Democratizing Chatbot Debugging: A Computational Framework for Evaluating and Explaining Inappropriate Chatbot Responses
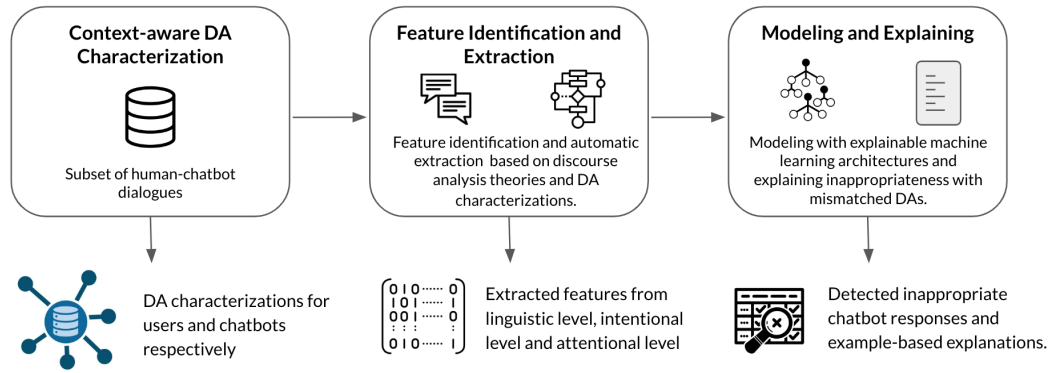
Xu Han
University of Colorado Boulder
Boulder, CO, USA
xuha2442@colorado.edu

Michelle Zhou
Juji, Inc.
San Jose, CA, USA
mzhou@acm.org

Yichen Wang
University of Colorado Boulder
Boulder, CO, USA
yichen.wang@colorado.edu

Wenxi Chen
Juji, Inc.
San Jose, CA, USA
wchen@juji-inc.com

Tom Yeh
University of Colorado Boulder
Boulder, CO, USA
tom.yeh@colorado.edu

Figure 1: The three-step modeling workflow for our proposed computational framework for Evaluating and Explaining Inappropriate Chatbot Responses

ABSTRACT

Evaluating and understanding the inappropriateness of chatbot behaviors can be challenging, particularly for chatbot designers without technical backgrounds. To democratize the debugging process of chatbot misbehaviors for non-technical designers, we propose a framework that leverages dialogue act (DA) modeling to automate the evaluation and explanation of chatbot response inappropriateness. The framework first produces characterizations of context-aware DAs based on discourse analysis theory and real-world human-chatbot transcripts. It then automatically extracts features to identify the appropriateness level of a response and can explain the causes of the inappropriate response by examining the DA mismatch between the response and its conversational context. Using interview chatbots as a testbed, our framework achieves comparable classification accuracy with higher explainability and fewer computational resources than the deep learning baseline, making it

the first step in utilizing DAs for chatbot response appropriateness evaluation and explanation.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.
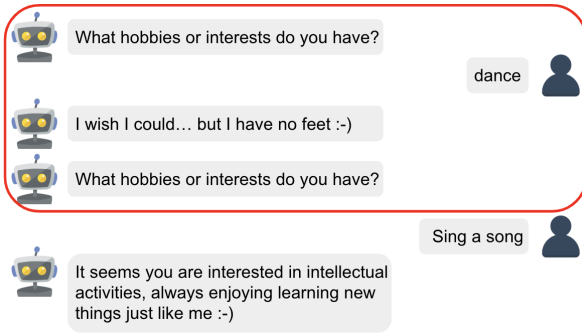
KEYWORDS

Conversational AI Agents; Interview Chatbot; Chatbot Debugging; Automatic Chatbot Evaluation and Explanation Framework

1 INTRODUCTION

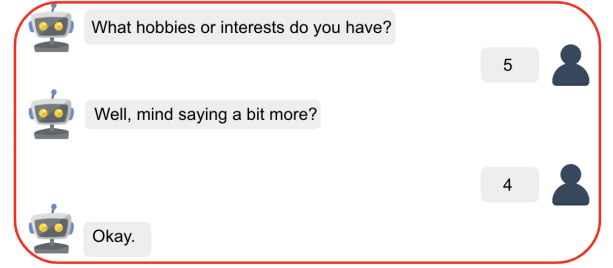Chatbot systems like ChatGPT[9] engage users in one-on-one text-based conversations by responding to user inputs correspondingly. While natural language generation (NLG) approaches, such as

