

# Automatic Personalized Evaluation

Jessica Huynh

CMU-LTI-26-008

April 2026

School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213

**Thesis Committee:**

Jeffrey P. Bigham (co-chair)

Fernando Diaz (co-chair)

Dilek Hakkani-Tür (University of Illinois Urbana-Champaign)

Ryen W. White (Microsoft Research)

Sherry Tongshuang Wu (Carnegie Mellon University)

*Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy.*

**Keywords:** dialog, personalized evaluation, personalization

*To my parents.*



## Abstract

Using large language models (LLMs) to simulate human ratings is a popular topic of study because human ratings are expensive and difficult to collect, but it is important to obtain a measurement of performance for consumer-facing dialog systems such as Alexa or Siri before releasing that system to the public. Objective measurements such as ensuring the system generates understandable text may not be enough as a measurement of performance, but these are metrics that can be, to some extent, automatically calculated. Subjective measures such as user satisfaction are just as important, if not more, because user satisfaction often drives user retention for these products. User satisfaction is derived from the system both satisfying the user’s needs and adjusting to the user’s preferences, and can vary greatly from user to user. Subjective measures often require human ratings, thus, LLMs are starting to be used as a proxy. The goal is to align these LLMs with humans such that the ratings are comparable to human ratings. Developers can then use these LLMs to both evaluate more accurately and develop better systems.

Through this work, we first explore the efficacy of using LLMs to rate dialogs, where we find that LLMs have the potential to rate more subjective measures such as interestingness. Next, we explore prompting as a way to achieve higher human and LLM rating agreement, and find that contextualizing prompts is important. We then further study user attitude injection into prompting with personalized evaluation through simulation to determine how much that may affect ratings. Lastly, we collect a dataset for personalized evaluation for further understanding of how LLMs can simulate human ratings such that they can be deployed to evaluate real systems.



## Acknowledgments

First, I would like to thank my parents for supporting me through my entire journey through the PhD. I always know that you are a phone call away for anything I need to talk about, and for that I cannot be more grateful. I would also like to thank my 公公 and 婆婆; I would not be who I am today without you.

I would also like to thank everyone who has been my advisor at some point in the PhD: Maxine Eskenazi, Jeffrey Bigham, Fernando Diaz, and Shinji Watanabe. Maxine, thank you for really teaching me how to think about research when I first started the Masters then the PhD - it has shaped a lot about who I am as a researcher. Jeff, thank you for sticking by me all throughout the PhD and reminding me to think about the big picture whenever I get stuck in the weeds. Fernando, thank you for teaching me about and enforcing the rigors of statistical testing that has become such a backbone for my research. Shinji, although we were only together for a short time, thank you for allowing me to collaborate and learn from you and many others in the lab. Additionally, thank you to the rest of my committee: Dilek Hakkani-Tür, Ryen White, and Sherry Wu. You all genuinely inspire me, and I greatly appreciate all the help you all have given me along the way.

I have also been extremely fortunate to have crossed paths with many wonderful people through my time at CMU, and I could not have done this without all of their support. This includes BIG Lab, Starlight, WAV Lab, all my collaborators across industry and academia, and all of the friends I have made along the way. A special shoutout to Cathy Jiao and Rosa Vitiello - you both made the office feel like a home away from home, and our cafe coworking sessions, dinners, and your unwavering friendship have meant more to me than I can put into words.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background and Related Work</b>	<b>5</b>
2.1	Autoraters . . . . .	5
2.1.1	What are they? . . . . .	5
2.1.2	How do you measure autorater performance? . . . . .	5
2.1.3	What makes a “good” autorater? . . . . .	6
2.1.4	Issues with Autoraters . . . . .	6
2.1.5	But why should we use them? . . . . .	7
2.2	Dialog . . . . .	7
2.2.1	Interaction and Evaluation . . . . .	7
2.2.2	User Models . . . . .	7
2.3	Personalization . . . . .	8
2.3.1	Conversation between Personalization and Personalized Evaluation	8
<b>3</b>	<b>How should we use autoraters?</b>	<b>11</b>
3.1	Introduction . . . . .	11
3.2	Understanding the Effectiveness of Very Large Language Models on Dialog Evaluation . . . . .	12
3.2.1	Introduction . . . . .	12
3.2.2	Related Work . . . . .	13
3.2.3	Evaluation Settings . . . . .	15
3.2.4	Experiments and Results . . . . .	17
3.2.5	Conclusion . . . . .	25
3.3	Quantifying the Statistical Effect of Rubric Modifications on Human-Autorater Agreement . . . . .	25
3.4	Introduction . . . . .	25
3.5	Related Work . . . . .	28
3.5.1	Rubrics in Context . . . . .	28
3.5.2	LLM-as-judges . . . . .	28
3.5.3	Automatic Essay Scoring (AES) and Instruction-Following (IF) . .	29
3.6	Experiment Setup . . . . .	29
3.6.1	Experiments . . . . .	30
3.7	Methods and Materials . . . . .	30

3.7.1	Datasets . . . . .	30
3.7.2	Autoraters . . . . .	31
3.7.3	Agreement Calculation . . . . .	32
3.8	Results . . . . .	34
3.8.1	Edited Rubrics . . . . .	34
3.8.2	Decomposition Level . . . . .	36
3.8.3	Agreement Level . . . . .	36
3.9	Conclusion . . . . .	37
<b>4</b>	<b>Autoraters for Simulated Personalized Evaluation</b>	<b>39</b>
4.1	Introduction . . . . .	39
4.2	Sensitivity of Automatic Dialog Evaluation with Simulated Attitudes . . . . .	40
4.2.1	Introduction . . . . .	40
4.2.2	Related Work . . . . .	41
4.2.3	Studying Personalized Evaluation in Dialog . . . . .	42
4.2.4	Methodology . . . . .	44
4.2.5	Results . . . . .	45
4.2.6	Conclusion . . . . .	60
<b>5</b>	<b>Personalized Evaluation in Real World Systems</b>	<b>63</b>
5.1	Introduction . . . . .	63
5.2	Personalized Evaluation in Natural Interaction . . . . .	63
5.2.1	Motivation . . . . .	63
5.2.2	Analytical Model . . . . .	65
5.2.3	Hypotheses . . . . .	67
5.2.4	Experiment Parameters . . . . .	67
5.2.5	Dataset Collection Methodology . . . . .	69
5.2.6	Analysis . . . . .	72
5.2.7	Results . . . . .	72
5.2.8	Conclusion . . . . .	86
<b>6</b>	<b>Conclusion</b>	<b>87</b>
<b>7</b>	<b>Appendix</b>	<b>91</b>
.1	Understanding the Effectiveness of Very Large Language Models on Dialog Evaluation Appendix . . . . .	92
.1.1	Prompt format examples FED . . . . .	92
.1.2	Prompt format examples DSTC10 . . . . .	93
.2	Influence of Prompt Design on Human-Autorater Agreement Appendix . . . . .	93
.3	Appendix A . . . . .	93
.3.1	Additional Other Observations . . . . .	93
.3.2	Agreement Level . . . . .	96
.4	Appendix B . . . . .	98
.4.1	Prompt for AES . . . . .	98

.4.2	Prompt for Instruction Following . . . . .	101
.5	Sensitivity of Automatic Dialog Evaluation with Simulated Attitudes Appendix . . . . .	107
.5.1	Limitations . . . . .	107
.5.2	Model Seeding . . . . .	107
.5.3	Attitude Definitions . . . . .	107
.5.4	Prompt for Conversation Creation - System . . . . .	109
.5.5	Prompt for Conversation Creation - User Static . . . . .	109
.5.6	Prompt for Conversation Creation - User Dynamic . . . . .	110
.5.7	Prompt for Conversation Rating . . . . .	111
.5.8	User Satisfaction Prompt . . . . .	111
.5.9	Believability . . . . .	113
.6	Personalized Evaluation In Real World Systems . . . . .	114
.6.1	Prompts for System Style . . . . .	114
.6.2	Prompts for the Tasks . . . . .	115

**Bibliography** **117**



# List of Figures

- 3.1 Large Language Models, comparison of select approximate sizes . . . . . 13
  
- 3.2 This diagram provides a walkthrough of the experimental setup as shown for automatic essay scoring. It represents comparisons made between human-  
autorater agreements  $\tau$  across holistic rubrics (left), in which all criteria are  
applied together in a single overall judgment, or analytic rubrics (right), in  
which criteria are evaluated separately, resulting in multiple scores. The  
original rubrics are given to a human and an autorater, while edited rubrics  
are given to autoraters only. Arrows in bold between  $\tau_1, \tau_2$  and  $\tau_3, \tau_4$   
represent comparisons for which statistical significance can be calculated.  
 $\Delta$ rater represents comparisons where the type of rater is changed while  
the type of rubric remains constant, while  $\Delta$ rubric represents comparisons  
where the type of rubric is changed and the type of rater remains constant. 26
  
- 4.1 Task-oriented and open-domain conversations are seeded for the simulated  
user and their initial attitude. Static conversations are generated with the  
user simulator using the initial attitude throughout the entire conversation,  
while dynamic conversations are generated with an attitude selector, which  
chooses an attitude based on the conversation history at each turn for the  
user simulator. The generated conversation is then rated both on the dialog  
and turn level, with and without an attitude. . . . . 43
  
