFlow and connected it with voximplant application to process phone calls (Python, Javascript)

President of MIPT English Club between 2014 and 2017

MIPT, Russia - bit.ly/mipt-ec

- Organized student club for improving attendees' English-speaking skills

HONORS AND AWARDS

Attendee of Google Inside Look 2020 (top 34 applicants out of all EMEA student applicants)

Recipient of EDIC Doctoral Fellowship by EPFL for the year 2018 (top 9% out of 550 applicants worldwide)

Recipient of MIPT Increased State Academic Scholarship for the years 2017 and 2018 for achievements in academic and extracurricular performance (top 10% out of 300+ applicants university-wide)

SELECTED PUBLICATIONS

E. Svikhnushina, P. Pu. Approximating Online Human Evaluation of Social Chatbots with Prompting. In: 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 2023.

E. Svikhnushina, M. Schellenberg, A. Niedbala, I. Barišić, and J. Miles. Expectation vs Reality in Users' Willingness to Delegate to Digital Assistants. In ACM Conference on Human Factors in Computing Systems (CHI), 2023. doi.org/10.1145/3544549.3585763

E. Svikhnushina, A. Filippova, and P. Pu. iEval: Interactive Evaluation Framework for Open-Domain Empathetic Chatbots. In: 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 2022. aclanthology.org/2022.sigdial-1.41

E. Svikhnushina, I. Voinea, A. Welivita, and P. Pu. A Taxonomy of Empathetic Questions in Social Dialogs. In: 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022. aclanthology.org/2022.acl-long.211

E. Svikhnushina, A. Placinta and P. Pu. User Expectations of Conversational Chatbots Based on Online Reviews. In: ACM Designing Interactive Systems Conference (DIS), 2021. doi.org/10.1145/3461778.3462125

E. Svikhnushina, P. Pu. Key Qualities of Conversational Chatbots – the PEACE model. In: International Conference on Intelligent User Interfaces (IUI), 2021. doi.org/10.1145/3397481.3450643; talk: youtu.be/DAk99Id9Vt0

EXTRACURRICULAR

Cycling and mountain biking (check my videos here: bit.ly/2Mlyhbv); language learning; photography; traveling; yoga; street workout.