(a)   (b)

Figure 2: Example human-chatbot dialogues. (a) The chatbot incorrectly recognizes the user's input ("*dance*") as a request and responds accordingly; (b) the chatbot mistakes the user's gibberish input ("*5*") for a legitimate answer and accepts it with an acknowledgment.

the use of large language models (LLMs), have made significant progress in generating syntactically well-formed chatbot responses [1, 18], it remains challenging to ensure that these responses are appropriate for the given conversational contexts [14]. The mismatch between the chatbot responses and the contexts can happen due to issues like context complexity[13], limitations in NLG model architectures[1], and dataset bias[15]. For example, the chatbot response *"I wish I could... but I have no feet :-)"* may be appropriate in the context of asking the chatbot to dance, but it's entirely inappropriate if the user indicates dancing as the hobby (Figure 2(a)). Such inappropriate responses (i.e., chatbot misbehavior) can lead to poor user experience or even abandoned conversations [7]. Therefore, it's critical for chatbot designers to ensure response appropriateness during the design processes.

In light of this, designers often conduct pilot studies to evaluate chatbot response appropriateness and iterate their designs accordingly (i.e., chatbot debugging) [7]. However, designers without technical backgrounds may face two challenges when it comes to **detecting** and **understanding** potentially inappropriate responses revealed by these studies. First, it can be difficult for them to detect inappropriate responses without adequate computational resources. Examining all chat transcripts collected from pilot studies to locate inappropriate responses, such as the example in Figure 2(a), is a laborious and time-consuming task to perform manually [7]. Even if designers opt to develop an automatic model for inappropriate response detection, they may be limited by a lack of access to training data and required computing power. Second, even if non-technical designers are able to locate all inappropriate responses, it can still be difficult for them to understand why they occur and how to address them. Given the wide variety of conversational contexts, chatbots can exhibit very different types of inappropriateness. For example, in Figure 2(a), the chatbot incorrectly recognizes the user's input (*"dance"*) as a request and responds accordingly, whereas in Figure 2(b), the chatbot mistakes the user's gibberish input for a legitimate answer and accepts it with an acknowledgment. This high degree of variability in chatbot inappropriateness can make it challenging for non-technical designers to understand and address them within a unified framework.

To democratize the debugging process of chatbot misbehaviors for non-technical designers, we propose a computational framework to evaluate and explain chatbot response inappropriateness through characterizing context-aware dialogue acts (DAs). Our framework draws inspiration from recent works that combine DA characterization and neural response generation tasks [12, 18, 19, 22]. These studies have shown the promise of utilizing DA modeling to enhance chatbot response quality, making them more controllable and interpretable. For example, Xu et al. [22] incorporate DAs as policies to improve their open-domain chatbot response generation model. With this in mind, our framework first guides the development of context-aware DA characterization of human-chatbot dialogues. Next, it identifies and extracts computational features based on the DA characterizations, and then trains automatic detection models to evaluate the appropriateness of a chatbot response. By utilizing DA characterization, our framework can explain the causes of inappropriate responses by examining the DA mismatch between the response and its conversational context.

To the best of our knowledge, our framework is the first to incorporate DA characterization into the evaluation and explanation of chatbot inappropriate responses. To test the framework, we used *interview chatbots* as a testbed and developed the first context-aware characterizations of DAs in human-interview chatbot interactions. It also achieved comparable accuracy in detecting inappropriate responses compared to the deep learning baseline, while offering greater explainability and requiring fewer computational resources.

2 METHODS

2.1 Testbed and Dataset

Testbed. To ensure practicality, we have selected interview chatbots as our testbed, given their widespread use in a variety of applications, including research and job interviews [20, 21]. Interview chatbots utilize generative AI technology to engage users in text-based, one-on-one conversations, making them an ideal testbed for our study. Specifically, they are suitable for our study for several reasons: firstly, they support both task-oriented and social dialogues, making them representative of current chatbot systems; secondly,

the dialogues between human and interview chatbots tend to follow a concise and controllable pattern of "*interview question* (from chatbot) - *answer* (from user) - *response* (from chatbot)" [7], which facilitates our analysis of response appropriateness. Importantly, findings from interview chatbots can potentially be generalized to other chatbot categories [7].