- 4.2 All attitudes rating the same set of open-domain conversations over all  
dialog- and turn-level metrics. The number of attitudes (positive or nega-  
tive) that have significantly different ratings from the current attitude for  
each metric are counted. Each number represented in the graph corresponds  
to the counts for one of the metrics. The darker colored bars represent com-  
parisons between that attitude and all positive attitudes, while the lighter  
colored bars represent comparisons between that attitude and all negative  
attitudes. The x-axis represents each attitude, and the y-axis represents  
how many metric comparisons were statistically significant. . . . . 46

4.3	All attitudes rating the same set of task-oriented conversations over all dialog- and turn-level metrics. The number of attitudes (positive or negative) that have significantly different ratings from the current attitude for each metric are counted. Each number represented in the graph corresponds to the counts for one of the metrics. The darker colored bars represent comparisons between that attitude and all positive attitudes, while the lighter colored bars represent comparisons between that attitude and all negative attitudes. The x-axis represents each attitude, and the y-axis represents how many metric comparisons were statistically significant. . . . .	53
4.4	Differences between the ratings of the no attitude and attitude ratings calculated pairwise across all conversations for open-domain conversations in the static case for coherence and interestingness. The attitudes which have no significant differences between the no attitude ratings and attitude ratings when a paired t-test is performed have data points indicated in the lighter color, while attitudes with significant differences are indicated in the darker color. . . . .	54
4.5	All attitudes rating the same set of open-domain and task-oriented conversations over all dialog-level metrics on Llama. The number of attitudes (positive or negative) that have significantly different ratings from the current attitude for each metric are counted. Each number represented in the graph corresponds to the counts for one of the metrics. The darker colored bars represent comparisons between that attitude and all positive attitudes, while the lighter colored bars represent comparisons between that attitude and all negative attitudes. The x-axis represents each attitude, and the y-axis represents how many metric comparisons were statistically significant. . . . .	57
4.6	Differences between the ratings of the no attitude and attitude ratings calculated pairwise across all conversations for open-domain and task-oriented conversations in the static case for coherence on Llama. The attitudes which have no significant differences between the no attitude ratings and attitude ratings when a paired t-test is performed have data points indicated in the lighter color, while attitudes with significant differences are indicated in the darker color. . . . .	59
5.1	Analytical model of the study . . . . .	66
5.2	Preferences shown to human annotator. . . . .	70
5.3	Example of system response preference shown to human annotator. . . . .	71
5.4	Heatmap showing the normalized distributions of our study participants (users and human raters), over our four questions related to participant attitude toward chatbots for the empathy and sociability conditions. . . . .	73
5.5	Heatmap showing the distribution of preferences across our user data from 0 as not preferring empathetic or sociable systems at all to 4 as preferring empathetic or sociable systems in all cases. . . . .	74

5.6	Heatmap showing the normalized distributions of our study participants over whether being understood and empathy/sociability are important for this type of task. The heatmaps on the left represent the empathy condition, and the heatmaps on the right represent the sociability condition. . . . .	81
1	This diagram represents comparisons made between human-atorater agreements $\tau$ across various conditions such as holistic rubrics (left side), in which all criteria are applied together in a single overall judgment, or analytic rubrics (right side), in which criteria are evaluated separately, and rubrics which have been edited for LLMs. Arrow in bold between $\tau_5$ and $\tau_6$ represent comparisons for which statistical significance can be calculated. . . . .	94



# List of Tables

- 3.1 LLMs with the datasets they were trained on. During training or fine-tuning: “Seen Dialog” indicates that the model has explicitly seen dialog datasets and therefore elements of casual language, and “fine-tuned” indicates that the model was fine-tuned on dialog data. TNLGv2 has not seen datasets explicitly categorized as having dialog, but elements of casual language may be included in the Common Crawl snapshots and other internet-based corpora. Symbols: ✓ means that the category is included and × means that the category is not included. . . . . 14
- 3.2 Turn-level fine-grained metrics on the FED dataset for manually, randomly, and BM25 chosen examples over the TNLGv2 6.7B and 530B models. BM25<sub>C+R</sub> stands for examples chosen by BM25 considering both the context and the response of the test point. BM25<sub>C</sub> stands for examples chosen by BM25 considering the context and BM25<sub>R</sub> stands for examples chosen by BM25 considering the response. **Bold** values indicate the best value for the metric and \* values indicate correlations that are not statistically significant. . . . . 20
- 3.3 Dialog-level fine-grained metrics on the FED dataset for manually, randomly, and BM25 chosen examples over the TNLGv2 6.7B and 530B models. BM25<sub>C</sub> stands for examples chosen by BM25 considering only the context of the test point. . . . . 20
- 3.4 Turn-level fine-grained metrics on the FED dataset for manually chosen examples over the TNLGv2, BLOOM, OPT, Flan-T5, and InstructGPT models. . . . . 21
- 3.5 Dialog-level fine-grained metrics on the FED dataset for manually chosen examples over the TNLGv2, BLOOM, OPT, Flan-T5, and InstructGPT models. . . . . 21
- 3.6 Turn-level fine-grained metrics on the FED dataset for manually chosen examples over the smaller sizes of BLOOM and OPT. . . . . 21
- 3.7 Dialog-level fine-grained metrics on the FED dataset for manually chosen examples over the smaller sizes of BLOOM and OPT. . . . . 22

3.8	Spearman correlation of model predictions with human ratings for TNLGv2 6.7B model with algorithmically chosen examples. TU, PU, PZ, DZ, CG, DGU, DGR, EG, FT and FD are abbreviations for TopicalChat-USR, PersonaChat-USR [121], PersonaChat-Zhao [204], DailyDialog-Zhao [204], ConvAI2-GRADE [71], DailyDialog-Gupta [59], DailyDialog-GRADE [71], Empathetic-GRADE [71], FED-Turn and FED-Dial [120]. . . . .	23
3.9	Spearman correlation of model predictions for overall quality with human ratings for TNLGv2 530B model with algorithmically chosen examples. TU, PU, PZ, DZ, CG, DGU, DGR, EG, FT and FD are abbreviations for TopicalChat-USR, PersonaChat-USR [121], PersonaChat-Zhao [204], DailyDialog-Zhao [204], ConvAI2-GRADE [71], DailyDialog-Gupta [59], DailyDialog-GRADE [71], Empathetic-GRADE [71], FED-Turn and FED-Dial [120]. . .	24
3.10	Spearman correlation of model predictions for overall quality with human ratings with 4 examples chosen with BM25 using context. Macro average scores are also shown. . . . .	24
3.11	Kendall’s $\tau$ results on AES with GPT-4o and Llama for $\Delta$ rater. P. indicates the essay prompt, and C. indicates what ratings are being compared, with ideas, organization, word choice, sentence fluency, and conventions compared for prompt 1, and content, prompt adherence, language, and narrativity compared for prompts 4 and 6. $\tau$ is calculated with singular numerical values for $\Delta$ rater and calculated through Pareto dominance comparison for preferences for $\Delta$ rubric. Significance tests between separate (sep.), batch (bat.), and edited prompts are performed, where $s$ and $b$ in the edited prompt column represents significant differences with separate and batch prompts respectively. † is indicated next to comparisons that are significantly larger within holistic prompts. * is indicated next to comparisons that are significantly larger between separate and batch comparisons. † and ‡ represent that the $\tau$ value for edited prompts is significantly larger or smaller respectively with the separate $s$ or batch $b$ prompts’ $\tau$ . The lack of any dagger, star, or arrow denotes no statistical significance. H is a shortened form for Human. . . . .	33
3.12	Kendall’s $\tau$ results on IF with GPT-4o and Llama. $\Delta$ rubric is calculated through instruction following ratio comparison for preferences. All other calculations and significance follow the methodology of Table 3.11. . . . .	34

4.1	Statistical significance for attitudes rating the same conversations, created with no attitude indicators and rated by GPT-4o. OD represents open-domain, TO represents task-oriented conversations. The positive column represents significant differences between the given attitude’s ratings and other positive attitudes’ ratings for each metric. The negative column represents significant differences between the given attitude’s ratings and other negative attitudes’ ratings for each metric. Bold and starred entries represent entries that deviate greatly from the observation that positive attitudes are different from all other negative attitudes and vice versa. The four experimental conditions are outlined on the left, along with the list of metrics under each condition. . . . .	47
4.2	All attitude rating the same set of open-domain conversations. “+” signs indicate that the set of ratings for the attitude on the metric, when compared with each set of ratings for all positive attitudes, was found significantly different in every case. “-” signs indicate similarly for the set of ratings when compared with each set of ratings for all negative attitudes. . . . .	49
4.3	All attitude rating the same set of task-oriented conversations. “+” signs indicate that the set of ratings for the attitude on the metric, when compared with each set of ratings for all positive attitudes, was found significantly different in every case. “-” signs indicate similarly for the set of ratings when compared with each set of ratings for all negative attitudes. . . . .	50
4.4	Ratings for conversations for (no attitude, attitude) on GPT-4o for dialog-level metrics. Pairs of numbers in bold represent statistically significant differences when a paired t-test is conducted between no attitude ratings and attitude ratings for the same metric on the same set of conversations.	51
4.5	Ratings for conversations for (no attitude, attitude) on GPT-4o for turn-level metrics. Pairs of numbers in bold represent statistically significant differences when a paired t-test is conducted between no attitude ratings and attitude ratings for the same metric on the same set of conversations.	52
4.6	Statistical significance for attitudes rating the same conversations, created with no attitude indicators and rated by Llama2 70b. The positive column represents significant differences between the given attitude’s ratings and other positive attitudes’ ratings for each metric. The negative column represents significant differences between the given attitude’s ratings and other negative attitudes’ ratings for each metric. Bold and starred entries represent entries that deviate greatly from the observation that positive attitudes are different from all other negative attitudes and vice versa. The two experimental conditions are outlined on the left, along with the list of metrics under each condition. . . . .	58
4.7	All attitudes rating neutral conversation dialog opendomain, + = significant across all positive, - = significant across all negative - Llama . . . . .	58
4.8	All attitudes rating neutral conversation turn opendomain, + = significant across all positive, - = significant across all negative - Llama . . . . .	58