Table 1: Interview Topics Used in Our Dataset

Q1	What hobbies or interests do you have?
Q2	What do you do now for a living?
Q3	What are your strongest qualities as a friend?
Q4	Tell me about a time when you didn't know if you would make it. How did you overcome that challenge?

Dataset. We study real-world dialogues collected through the interview chatbots supported by Juji [1], a publicly available chatbot platform where chatbot designers can create, customize, and deploy a chatbot with either a graphical user interface (GUI) or an interactive development environment (IDE) [17]. We analyzed a dataset of 5342 real-world human-chatbot dialogues with 8987 chatbot responses in total, accumulated from various interview chatbots developed by Juji's designers, including personality survey bots [4]. These chatbots were active in the wild for dialogue transcript collection from February 2021 to July 2021. Each dialogue in the dataset was associated with one of the Juji built-in topics shown in Table 1. To ensure quality, we manually reviewed each dialogue, excluding those without any end-users inputs. The collection of these 5342 dialogues involves 2155 participants, most of whom are university students and their families with various backgrounds. For our study, we recruited two dialogue researchers to annotate all 8987 chatbot responses using three labels: *Inappropriate*, *Appropriate*, and *Neutral*. Overall, the two annotators had achieved an inter-annotator agreement of 0.795 (Cohen's κ), which indicates a level of substantial agreement. When there were disagreements, the two annotators resolved the disagreement together through a discussion.
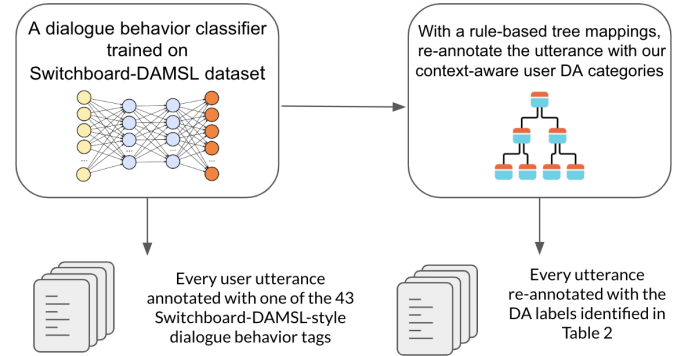
2.2 Computational Framework for Evaluating and Explaining Chatbot Response Inappropriateness

Our computational framework has two primary goals: 1) to provide a modeling workflow that enables non-technical designers to automatically detect and understand inappropriate chatbot responses; 2) to provide example-based explanations that facilitate a better understanding of chatbot inappropriateness. The framework addresses the problem of chatbot response appropriateness by formulating it as a three-class classification problem that distinguishes between appropriate, inappropriate, and neutral responses. Building upon prior research in DA characterization [2, 12, 16, 19, 22], discourse theories [5], and DA classification [13], we have developed a three-step modeling workflow for our proposed framework.

[1]https://juji.io/

This workflow consists of context-aware DA characterization, feature identification and automatic extraction with the characterized DAs, and modeling and explaining. An overview of our framework can be found in Figure 1. In the following sections, we provide a detailed description of the framework and illustrate it with the interview chatbot dataset as a case study.

2.2.1 Context-Aware DA Characterization in Human-Chatbot Interactions. DA characterization is to model a single utterance in a dialogue with functional tags which represent the communicative intentions behind it. Since the same utterance can reflect different intentions due to different contexts, determining the DA category of one utterance requires context-aware modeling based on the preceding and following context [13]. With this in mind, the first step of our framework is an open coding process [8] to investigate the DAs that are frequently associated with users and chatbots in different conversational contexts in the dataset. Specifically with our dataset, we analyzed a subset of the dataset consisting of dialogues associated with the four interview topics presented in Table 1. To achieve this, we randomly selected 100 dialogues belonging to each interview topic, and an expert evaluator manually annotated each utterance in the subset with a label that best describes its DAs considering the contexts. After analyzing the occurrences of DAs and grouping similar DAs into categories, we identified 12 user DAs and 14 chatbot DAs. Tables 2 and 3 present the DA characterization for users and chatbots in human-chatbot interactions, respectively.

Figure 3: The cascading method we used to realize DA auto-annotation for an utterance

2.2.2 Identifying and Automatically Extracting Features with DA Characterization. Drawing on previous discourse analysis theories by Grosz and Sidner [5], our framework identifies key features in human-chatbot dialogues from three different levels: *linguistic level*, *intentional level* and *attentional level*. At the *linguistic level*, our framework identifies specific linguistic markers, such as words or phrases, that contribute to the *shallow discourse structure* [2, 10, 16]. In our case study with interview chatbots, we utilized the interview topic's and target chatbot responses' unigram bag of words as *linguistic-level* features. Meanwhile, the *intentional level* captures the utterance-level DAs. We thus encoded the *intentional-level* features through one-hot categories of various dialogue components

Table 2: Context-aware Characterization of User DAs When Interacting with Interview Chatbots

DA Categories	Synopsis	Typical Examples *
user-answer-relevant	Giving relevant answers to the interview questions	*"S: What things frighten you now?"* *"U: **My future is the most terrifying.**"*
user-question-relevant	Questioning for further details or starting chitchat under the same topic	*"S: What things frighten you now?"* *"U: **Why are you asking ?**"*
user-respond-irrelevant	Responding with irrelevant information to the interview questions	*"S: What hobbies or interests do you have?"* *"U: **I like blue most.**"*
user-question-irrelevant	Questioning about different topics to start chitchat	*"S: What hobbies or interests do you have?"* *"U: **I like swimming. What are your capabilities?**"*
user-excuses	Dodging answering the interview questions/digression with various excuses	*"S: What were the worst parts of your childhood?"* *"U: **This is personal.**"*
user-acknowledge	Accepting or admitting the chatbot's utterances	*"U: How long is our chat gonna last?"* *"S: If part of the chat progress bar is still red, it indicates that our chat is still in progress. It will end before you know it."* *"U: **Got it!**"*
user-request	User's requests to the chatbot	*"U: **Tell me a joke.**"*
user-command	User's commands on managing the chat-flow	*"U: **Next question.**"*
user-complain	Complaining about the chatting experience or else	*"S: What do you do now for a living?"* *"U: **You didn't listen. I just answered it.**"*
user-social-obligations	Apology, greeting, thanking and etc.	*"S: I hear you... would love to help when I have the power to do so."* *"U: **Thank you!**"*
user-gibberish	user gives gibberish	*"S: What hobbies or interests do you have?"* *"U: **blea blahe**"*
user-other	Sentences do not belong to any of the categories above	*"S: What hobbies or interests do you have?"* *"U: **Wow.**"*

* Note: "S" denotes the chatbot system while "U" denotes the user.

Examples are for demonstration purposes only, not necessarily from the original transcripts.

including the target chatbot utterance, all previous chatbot utterances, all following chatbot utterances, the most recent user utterance before the target, all previous user utterances, the next user utterance after the target, all following user utterances. The *attentional level* models the dynamic focus of attention as the dialogue unfolds and the relationships between utterances, contributing to the *deep discourse structure* [2, 10, 16]. For simplicity, we utilized the user-chatbot exchange DA pairs and ordinal index of the target chatbot response to describe the *attentional-level* features.

Although most of the identified features mentioned above can be extracted computationally, the categorization of utterance-level user DAs still requires additional annotation efforts. To automate this process, we propose a two-stage cascading method for auto-annotating each user utterance's DAs (Figure 3). The first stage employs a dialogue behavior classifier that is trained on a large-scale open-sourced dataset, specifically the Switchboard-DAMSL dataset[10, 16], to assign Switchboard-DAMSL-style dialogue behavior tags to the utterances (pre-annotation). The Switchboard-DAMSL dataset contains a tag set of 43 mutually exclusive dialogue behaviors with the intention of building better language models for conversations. We directly utilized the dialogue behavior classifier trained by Raheja and Tereault [13] in this stage. Subsequently, in the second stage, these pre-annotated utterances are re-annotated automatically, following a rule-based tree mapping between the 43

Switchboard-DAMSL-style dialogue behavior tags and our characterized context-aware user DA categories from Table 2. An expert evaluator formulated the rule-based tree mapping manually following an open coding process [8]. The mapping was created with the same subset of the dataset used during the context-aware DA characterization phase.