4.9	Ratings for conversations for (no attitude, attitude) on Llama2 70b annotations. Pairs of numbers in bold represent statistically significant differences when a paired t-test is conducted between no attitude ratings and attitude ratings for the same metric on the same set of conversations. . . . .	60
5.1	This is comparing between the same rater, with and without context. The reported numbers are the Cohen’s weighted $\kappa$ between the user ratings and the rater. $\uparrow$ represents which rating between with context and without context was significantly higher when comparing the bootstrapped distributions. Bold represents which rater out of human, GPT, or Llama performs significantly the best out of the three when keeping the context or lack thereof the same over the bootstrapped distributions. All comparisons between human, GPT, and Llama are significant with the exception of human/Llama without context with the enjoyment metric on the info-seeking task in the empathy condition (0.004, -0.006), GPT/Llama without context with the empathy metric on the writing task in the empathy condition (0.187, 0.229), and GPT/Llama with context with the sociability metric on the writing task in the sociability condition (-0.085, -0.058). . . . .	75
5.2	QWK between raters when context or no context is kept the same. The reported numbers are the QWK between the each of the two raters. $\uparrow$ represents if there is significance between the comparison of the raters with context and without context using bootstrapping. . . . .	77
5.3	Using user attitude survey questions with models to predict user annotations. Best refers to the best correlation from human raters, GPT, and Llama from Table 5.1, and best + c refers to the best correlation that was using context. The bold result represents the best correlation out of all values, and the * represents the best correlation out of all the models using the user attitude survey questions. . . . .	79
5.4	User and human rater normalized consistency across overall satisfaction and feeling understood. C stands for consistent, and C + R stands for consistent and resolved from asking the user and raters additional questions to resolve conflicts. - represents that there was no change after resolution. . . . .	84
5.5	User and human rater consistency across overall satisfaction and empathy/sociability preference. C stands for consistent, and C + R stands for consistent and resolved from asking the user and raters additional questions to resolve conflicts. - represents that there was no change after resolution. +pref represents adding the user’s preferences in mind - does overall satisfaction agree with the rater’s views on the more empathetic/sociable system and the user’s initial preference for empathy/sociability. . . . .	84

5.6	User and human rater consistency across overall satisfaction and empathy/sociability preference. C stands for consistent, and C + R stands for consistent and resolved from asking the user and raters additional questions to resolve conflicts. - represents that there was no change after resolution. This table answers if people are overall more satisfied with the systems we designed to be more empathetic/sociable. . . . .	85
5.7	Autorater consistency across overall satisfaction and empathy and understood preference. We report the autorater consistencies. . . . .	85
1	An example of a prompt with one example from FED [120]. Interestingness was rated in FED over three values corresponding to 0/2, 1/2, and 2/2. The resulting output is truncated to the integer value of 0, 1, or 2 to be used in evaluation. . . . .	92
2	An example of a prompt with examples from DSTC 10. . . . .	93
3	Kendall’s $\tau$ between GPT and Llama. All abbreviations follow Table 3.11. . . . .	95
4	Kendall’s $\tau$ results for $\Delta$ rater+rubric with GPT-4o and Llama on AES and IF. For AES, the number indicates the essay prompt, with ideas, organization, word choice, sentence fluency, and conventions compared for prompt 1, and content, prompt adherence, language, and narrativity compared for prompts 4 and 6. $\tau$ is calculated through Pareto dominance comparison for AES and through instruction following ratio comparison for IF. Significance tests between separate (sep.), batch (bat.), and edited prompts are performed, where $s$ and $b$ in the edited prompt column represents significant differences with separate and batch prompts respectively. $\dagger$ is indicated next to comparisons that are significantly larger within holistic prompts. $\star$ is indicated next to comparisons that are significantly larger between separate and batch comparisons. $\uparrow$ and $\downarrow$ represent that the $\tau$ value for edited prompts is significantly larger or smaller respectively with the separate $s$ or batch $b$ prompts’ $\tau$ . The lack of any dagger, star, or arrow denotes no statistical significance. H is a shortened form for Human. . . . .	95
5	ASAP Chosen Essay Prompts . . . . .	97
6	Kendall’s $\tau$ results on AES with GPT-4o and Llama. The number indicated by the $\dagger$ represents that the $\tau$ value on that subset of data (in all cases here, where raters agreed with each other) is significantly larger than the $\tau$ value on the other subset of data (where raters disagreed with each other). The lack of any arrow denotes no statistical significance. . . . .	97
7	Kendall’s $\tau$ results on IF with GPT-4o for $\Delta$ rater. Significance tests between data with full agreement, partial agreement, and full disagreement are performed, where $p$ and $f$ in the full agreement row represents significant differences with partial disagreement and full disagreement data respectively. $\uparrow$ and $\downarrow$ represent that the $\tau$ value is significantly larger or smaller respectively with compared $\tau$ . The lack of any arrow denotes no statistical significance. . . . .	98



# Chapter 1

## Introduction

With the proliferation of many consumer-focused systems that can book travel, chat, assist with other customer service tasks, or be a general purpose system like ChatGPT <sup>1</sup>, it is more important than ever to understand how users interact with these systems. Free-form chat has become a natural form of communication with these products, as opposed to some phone agents which only offer multiple choice options for structured conversations. Successful systems include ELIZA in the 1960s [186] (where keywords and rule-based methods governed ELIZA’s responses), pre-LLM chatbots [2, 211] (templates or other restrictions on conversation), and now Microsoft CoPilot <sup>2</sup> (where users can use natural language to interact with language models that can connect to and modify information in Microsoft apps).

However, chatting using natural language presents additional issues for a system as compared to interaction with a pre-defined pattern. These issues can vary in importance depending on the type of system and the goal of the user. Goals can consist of entertainment [16] with open-domain chit-chat bots, conversational recommendations [79] with conversational recommender systems, and information-seeking [16, 66] with task-oriented chatbots. Many older systems rely on a pipeline architecture that has separate modules for natural language understanding, dialog management, and natural language generation [27, 127]. The systems must understand the intent of the user through the intent classification task and extract the important information from the conversation through the slot-filling task using datasets such as ATIS [66] and then execute a set of steps to complete the user’s request. These tasks can be evaluated through objective measures such as F1, accuracy, and task completion. Older systems used pre-written templates for text generation [127], which heavily restricted the diversity of system outputs. The evaluation of these systems can be broken down into individual components, with system parts being interchangeable and able to be developed separately.

The introduction of language models, especially large language models (LLMs) with enhanced capabilities, has changed the dialog space. Instead of pipeline systems, work has moved toward end-to-end neural systems for task-oriented conversations with GPT-2 [63],

<sup>1</sup><https://chatgpt.com/>

<sup>2</sup><https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>

or more recently, general purpose chat models like ChatGPT and GPT-4 [129]. The general purpose chat models have increasingly higher performance on language and reasoning benchmarks like HellaSwag [195], but bring additional issues such as hallucination during generation. With these end-to-end systems, evaluation has moved from objective measures of system components to focusing on subjective measures of the overall system generation. Measures of interestingness or relevance [120] are now possible with generations from LLMs.

These systems (both pipeline and end-to-end) can also be personalized for improved user satisfaction [95] which would lead to higher user retention in these products [14]. We define personalization in this thesis as referring to a system using explicit information given by a user or implicit information taken from conversation history to modify its generation. Similar definitions of personalization have been used in marketing for customers [138]. A further specification is commercial personalization, defined in Fan and Poole as varying the content for individuals, and segmentation, defined in Fan and Poole as focusing on specific parts of the population. This personal experience has been shown to be a desire from users of chatbots [15]. Most recently, ChatGPT introduced a new feature called “Memory”<sup>3</sup> where users can customize information that ChatGPT can use to tailor responses. Previous work has also suggested using natural language user profiles in recommendation systems [143]. However, these systems have been evaluated without additional knowledge of who the users are - Bowden and Walker evaluates system personalization with real user interactions and ratings, but these users are completely anonymous. Therefore, although insights from system personalization are useful for improving user experience overall, it is often unknown how system personalization affects individual users’ experiences.

Users provide a large amount of information that can be used during personalization; various algorithms may take advantage of different pieces of that information [67]. However, testing these algorithms would require obtaining real user ratings for user satisfaction as a proxy for user retention, and it would be extremely expensive and time-consuming to collect user satisfaction ratings for every algorithm to determine the best-performing one. Also, not every user may want to rate their interaction with a system [171]. Therefore, because of the increasing capabilities of LLMs, there has been a growing interest in using LLMs to not only generate text, but to also rate text in place of human annotators (LLM-as-a-judge [207] or autoraters). Using LLMs to rate is less expensive and quicker [30], but it is important that the autoraters align with human raters to provide a user accurate assessment of the system. Additional model biases [207] may affect this alignment. Typically, the performance of an autorater is determined by the correlation with a small set of existing human ratings. High correlation suggests better ability of the autorater to mimic human ratings. A well-performing autorater would allow for developers to have an idea of system performance and quickly iterate on their systems before consumer deployment.