2.2.3 Modeling and Explaining with DA Characterization. The framework then utilizes the extracted features to train classifiers for automatic detection of chatbot response appropriateness. To democratize the evaluation process, the framework opts to employ simple and interpretable machine learning models such as random forest (RF), instead of deep learning models that are resource-intensive and opaque. Following the common practice, the framework utilizes grid search to select the hyperparameters for the model. For the evaluation, the framework assesses model performance using four standard performance metrics, namely precision, recall, F1, and accuracy. To adjust for class imbalance, the framework weights all the metrics by the number of samples in each class when reporting the overall model performance. With the auto-annotated utterance DAs and detection results, the framework further examines the mismatch between the contextual utterances' DAs and the target chatbot response DAs to explain the inappropriateness. We present the results of modeling and explaining in the following section.

Table 3: Context-aware Characterization of Interview Chatbot DAs to Previous User Inputs

Chatbot Behavior Categories	Synopsis	Typical Examples *
chatbot-respond-relevant	Responding relevantly and empathetically to user's relevant answers or questions	*"U: I lost my dog when I was eight."* *"S:* ***Thanks for sharing.*** ***I'm sorry you had to go through that."***
chatbot-acknowledge	Accepting and admitting user's inputs	*"U: I don't like rollercoasters. "* *"S:* ***Okay. "***
chatbot-ask-followup	Asking follow-up questions based on recognized user intents	*"U: Least favorite was physics. "* *"S:* ***Do you mind saying a bit more? "***
chatbot-handle-user--question-irrelevant	Handling user's questions about topics different from the current interview topic	*"U: How are you feeling?"* *"S:* ***"Thank you for asking. I feel great!"***
chatbot-handle-user-excuses	Handling user's excuses to avoid answering the question/digression	*"S: "What were the worst parts of your childhood? "* *"U: This is personal. "* *"S:* ***"I understand. We'll continue then."***
chatbot-handle-user-request	Handling user's various requests	*"U: Tell me a joke. "* *"S:* ***Why was the math book sad?*** ***Because it had too many problems.."***
chatbot-handle-user-command	Handling user's commands on managing the chat-flow	*"U: I want to skip the current questions. "* *"S:* ***That's okay. Let's move on then."***
chatbot-echo-user--respond-irrelevant	Responding to user's irrelevant responses relevantly and empathetically	*"S: What do you do now for a living?"* *"U: I felt lonely sometimes."* *"S:* ***If you need urgent help, please call 911 or*** ***your doctor directly. I'd love to cheer you up if I could. "***
chatbot-handle-user-complain	Handling user's complaints	*"U: "You didn't listen. I just answered it. "* *"S:* ***Sorry, I must have missed it."***
chatbot-social-obligations	Handling user's acknowledging or social obligation inputs	*"U: Thank you. "* *"S:* ***"You're most welcome, {user's first name}.***
chatbot-respond-default-fallback	Not understanding user inputs and responding with default fallback messages	*"S:* ***My bad, I didn't recognize your inputs.*** ***Let's try again."***
chatbot-repeat	Not understanding user's answers and repeat the same utterance again	*"S: What things frighten you now?"* *"U: nothing."* *"S:* ***What things frighten you now?"***
chatbot-handle-gibberish	Handling user's gibberish	*"U:blea blahe"* *"S:* ***Sorry I didn't understand. Please use English."***
chatbot-other	Chatbot responses do not belong to any of the categories above	*"S:* ***Sorry I got disconnected. Let's continue."***

* Note: "S" denotes the chatbot system while "U" denotes the user.

Examples are for demostration purpose only, not necessarily from the original transcripts.

3 RESULTS

3.1 Modeling

Following the modeling practice in the framework, we trained an RF model to detect and explain the chatbot response inappropriateness. Using our interview chatbot dataset, we allocated 80% of the data to the training set and the remaining 20% to the test set. Our framework has achieved an accuracy of 91.0%. To further validate the effectiveness of our framework, we compared our RF model with a baseline model on the same dataset. The baseline is a RoBERTa classifier fine-tuned on our dataset, which is a complex deep learning model that has demonstrated top performance in many natural language processing (NLP) tasks [11]. We used 10% of the dialogues in the training set as the development set for hyperparameter selection. Table 4 shows the performance of the two models.