A natural next question given the user of LLMs for both system personalization and rating is if it is possible to use LLMs to simulate ratings based on individual users - automatic personalized evaluation. Personalization of a system and personalized evaluation, which refers to using humans to rate similarly to individual users, have different goals,

<sup>3</sup><https://help.openai.com/en/articles/8590148-memory-faq>

which will be discussed in Section 2.3.1. Challenges for personalized evaluation include the aforementioned difficulty of collecting real user ratings, and also the extent to which segmentation is performed. To explore automatic personalized evaluation, we explore three main areas: the soundness of using autoraters as a proxy for human judgment, automatic personalized evaluation in a simulated setting, and personalization and automatic personalized evaluation in real world systems.

Chapter 3 discusses the efficacy of using autoraters as proxies for human judgments. First, we will compare the performance of various autoraters for open-domain dialog evaluation in section 3.2. Next, we study how prompt variation affects human-autorater alignment in the automatic essay scoring and instruction following domains in section 3.3. Chapter 4 studies automatic personalized evaluation in a simulated setting. Section 4.2 creates simulated conversations with injected user attitudes and quantitatively analyzes the differences using user attitude during autorating. Lastly, Chapter 5 focuses on personalized and automatic personalized evaluation in real world systems.

Through developing an evaluation framework for studying automatic personalized evaluation, collecting a dataset, and performing subsequent quantitative analysis, we see that **automatic personalized evaluation mostly provides ratings significantly more similar to user ratings for system performance as compared to personalized evaluation or evaluation without context**, in real world dialog systems, dependent on the task and contextual user information.



# Chapter 2

## Background and Related Work

### 2.1 Autoraters

#### 2.1.1 What are they?

Through this work we define autoraters (or LLM-as-a-judge [207]) as LLMs that take natural language prompts as input along with a system output, and rate the system output based on some criteria. Researchers use autoraters to compare outputs from systems A and B, allowing them to determine which output should be selected, and, by extension, which system should be used. The ratings can be in the integer numerical form (ex. between 1 and 5), preference form (ex. output A < output B), and can include additional reasoning about the response. Additionally, autoraters provide the log-probabilities of the potential generated tokens, which makes it possible to take a weighted sum of the log probabilities with the ratings [74] to obtain a more accurate measure of what the autorater believes the rating to be. Other work has looked at using a suite of autoraters to obtain a final judgment [175]. We distinguish autoraters from models that return a rating *without* natural language prompting - for example through training data such as COMET [145].

#### 2.1.2 How do you measure autorater performance?

Autorater performance is typically measured as correlation with human numerical ratings or similarity with human rationale for answer choices. For numerical ratings, we use correlation metrics such as Pearson’s correlation (ex. [50]) or Spearman’s coefficient (ex. [120]). These metrics provide a direct comparison between the ratings from the autorater and the humans. There are several reasons why different correlation or agreement metrics may be picked in the numerical case. Pearson correlation is for determining a linear relationship between two groups - which is useful if the autorater ratings have a simple linear relationship with human ratings. However, this is often not the case. Therefore, Kendall’s Tau [88] may be a suitable alternative because it calculates ratios using the concordant, discordant, and tied pairs of data points - which results in calculating preference agreement between humans and autoraters. This may be useful when the relationship between the autorater and human ratings is unclear.

Humans and autoraters can also align in the rationale (explanations) that they provide for certain ratings. This is important because if humans and autoraters use the same explanations for the same ratings, we would trust the autorater performance more than if they did not use the same explanations for the same ratings. Fayyaz et al. found that GPT-4-Turbo had higher alignment than other LLMs when computing F1 score over human and autorater explanations. Other works study the possibility that the rationale provided by the autorater is not what the autorater uses to obtain answers (faithfulness) [10, 111], which may be a problem. Chain-of-thought [185] is also often used as a tool to improve human and autorater alignment on tasks such as arithmetic and commonsense reasoning. It is important to ensure that alignment with numerical ratings and rationale is achieved to truly trust autoraters as a proxy for human judgment.

### 2.1.3 What makes a “good” autorater?

An autorater should be both consistent and aligned with humans. Consistency consists of several parts: robustness to output variation [194], prompt sensitivity [194], and process consistency [19]. Robustness to output variation refers to how the ratings of an autorater changes based on small variations to the system output that is being evaluated (which could include typos [194]). Prompt sensitivity refers to how the ratings of an autorater changes based on small variations to the instructions given to the autorater (including synonyms and typos [194]). Lastly, Bubeck et al. discusses how GPT-4 excels at being output-consistent (generating plausible explanations given the input, output, and context), but falls short in being process-consistent (given different inputs and contexts, the output should be consistent with the output given by following the logic of the model explanation). They note that the task should both be easily explainable and be able to be simulated properly, while the explanation should make sense such that process-consistency can be achieved. With consistency and alignment, we would be able to confidently use autoraters in deployment settings.

### 2.1.4 Issues with Autoraters

Including the consistency and faithfulness issues mentioned, there are unsolved autorater bias issues including egocentric bias [207], lost-in-the-middle context forgetting [105], position bias [107, 181], selection bias [206], and verbosity bias [207], which affect the autorater judgment based on prompt style and output style preferences.

Egocentric bias (or self-enhancement bias) is defined in [207] as “the effect that LLM judges may favor the answers generated by themselves”. By not correcting for egocentric bias, some outputs being judged may have inflated scores. Typically, this is solved by using a model that was not used to generate any system outputs that are being evaluated. This issue can also be extended to models in the same family.

Position bias refers to the effect that the order of system outputs presented for comparison affects the decision an autorater makes [107, 181]. Selection bias studies this affect with multiple choice questions and the order of options presented [206]. Lastly, context forgetting refers to information that is in the middle of the prompt may be lost, which

could cause an autorater to place more emphasis on evaluation criteria that were noted at the beginning or end of a prompt. This has been studied and shown in relation to giving models long contexts in multi-document question answering [105].

Works using autoraters should be aware of and ensure that the task setup does not incorporate these biases before real world deployment.

### **2.1.5 But why should we use them?**

With the plethora of considerations and issues that arise when using autoraters, there are still a multitude of reasons why autoraters should be used. Autoraters are scalable, cost-effective compared to human annotations, and if configured properly to reduce bias and increase consistency, can be a strong indicator of whether a system performs up to some standard before being released for human use.

## **2.2 Dialog**

### **2.2.1 Interaction and Evaluation**

The main difference between dialog and other tasks is the concept of “interaction”. Interaction means that user goals and satisfaction can shift within a conversation, and that the conversational context affects the current conversational turn. This leads to evaluation of both a system’s output turn-by-turn, as well as the system’s performance overall [120]. Additionally, the evaluation of any system output is not straightforward - any conversational context can have multiple “correct” responses, which is the one-to-many problem. Therefore, it is important to incorporate the aspect of interaction when studying dialog evaluation.

Dialog evaluation can be done both with humans and models. Human evaluation has been done in both the first, second, and third-party setting with debates and emotion annotations, and Park et al. showed some significant differences between the first and second party ratings. Sentiment annotations between first-party and third-party ratings also show differences [91]. Model evaluation has consisted of analyzing a model’s next response to the dialog [120] to using autoraters [74].

### **2.2.2 User Models**

Within interaction, it is also important to understand user models. Each user interacting with the system may have a different internal and external model of themselves and the world, which affects the conversation they have with the system and perhaps their evaluation of the system. These models can either be interpretable or not.

An interpretable user model is defined as “explicit assumptions on all aspects of the user that may be relevant for the dialog behavior of the system”, and can consist of “general knowledge, belief about individual facts, and goals” [178]. These can be represented by enumerable relationships or criteria [178] or through query and document weights [159].

The benefit of these approaches is that a system can pre-determine what information it will receive about the user and plan generation accordingly, and this user model will be standardized across users and visible to external system developers. In practice, explicit user modeling and the use of user models to adapt systems is useful [77]. Some works have used stereotyping as a way to begin user modeling [146, 161]. Additional adaptation to individual users improves performance over stereotyping users due to the potential non-homogeneous behavior or incorrect stereotyping [77]. Uninterpretable models learn user embeddings [103]. However, implicit models have added flexibility, as the model does not need to know what kinds of explicit information it needs to obtain from the user. As referenced in 1, we will be focusing on explicit user modeling.

However, user modeling has some important considerations - awareness of the user to the user modeling, inaccurate representations, and misuse and the legality of using and keeping user information [90]. Throughout this thesis, we choose to use interpretable user models to determine trends among users with similar user models and understand how that affects their system evaluation.

## 2.3 Personalization

Personalization can have multiple definitions. One definition of personalization may be applied in the RLHF setting, where feedback from the users can be used to train the model to have better performance [130]. Different parts of a dialog system can be personalized as well - the data that a system is trained on (ex. a user's data can be used to fine-tune a generic model for personalized performance), or the algorithm (ex. a sub-criteria could be weighted different for two users when calculating a final score for system performance). We will define personalization as the user data that affects model output, keeping algorithms the same, and focus on user segmentation as mentioned in Chapter 3.