3.2 Explaining

Our detection model enables us to generate example-based explanations by examining the mismatch between the DAs behind the inappropriate chatbot response and the DAs behind the contextual utterances. These explanations remind chatbot designers of the probable causes of inappropriate responses generated by their chatbot designs. For instance, in Figure 4, we can observe that the chatbot incorrectly recognized the user's input (*"dance"*) as a "*user-request*" (from Table 2) and responded accordingly ("*chatbot-handle-user-request*" from Table 3). However, since the chatbot did not perceive the user's input as an answer to its question, it repeated the question ("*chatbot-repeat*" from Table 3), resulting in an inappropriate response.
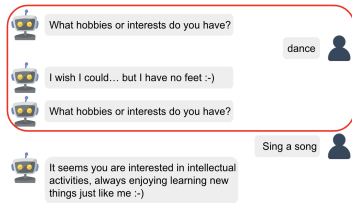
4 DISCUSSION

We discovered that our model achieved comparable performance to RoBERTa (91.0% vs. 90.6%) while utilizing fewer computational

Table 4: Evaluation Results of Chatbot Response Inappropriateness

Model	Class	Precision	Recall	F1	Accuracy	Required Computational Resources	Explainability
Fine-tuned RoBERTa (Baseline)	Inappropriate	0.810	0.781	0.795	0.906	High training/prediction & High storage efficiency ✗	Low ✗
	Neutral	0.942	0.931	0.936			
	Appropriate	0.909	0.965	0.936			
RF with our proposed framework	Inappropriate	0.789	0.819	0.804	**0.910**	**Low training/prediction & Low storage efficiency** ✔	**High** ✔
	Neutral	0.959	0.928	0.943			
	Appropriate	0.916	0.955	0.935			

Figure 4: An example-based explanation generated by our framework through examining the mismatch between the DAs behind the inappropriate chatbot response and the DAs behind the contextual utterances.

resources and offering greater model simplicity, resulting in higher interpretability. This comparison highlights the effectiveness of our proposed features and the potential of incorporating DA modeling in detecting inappropriate chatbot responses. During our experiments, RoBERTa required significantly more computational resources for both training and prediction than our model. It took us 1 hour and 52 minutes to fine-tune RoBERTa (4 epochs) on single NVIDIA Tesla K80 GPU and 1 minute and 49 seconds to make predictions, while our model required no specialized hardware and only needed 11.6 seconds to finish training and less than 0.5 seconds to make predictions. Our model's storage efficiency is also much higher than RoBERTa since RF's storage efficiency is proportional to the number of decision trees (500) in the ensemble and the maximum depth of each tree (45), whereas RoBERTa has hundreds of millions of parameters (123 million) that need to be stored. Additionally, the RF model's simpler architecture enables us to provide easy-to-interpret features and decision paths associated with specific chatbot responses. In contrast, RoBERTa is often considered a black box [3], making it difficult to interpret how it makes predictions. Benefiting from such high explainability, our framework offers example-based explanations with corresponding DA tags and contexts to guide chatbot designers in the next design iteration. With the explanations, chatbot designers can better understand the probable causes and devise appropriate strategies to fix any inappropriate responses that fall within the same mismatched DA categories. The comparison between our model and RoBERTa demonstrates that our framework can democratize chatbot inappropriateness debugging to non-technical users in terms of requiring

fewer computational resources and offering higher explainability while maintaining relatively good detection performance.

5 CONCLUSION AND FUTURE WORKS

Our findings indicate the feasibility and effectiveness of our proposed computational framework in evaluating and explaining chatbot inappropriateness. By incorporating DA modeling with just a simple RF model, our framework achieved comparable performance to top deep learning models while offering higher explainability and requiring fewer computational resources. In actual practice, our computational model can help chatbot designers identify the inappropriate responses from the pilot data and make corresponding revisions in further design iterations. These features make our framework an effective tool for non-technical chatbot designers to iteratively evaluate and improve their designs, which greatly democratizes the chatbot debugging process. However, we acknowledge some challenges and opportunities for future studies, such as:

- Exploring Framework Generalization Capability: While our results demonstrate promising performance in the context of interview chatbots, the generalizability of these findings to chatbots in diverse domains and with respect to other types of misbehaviors, such as toxic behaviors [6], remains uncertain. Additionally, it is worth investigating the necessary adaptations required to enhance the framework's applicability and generalizability.
- Interviewing Chatbot Designers: Since the target audience of our framework is non-technical chatbot designers, it is essential to test its usability and gather feedback from designers themselves to improve its practicality and effectiveness.
- Enhancing the Framework's Design-Assisting Capability: In its current stage, our framework provides example-based explanations of inappropriate chatbot responses with characterized DA tags and contexts. Inspired by previous work [7], we aim to provide more actionable design suggestions based on these examples to improve the democratization level of chatbot debugging.

REFERENCES

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* 1, 1 (2023), 1–155.

[2] Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, Vol. 56. Boston, MA, n.a., Rochester, NY, USA, 28–35.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[4] Jinyan Fan, Tianjun Sun, Jiayi Liu, Teng Zhao, Bo Zhang, Zheng Chen, Melissa Glorioso, and Elissa Hack. 2023. How Well Can an AI Chatbot Infer Personality? Examining Psychometric Properties of Machine-inferred Personality Scores. https://doi.org/10.31234/osf.io/pk2b7

[5] Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12, 3 (1986), 175–204. https://aclanthology.org/J86-3001

[6] Xu Han and Tom Yeh. 2019. Evaluating Voice Skills by Design Guidelines Using an Automatic Voice Crawler. *arXiv preprint arXiv:1906.01122* n.a. (2019), n.a.

[7] Xu Han, Michelle Zhou, Matthew J Turner, and Tom Yeh. 2021. 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*. Association for Computing Machinery, New York, NY, USA, 1–15.

[8] Judith A Holton. 2007. The coding process and its challenges. *The Sage handbook of grounded theory* 3 (2007), 265–289.

[9] OpenAI 2023. *Introducing ChatGPT*. OpenAI. https://openai.com/blog/chatgpt

[10] Dan Jurafsky. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report* 1, 1 (1997), 1–61.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* 1, 1 (2019), 1–13.

[12] César Montenegro, Asier López Zorrilla, Javier Mikel Olaso, Roberto Santana, Raquel Justo, Jose A Lozano, and María Inés Torres. 2019. A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. *Multimodal Technologies and Interaction* 3, 3 (2019), 52.

[13] Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3727–3733. https://doi.org/10.18653/v1/N19-1373

[14] Timo Spring, Jacky Casas, Karl Daher, Elena Mugellini, and Omar Abou Khaled. 2019. Empathic response generation in chatbots. In *Proceedings of 4th Swiss Text Analytics Conference (SwissText 2019), 18-19 June 2019, Wintherthur, Switzerland*. 18-19 June 2019, CEUR-WS, Sun SITE Central Europe, 1–10.

[15] Ramya Srinivasan and Ajay Chander. 2021. Biases in AI Systems: A Survey for Practitioners. *Queue* 19, 2 (may 2021), 45–64. https://doi.org/10.1145/3466132.3466134

[16] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26, 3 (2000), 339–373.

[17] Juji 2020. *Juji*. Juji. https://juji.io/no-coding-ai-chatbot-builder

[18] Anuradha Welivita and Pearl Pu. 2020. A Taxonomy of Empathetic Response Intents in Human Social Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 4886–4899. https://doi.org/10.18653/v1/2020.coling-main.429

[19] Chen Henry Wu, Yinhe Zheng, Yida Wang, Zhenyu Yang, and Minlie Huang. 2021. Semantic-Enhanced Explainable Finetuning for Open-Domain Dialogues. *arXiv preprint arXiv:2106.03065* 1, 1 (2021), 1–10.

[20] Ziang Xiao, Michelle X Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[21] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.

[22] Can Xu, Wei Wu, and Yu Wu. 2018. Towards explainable and controllable open domain dialogue generation with dialogue acts. *arXiv preprint arXiv:1807.07255* 1, 1 (2018), 1–15.