Some works have studied how far personalization can improve performance with objective metrics. Teevan et al. studies the "potential for personalization", in which they re-rank search results based on a joint discounted cumulative gain to determine what the best performance of a non-personalized system could be over a group of users, and calculate the difference between that performance and the performance of personalized rankings for each user. Knowing that personalization has the potential to improve performance, personalized evaluation seeks a similar goal - understanding how personalized metrics can improve metric accuracy over a metric used over all users.

### 2.3.1 Conversation between Personalization and Personalized Evaluation

Personalization and personalized evaluation are two inter-related but different aspects. In prior sections, we discussed personalization within systems, and how user modeling has been used heavily to create this personalization. Due to human interaction relying heavily on context and user modeling, evaluation should also be personalized. [9] introduces the concept of personalized metrics, which edit online metrics based on a long period of online

individual user history. This work focuses on offline user information - personalized evaluation is the act of accounting for contextual information about the user during evaluation. Personalized evaluation can be applied to both generic and personalized systems. The difference between personalized systems and personalized evaluation is that personalized systems are often holistically scored - raters provide scores for system interactions, and the scores are analyzed holistically for an assessment of system performance. Personalized evaluation takes into account who the rater that is providing scores for the system interactions is - this could be individual information, or group belonging information. This is grounded in the context of autorating, where information about a user is injected into the prompt.

We can imagine a scenario of a generic chatbot with two users - an expert on machine learning who enjoys concise answers to their questions, and a novice on machine learning who enjoys verbose answers to their questions. These users rate the chatbot for conciseness and fluency. In the case of generic evaluation, a developer may take these user ratings at surface level, and create an autorater that has moderate correlation with the human ratings as a whole. However, with personalized evaluation, a developer can inject the preferences of the expert or novice into the autorater, which would create higher correlation with the human ratings - it is hypothesized that intra-group variance of ratings for a metric is lower than inter-group variance, where the groups are based on the preferences of the users for specific stylistic traits.



# Chapter 3

## How should we use autoraters?

### 3.1 Introduction

In Chapter 2, we discussed autoraters, and noted that one important condition to using autoraters as a proxy for human judgment is to ensure numerical alignment between humans and autoraters. There are a plethora of LLMs of different sizes with different use cases that can be used as autoraters, including BLOOM [188] - a large multilingual model, ChatGPT - a consumer-facing multi-purpose agent, and GPT-4o<sup>1</sup> - a multi-modal large model. The wording, examples, and the task being evaluated can all affect the prompting of these LLMs as well. Therefore, it is important to experiment which LLMs along with what kinds of prompts are suitable to achieve high numerical agreement for the task in question.

In this chapter, we explore the efficacy of using autoraters for dialog evaluation and other tasks. First, we find that even before the proliferation of ChatGPT, InstructGPT [130] could perform dialog evaluation tasks. However, larger models do not always mean better performance on dialog evaluation - rather the training data and examples given drive better performance. Next, we explore how prompting can further affect autorating performance for automatic essay scoring and instruction following. The evaluation criteria, examples, and complexity of the evaluation all factor in to autorater performance. We conclude by stating in order to start using autoraters in production settings, we must carefully construct prompts such that autoraters and humans both align in their evaluation criteria and have high correlation in ratings.

<sup>1</sup><https://platform.openai.com/docs/models/gpt-4o>

## 3.2 Understanding the Effectiveness of Very Large Language Models on Dialog Evaluation

### 3.2.1 Introduction

In recent years, language models such as GPT-3 [18] have grown larger, and their performance on downstream natural language processing (NLP) tasks has significantly improved in low-resource settings where only a few instances per task are available (few-shot). The larger these models are, the higher their performances trend on tasks such as language generation and evaluation [184]. They can generate coherent, fluent and interesting responses. However, they can also produce responses that are repetitive and un-engaging [151], in addition to being hard to control. Dialog evaluation is the task of assessing the quality of responses generated by dialog models in terms of properties like those mentioned above. However, one significant impediment for open-domain dialog generation research is the lack of meaningful automatic metrics for open-domain dialog evaluation. Standard language generation metrics have been shown to be ineffective for dialog evaluation [59], a large part of which is because conversations can be followed by *multiple valid* responses. Standard automatic metrics (e.g. BLEU [133]), which use references for evaluation, cannot deal with this quality, known as the *one-to-many* response problem. Many recently introduced automatic metrics for dialog evaluation [60, 120] have attained increasingly stronger correlations with human judgment. Since human dialog evaluation typically measures multiple fine-grained properties (e.g. appropriate, interesting, consistent), automatic evaluation metrics should be expected to do the same. This paper explores several fine-grained metrics that are measured both at turn-level (i.e. relevance and fluency), and dialog-level (i.e. consistency and coherence).

Automatic dialog evaluation continues to be an evolving topic, but with fine-grained metrics and definitions varying across different human-annotated datasets [121, 204], it is important to be able to create reasonable automatic metrics with limited data. Large language models (LLMs) that have been pre-trained on large-scale datasets are able to perform zero and few-shot inference [142, 155], and they have exhibited good reasoning skills [18, 184] in addition to having implicitly learned some notion of dialog quality [120]. This makes them suitable for open-domain dialog evaluation in zero-shot and extreme few-shot settings. While there have been a few attempts to use LLMs for dialog evaluation [170], there has not, to our knowledge, been a systematic study of LLMs for this task. This paper explores several aspects of LLM use in dialog evaluation: the effect of model type and size and the choice of training data as well as the use of in-context examples for dialog evaluation (the number and quality of the examples used). The experiments herein employ benchmarks to test both how well LLMs can be used for fine-grained evaluation, and how generalizable the models' performance is across multiple domains and datasets.

### 3.2.2 Related Work

#### LLMs

Several LLMs have been released recently: T5 [144], GPT-3 [18], BLOOM [188], OPT [199], and TNLGv2 [163]. The following models, the sizes of which are shown in Figure 3.1, are explored here:

- T5, trained on the 750B Colossal Clean Crawled Corpus (C4) contains heuristically cleaned natural language English text from the web. Specific models considered are:
  - Flan-T5 [34], T5 fine-tuned on 1836 tasks, including dialog tasks and data.
  - InstructDial [61], T5 fine-tuned specifically on 48 dialog tasks.
- GPT-3 includes a 570B filtered CommonCrawl corpus [144] in addition to WebText [142], Books1, Books2, and Wikipedia [87].
  - InstructGPT (text-davinci-002) [130], GPT-3 fine-tuned with a prompting dataset and 175B parameters.
- BLOOM was trained on 46 languages and 13 programming languages with a multi-lingual focus.
- OPT contains data from the RoBERTa corpus [106], the Pile [53], and PushShift.io Reddit [11, 151].
- TNLGv2 is trained on a subset of the Pile (notably excluding corpora classified as having natural dialog), two CommonCrawl snapshots [144], RealNews [196], and CC-Stories [173].

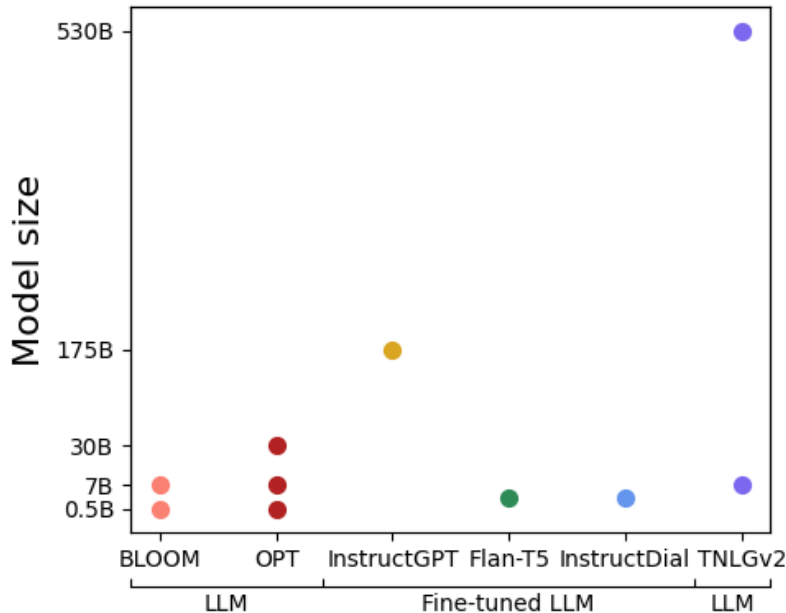


Figure 3.1: Large Language Models, comparison of select approximate sizes

	Seen Dialog Fine-tuned	
Flan-T5	✓	✓
InstructDial	✓	✓
InstructGPT	✓	✓
BLOOM	✓	×
OPT	✓	×
TNLGv2	×	×

Table 3.1: LLMs with the datasets they were trained on. During training or fine-tuning: “Seen Dialog” indicates that the model has explicitly seen dialog datasets and therefore elements of casual language, and “fine-tuned” indicates that the model was fine-tuned on dialog data. TNLGv2 has not seen datasets explicitly categorized as having dialog, but elements of casual language may be included in the Common Crawl snapshots and other internet-based corpora. Symbols: ✓ means that the category is included and × means that the category is not included.

As the number of parameters in these models increases, performance also increases: TNLGv2 530B, with around three times the number of parameters, outperforms the original GPT-3 on a variety of NLP tasks [163]. LLMs are also generalizable; they perform well on many NLP tasks in few-shot settings and zero-shot settings [155, 184]. However, several drawbacks and areas for exploration remain for LLMs that should be noted. Recent work has shown that performance on certain zero-shot tasks plateaus as model parameter size grows exponentially [18]. LLMs also struggle with parsing social situations [156] and correctly using context [3], which are important in dialog settings. This raises questions on the performance of LLMs for dialog evaluation, and how an LLM’s performance changes as it increases in size.

The data that a model is trained on also influences the performance of downstream tasks. T5 is fine-tuned on various subtasks, but pre-trained with C4. When pre-trained with domain-specific data, T5 performs better on tasks in that domain [12, 144]. Furthermore, adding several domains of data during pre-training makes the model likely to perform better [32, 106, 199]. Notably, BLOOM, OPT, Flan-T5, InstructGPT, and InstructDial are partially trained or fine-tuned on dialog datasets. Table 3.1 shows details on the content of these datasets. This is important because natural dialog data is difficult to obtain, so either scripted conversations or Reddit threads are used since they are the most readily available. This dearth of data is the reason that few-shot prompting is of interest. While work such as [184] acknowledges emergent abilities in larger language models in few-shot prompting settings, this paper explores discrepancies in performance specifically for dialog evaluation.

## Dialog Evaluation

Dialog evaluation presents a unique combination of challenges; it must consider multiple speakers [202], context that informs the current dialog turn, and the one-to-many aspect

mentioned above [203].

Metrics such as USR [121] and FED [120] were created to address some of these challenges; they are reference-free, capture complex aspects of dialog, and have good correlation with human evaluation. These metrics use models such as RoBERTa (125 million parameters) [106] and DialoGPT (345 or 762 million parameters) [201] respectively. However, the best performing versions of these models are smaller than most models examined in this paper, and are fine-tuned on dialog data or on a specific dialog task. Other automatic evaluation metrics include GRADE [71] and DEB [154]. With current LLMs’ large increase in hyperparameters, their plethora of training data, and their promising generalizable performance on NLP tasks, these model-based metrics should improve as well.

### Example selection for few-shot learning

The example selection process for prompting LLMs is of great interest. Prompting an LLM with a task and a few examples enables the model to adapt to a new task without completely fine-tuning it. In particular, in-context examples can provide important cues to help LLMs make predictions on tasks. Recent work has used a variety of methods to examine example selection. Common methods measure semantic similarity between example embeddings [104, 165]. Alternatively, retrieval methods (e.g. BM25 [147]) have been used directly, or as a precursor to training a selection retriever [152].

These example selection methods have shown promise in few-shot NLP tasks. In [165], the two-step framework for annotating and selecting in-context examples from large unlabeled data showed competitive performance across 10 tasks such as classification, common-sense reasoning, dialog state tracking, and code generation. [104] showed that selecting examples with similar sentence embeddings yields higher GPT-3 performance than random selection. However, the authors acknowledge that further investigation is required to find more efficient in-context example retrieval methods.

Moreover, the wording and order of examples presented in prompts can also affect model performance [54, 81, 104]. Lu et al. [109] observed order sensitivity across 0.1B to 175B parameter GPT-2 and GPT-3 models when the models were probed with different text classification tasks and several in-context examples. Also, the wording of the in-context examples depends on the data used for model training; for unfamiliar prompt formats, model performance may decrease [81]. Increasing the size of the model and the amount of data does not resolve the issue since the same instability is still prevalent [205]. Thus this paper studies the effect of example selection on dialog evaluation.

### 3.2.3 Evaluation Settings

Two settings for dialog evaluation are explored: fine-grained evaluation and multi-domain evaluation. In-context examples are explored in both.

## Fine-Grained Evaluation

Fine-grained metrics can be measured at both the turn level (e.g. informativeness and relevance), and the dialog level (e.g. coherence and diversity). The FED dataset [120] is used. It consists of 124 open-domain dialogs of humans with humans or with machines, for which each dialog has 3 responses that are chosen for annotation (8 turn-level and 10 dialog-level qualities along with overall turn- and dialog-level quality). This dataset was chosen due to the large number of previously studied fine-grained qualities as listed in Section 3.2.4, with the exception of correctness and error recovery, which are only specifically present in FED.

In the experiments, the LM is prompted to output a rating (an integer value - see Appendix .1.1) to evaluate each fine-grained quality in a response. The final rating for each fine-grained quality is a weighted sum of the  $K$ -top ratings outputted from the LM. Formally, given the  $K$ -top predicted ratings  $r_1, r_2, \dots, r_K$  along with their corresponding log probabilities,  $p_1, \dots, p_K$ , the weight,  $w_i$ , of each rating  $r_i$  is derived as:

$$w_i = \frac{p_i}{\sum_{j=1}^K p_j}$$

The final rating,  $r$ , is calculated as:

$$r = \sum_{i=1}^K r_i * w_i$$

In order to provide a more accurate view of the LM’s performance,  $K = 3$  in the following experiments. Additionally, this scoring mechanism converts the LM predictions onto a continuous scale, which more closely mirrors the average of human ratings. Results are reported with the Spearman correlations to the average human ratings for each fine-grained quality.

## Multi-domain Evaluation

This task tests automatic dialog evaluation metrics for robustness across multiple dialog domains. The analysis uses only the overall quality metric since many of the domain datasets do not have fine-grained annotations. The Spearman correlation is used between human ratings and model predictions on the evaluation sets released by DSTC 10 Track 5 [197] “Automatic Evaluation and Moderation of Open-domain Dialogue Systems”. These sets contain human judgement ratings for dialog responses. In this setting, a model is shown a dialog context and a response, and it outputs “yes” if the response is a good response to that context, otherwise it outputs “no”. An example can be seen in Appendix .1.2. The probability of the “goodness” of the response (i.e., the rating),  $g$ , is calculated as:

$$g = \frac{p_{model}(yes)}{p_{model}(yes) + p_{model}(no)}$$

where  $p_{model}(yes)$  and  $p_{model}(no)$  are the log probabilities of the model outputs for “yes” and “no”. Evaluation is carried out on 8 representative evaluation sets out of the 14 DSTC10 evaluation sets [197]. This subset was chosen because it covers multiple domains and datasets, such as persona, topic and chitchat-based responses. A robust dialog metric should perform well across all the domains and evaluation sets considered.

The evaluation sets used for fine-grained evaluation, FED-Turn (FT) and FED-Dial (FD) [120], are included as two of the eight datasets. The other datasets include: TopicalChat-USR (TU, knowledge-grounded open-domain conversations rated for six different dialog qualities) [121]; PersonaChat-USR (PU, persona-conditioned conversations annotated with the USR schema) [121]; DailyDialog-Zhao (DZ, more formal language conversations rated for appropriateness) [204]; DailyDialog-Gupta (DGU, rated for appropriateness) [59]; DailyDialog-GRADE (DGR; annotated for coherence) [71]; and Empathetic-GRADE (EG, emotionally grounded conversations annotated for coherence) [71]. Although some of these datasets are not directly annotated for whether a response is good, the metric they use remains a component for overall quality, and thus it is treated as the indicator of the overall quality of the response in the experiments.

### In-Context Examples

This paper uses two methods for example selection: random selection, and algorithmic selection using BM25 [115] which calculates document similarity. The examples remain consistent for each evaluation test point. The random selection experiment is run three times, and the mean and standard deviation of the runs are reported. There are three configurations for BM25 between the test point and each possible example point - comparing the context only (BM25<sub>C</sub>), the response only (BM25<sub>R</sub>), and the concatenated context and response together (BM25<sub>C+R</sub>).

With the FED dataset, an additional method, manual selection, is added for example selection. For each fine-grained dialog quality, a set of three dialogs which span a wide range of ratings is chosen that remains constant over every test point. In theory, the model should be able to show increased performance if it sees examples of very good, good and bad responses for fine-grained metrics. For the DSTC10 datasets, an additional experiment tested how the number of examples used affects model performance.

### 3.2.4 Experiments and Results

The in-context example experiments are carried out on the largest available model, 530B TNLGv2, to explore the ceiling of model performance on the dialog evaluation task. 6.7B TNLGv2 is used for a direct comparison of how much performance gain is provided by using more parameters.

BLOOM and OPT are examined up to 7B and 30B respectively for the fine-grained metric evaluation task.<sup>2</sup> Smaller LLMs do not perform as well with in-context examples

<sup>2</sup>Due to limitations in compute power, larger BLOOM and OPT models were not explored. However, as the largest available GPT-3 model is explored, the comparisons appear sufficient to show the performance of a variety of LLMs.

unless they have been specifically tuned for the task, so only the 7B and 6.7B models for BLOOM and OPT respectively are explored for the DSTC10 datasets. Flan-T5 and InstructDial are analyzed in the 3B setting for consistency. Lastly, InstructGPT (text-davinci-002) is used, which has 175B parameters.

### Fine-grained Metric Evaluation

FED is separated into turn-level and dialog-level metrics. The dataset has annotations for 8 different turn-level metrics, consisting of *interestingness*, *engagingness*, *specificity*, *relevance*, *correctness*, *semantic appropriateness*, *understandability*, and *fluency*, with the addition of *overall quality*. FED annotates three different responses for each dialog context; one FED dialog is treated as one example. The corresponding rating is inserted after the response statement in the prompt, an example of which can be seen in Appendix .1.1. FED also looks at 10 different dialog-level metrics for a system’s responses: *coherence*, *error recovery*, *consistency*, *diversity*, *topic depth*, *likeability*, *understandingness*, *flexibility*, *informativeness*, and *inquisitiveness*, with *overall quality* included. The model is prompted with the full dialog context with the rating.

The FED metric was previously evaluated with both fine-tuned (ft) and from-scratch 345M and 762M DialoGPT [201] models. In the following experiments on FED, 3 in-context examples were used for prompting in Tables 3.2, 3.3, 3.4, 3.5, 3.6, and 3.7.

### In-Context Example Selection

This setting evaluates 2 versions of the TNLGv2 model: 6.7B and 530B. These models are compared to the 762M ft DialoGPT model and the results are shown in Tables 3.2 and 3.3.

First, the performances of these models are compared over the three example selection methods: manual, random, and algorithmic. With manually chosen in-context examples, the 530B TNLGv2 model outperforms the DialoGPT model on almost all turn-level metrics except for *understandability* and *fluency*. There are significant gains in all of the dialog-level metrics as well. Since DialoGPT is fine-tuned on Reddit threads, more casual language is expected, compared to models like TNLGv2 where many of the training datasets consist of more formal language. Since the wording of conversational responses tends to be more casual, it is not surprising that the fine-tuned DialoGPT model outperforms even the largest TNLGv2 model for *fluency* and *understandability*. However, the TNLGv2 models show large improvement on predicting *turn-* and *dialog-level quality*. This suggests that the TNLGv2 models have a strong grasp on overall quality, which may be due to training on more formal language.

BM25<sub>C+R</sub> generally outperforms BM25<sub>C</sub> and BM25<sub>R</sub>. However, when choosing examples with BM25<sub>C+R</sub>, the correlation of *understandability* with human annotations increases significantly when using the 6.7B TNLGv2 model. 6.7B TNLGv2 consistently outperforms 530B TNLGv2 in this aspect with any BM25 method. It appears that the smaller model is more influenced by the similarity of language in the examples than the larger one.

Even when given random examples, the TNLGv2 models outperform the 762M ft DialoGPT model on a majority of the fine-grained metrics. This shows that larger models can better detect what constitutes a good response based on these metrics even if they are not given hand-picked examples. However, they generally do not outperform the manually or algorithmically chosen examples as expected.

An additional observation is that there are certain factors that cause models to perform better or worse on specific metrics: number of parameters the model has, the type of training data, and the difficulty of the task. LLMs are able to provide increases in performance of over 50% for 15 out of 20 turn- and dialog-level metrics compared to DialoGPT with 530B TNLGv2 and manually-chosen examples. However, if the 530B TNLGv2 model is compared to the 6.7B TNLGv2 model, this increase is only observed for 2 out of the 20 metrics: *correctness* and *understandability*. LLMs can achieve high correlations with human judgement, but there is a limit to how much more performance gains can increase with extremely large models.

*Specificity*, *relevance*, and *correctness* all relate to the context of the conversation while the other metrics are more turn-specific. It follows that *relevance* and *correctness* with BM25<sub>C+R</sub> on the 6.7B TNLGv2 model outperform the 530B TNLGv2 model with manual examples. However, *specificity* performs worse. Choosing both diverse ratings and similar example points are important. This finding further supports the idea that the nature of the data used to train these LLMs is important. Had the training data been more similar to conversational language, an increase could have been observed in the correlations for these metrics without choosing algorithmically similar examples.

TNLGv2 struggles with *understandability*; it performs the worst at the highest correlation of 0.193. It also has unstable performance; performing at significance with random examples and with algorithmically chosen examples on 6.7B, but not with manually chosen ones. This shows that choosing examples with diverse ratings helps a model less for metrics that it already performs poorly on; it would better benefit from examples that are similar.

In general, even with the difference in training data, it is easier to obtain an overall sense of the conversation than a metric for a single turn for the larger models due to the large amount of parameters and variety of data that they have seen. When choosing examples based on context, the larger models generally perform worse; it appears that having different examples is more important for dialog-level metrics than for turn-level metrics.

## Comparisons Across LLMs

These model comparisons are performed using manually chosen in-context examples, since that is what generally performed the best in both turn-level and dialog-level metrics in Tables 3.4 and 3.5. Comparisons across smaller versions of BLOOM and OPT can be found in Tables 3.6 and 3.7.

Even though the large versions of BLOOM and OPT could not be run, it is apparent that both of these models outperform TNLGv2 on *understandability*, and that OPT 6.7B can outperform TNLGv2 530B on *fluency*. Data dissimilarities were noted above in Section 3.2.4 between the TNLGv2 model and the FED data. Although BLOOM was only trained

Quality	762M ft	manual		random		BM25 <sub>C+R</sub>		BM25 <sub>C</sub>		BM25 <sub>R</sub>	
		6.7B	530B	6.7B	530B	6.7B	530B	6.7B	530B	6.7B	530B
Interesting	0.408	0.455	<b>0.474</b>	0.293 ± 0.03	0.398 ± 0.02	0.358	0.383	0.336	0.389	0.355	0.385
Engaging	0.318	0.459	<b>0.484</b>	0.235 ± 0.04	0.352 ± 0.02	0.378	0.383	0.308	0.332	0.328	0.389
Specific	0.267	0.305	<b>0.450</b>	0.188 ± 0.02	0.289 ± 0.01	0.268	0.322	0.217	0.224	0.297	0.329
Relevant	0.152	0.214	0.300	0.179 ± 0.04	0.299 ± 0.03	<b>0.392</b>	0.357	0.338	0.314	0.311	0.356
Correct	0.133	0.195	0.393	0.171 ± 0.04	0.338 ± 0.04	<b>0.399</b>	0.377	0.333	0.341	0.300	0.383
Sem. Approp.	0.155	0.292	<b>0.395</b>	0.163 ± 0.03	0.270 ± 0.01	0.291	0.294	0.261	0.270	0.287	0.337
Understandable	0.111	0.021*	0.036*	0.146 ± 0.02	0.129 ± 0.02	<b>0.193</b>	0.062*	0.141	0.028*	0.169	0.029*
Fluent	<b>0.224</b>	0.164	0.195	0.052* ± 0.03	0.112* ± 0.01	0.096*	0.178	0.106	0.147	0.096*	0.121
Overall	0.209	0.371	0.475	0.256 ± 0.02	0.380 ± 0.01	0.474	<b>0.514</b>	0.435	0.438	0.360	0.407

Table 3.2: Turn-level fine-grained metrics on the FED dataset for manually, randomly, and BM25 chosen examples over the TNLGv2 6.7B and 530B models. BM25<sub>C+R</sub> stands for examples chosen by BM25 considering both the context and the response of the test point. BM25<sub>C</sub> stands for examples chosen by BM25 considering the context and BM25<sub>R</sub> stands for examples chosen by BM25 considering the response. **Bold** values indicate the best value for the metric and \* values indicate correlations that are not statistically significant.

Quality	762M ft	manual		random		BM25 <sub>C</sub>	
		6.7B	530B	6.7B	530B	6.7B	530B
Coherent	0.251	0.599	<b>0.727</b>	0.443 ± 0.03	0.533 ± 0.02	0.618	0.512
Error Recovery	0.165*	0.474	<b>0.578</b>	0.348 ± 0.04	0.463 ± 0.06	0.492	0.419
Consistent	0.116*	0.276	<b>0.382</b>	0.270 ± 0.02	0.205* ± 0.04	0.238	0.046*
Diverse	0.420	<b>0.625</b>	0.620	0.434 ± 0.06	0.490 ± 0.02	0.496	0.548
Topic Depth	0.476	0.640	<b>0.659</b>	0.361 ± 0.03	0.531 ± 0.04	0.559	0.472
Likeable	0.262	0.619	<b>0.686</b>	0.511 ± 0.03	0.580 ± 0.01	0.568	0.515
Understanding	0.306	0.517	<b>0.638</b>	0.479 ± 0.06	0.496 ± 0.02	0.567	0.428
Flexible	0.293	0.617	<b>0.656</b>	0.491 ± 0.05	0.553 ± 0.03	0.614	0.451
Informative	0.288	<b>0.569</b>	0.547	0.391 ± 0.04	0.452 ± 0.04	0.523	0.419
Inquisitive	0.163	<b>0.537</b>	0.527	0.436 ± 0.05	0.444 ± 0.02	0.334	0.252
Overall	0.443	0.630	<b>0.688</b>	0.479 ± 0.05	0.570 ± 0.02	0.607	0.531

Table 3.3: Dialog-level fine-grained metrics on the FED dataset for manually, randomly, and BM25 chosen examples over the TNLGv2 6.7B and 530B models. BM25<sub>C</sub> stands for examples chosen by BM25 considering only the context of the test point.

on some English data, it has still seen some casual language, while OPT was partially trained on Reddit data. Thus the language appearing in the BLOOM and OPT training sets more closely matches that of the conversations used here. This explains the increase in performance.

BLOOM 7B outperforms 6.7B TNLGv2 on *correctness*, while OPT 6.7B outperforms 6.7B TNLGv2 on *relevance*, *correctness*, *semantic appropriateness* and *fluency* in addition. As previously noted, *relevance* and *correctness* are turn-level metrics that take more of the context into account, so with training data that is more similar to casual language, these models perform better. It should be noted that the *overall turn-* and *dialog-level quality* results were not surpassed by any smaller model, thus the very large models will have an advantage for overall metrics.

Flan-T5 outperforms the largest model, TNLGv2 530B, on *interestingness*, *relevance*, and *understandability* at turn level and *coherence*, *consistency*, and *informativeness* at

	TNLG		BLOOM	OPT		Flan-T5	InstructGPT
	6.7B	530B	7B	6.7B	30B	3B	175B
Quality	0.455	0.474	0.291	0.429	0.399	0.519	<b>0.551</b>
Interesting	0.459	0.484	0.435	0.446	0.349	0.425	<b>0.489</b>
Engaging	0.305	<b>0.450</b>	0.296	0.275	0.207	0.433	0.421
Specific	0.214	0.300	0.109	0.272	0.289	0.435	<b>0.471</b>
Relevant	0.195	<b>0.393</b>	0.235	0.342	0.354	0.378	0.376
Correct	0.292	<b>0.395</b>	0.258	0.371	0.382	0.277	0.374
Sem. Approp.	0.021*	0.036*	0.159	0.131	0.073*	0.297	<b>0.382</b>
Understandable	0.164	0.195	0.111	0.201	0.188	0.200	<b>0.204</b>
Fluent	0.371	0.475	0.274	0.368	0.433	0.445	<b>0.536</b>
Overall							

Table 3.4: Turn-level fine-grained metrics on the FED dataset for manually chosen examples over the TNLGv2, BLOOM, OPT, Flan-T5, and InstructGPT models.

	TNLG		BLOOM	OPT		FLAN-T5	InstructGPT
	6.7B	530B	7B	6.7B	30B	3B	175B
Quality	0.599	0.727	0.613	0.558	0.584	<b>0.730</b>	0.707
Coherent	0.474	<b>0.578</b>	0.474	0.377	0.479	0.398	0.560
Error Recovery	0.276	0.382	0.323	0.237	0.309	0.410	<b>0.517</b>
Consistent	0.625	0.620	0.498	0.454	0.607	0.544	<b>0.628</b>
Diverse	0.640	0.659	0.637	0.544	0.609	0.650	<b>0.680</b>
Topic Depth	0.619	<b>0.686</b>	0.566	0.544	0.571	0.659	0.672
Likeable	0.517	0.638	0.484	0.505	0.483	0.637	<b>0.694</b>
Understanding	0.617	0.656	0.499	0.528	0.592	0.595	<b>0.688</b>
Flexible	0.569	0.547	0.462	0.497	0.522	<b>0.662</b>	0.647
Informative	0.537	0.527	0.539	0.461	0.537	0.487	<b>0.578</b>
Inquisitive	0.630	0.688	0.531	0.374	0.530	0.585	<b>0.690</b>
Overall							

Table 3.5: Dialog-level fine-grained metrics on the FED dataset for manually chosen examples over the TNLGv2, BLOOM, OPT, Flan-T5, and InstructGPT models.

	BLOOM				OPT			
	560M	1.1B	1.7B	3B	125M	350M	1.3B	2.7B
Quality	0.282	0.331	0.336	0.328	0.187	0.186	0.388	0.245
Interesting	0.217	0.320	0.278	0.418	0.121	0.252	0.398	0.292
Engaging	0.030*	0.065*	0.204	0.353	0.197	0.004*	0.217	0.222
Specific	0.076*	0.056*	0.072*	0.091*	0.146	0.105	0.231	0.177
Relevant	0.106	0.146	0.124	0.173	0.119	0.152	0.327	0.270
Correct	0.048*	0.228	0.205	0.265	0.148	0.278	0.274	0.296
Sem. Approp.	-0.017*	0.043*	-0.005*	0.087*	0.058*	0.021*	0.189	0.205
Understandable	0.158	<b>0.223</b>	0.097*	0.091*	0.109	0.087*	0.158	0.163
Fluent	0.086*	0.179	0.076*	0.285	0.134	0.219	0.338	0.197
Overall								

Table 3.6: Turn-level fine-grained metrics on the FED dataset for manually chosen examples over the smaller sizes of BLOOM and OPT.

dialog level. There is a larger performance drop for the *semantic appropriateness*, *error recovery*, and *overall dialog-level quality* metrics. *Error recovery* is a relatively new metric [120]. Even though Flan-T5 was fine-tuned on many dialog tasks, it may not have seen data that addresses this specific metric. Flan-T5 only has 3B parameters, and the fact that it outperforms 530B TNLGv2 shows the importance of use of dialog data during pre-training or fine-tuning.

Quality	BLOOM				OPT			
	560M	1.1B	1.7B	3B	125M	350M	1.3B	2.7B
Coherent	0.499	0.533	0.531	0.531	0.490	0.514	0.528	0.435
Error Recovery	0.293	0.298	0.322	0.448	0.168	0.380	0.342	0.348
Consistent	0.217	0.238	0.129*	0.264	0.193	0.191	0.250	0.268
Diverse	0.345	0.430	0.461	0.518	0.451	0.304	0.491	0.531
Topic Depth	0.418	0.414	0.519	0.462	0.228	0.302	0.462	0.454
Likeable	0.310	0.374	0.421	0.476	0.467	0.395	0.462	0.535
Understanding	0.276	0.312	0.257	0.371	0.389	0.283	0.414	0.494
Flexible	0.269	0.432	0.400	0.441	0.458	0.377	0.460	0.432
Informative	0.149*	0.384	0.372	0.537	0.378	0.402	0.381	0.544
Inquisitive	0.198	0.350	0.318	0.339	0.489	0.300	0.439	0.413
Overall	0.262	0.146*	0.207	0.261	-0.000*	0.319	0.452	0.437

Table 3.7: Dialog-level fine-grained metrics on the FED dataset for manually chosen examples over the smaller sizes of BLOOM and OPT.

InstructGPT, being fine-tuned with prompting at 175B parameters, is more suitable for the present experiments. It performs very well on both turn- and dialog-level metrics, outperforming 530B TNLGv2 on almost all metrics. Since InstructGPT has already seen prompting, the model can better understand a task through only instructions or combinations of instructions and in-context examples.

## DSTC10 Datasets

The same set of experiments were carried out on the 8 datasets in the DSTC10 challenge in Tables 3.8, 3.9 and 3.10. The previous best performing metrics on DSTC10 are compiled from [61], which include both reference-free and fine-tuned metrics, of which over 6 of the 8 datasets, the best performing metric was based on a model fine-tuned on some subset of DSTC10 datasets. Quality is evaluated in terms of how good a response is to the context.

## In-Context Example Selection

Experiments are performed with randomly chosen examples and examples that were chosen by BM25 over 4, 8, and 12 examples in Tables 3.8 and 3.9. Higher correlation results are obtained on 4 datasets (DZ, DGU, DGR, and FT) with comparable results on 3 datasets (TU, PU, and FD), as compared to the best DSTC10 baselines. Most of the best results are on the 530B TNLGv2 model, which will be discussed in this section, as compared to the 6.7B TNLGv2 model. Several factors are relevant here: the language of the dataset, the way the dataset was created, and how the dataset was annotated.

DailyDialog contains more formal language, thus TNLGv2 should perform well since its training dataset includes data sources with formal language. DZ, DGU, and DGR almost always perform the best when examples are chosen from looking at the context; adding the response generally leads to poorer performance. Since these datasets are annotated for *appropriateness* and *coherence*, context is more important than a more turn-specific metric.

TopicalChat was created through knowledge-grounding. The conversations could thus have more substance than a purely open-domain un-prompted conversation. It thus follows

Model	TU	DZ	PU	DGU	DGR	FT	EG	FD
<i>Experiments with Random Examples</i>								
4ex	0.034* $\pm$ 0.05	0.117 $\pm$ 0.02	0.206 $\pm$ 0.02	0.080* $\pm$ 0.05	0.121 $\pm$ 0.05	0.191 $\pm$ 0.06	0.005* $\pm$ 0.04	0.228 $\pm$ 0.03
8ex	0.054* $\pm$ 0.05	0.160 $\pm$ 0.02	0.206 $\pm$ 0.03	0.109* $\pm$ 0.03	0.139 $\pm$ 0.08	0.178 $\pm$ 0.02	0.060* $\pm$ 0.06	0.238 $\pm$ 0.11
12ex	0.063* $\pm$ 0.03	0.149 $\pm$ 0.00	0.225 $\pm$ 0.01	0.114 $\pm$ 0.05	0.143 $\pm$ 0.06	0.210 $\pm$ 0.03	0.052* $\pm$ 0.02	0.127 $\pm$ 0.04
<i>Experiments with Algorithmically Retrieved Examples</i>								
4ex BM25 <sub>R</sub>	0.148	0.218	0.223	0.202	0.094*	0.273	-0.012*	0.335
4ex BM25 <sub>C</sub>	0.124	0.198	0.237	0.209	0.214	0.296	0.057*	0.314
4ex BM25 <sub>C+R</sub>	0.05*	0.142	0.169	0.167	0.083*	0.274	0.038*	0.339
8ex BM25 <sub>R</sub>	0.077*	0.270	0.203	0.222	0.128	0.199	0.042*	0.335
8ex BM25 <sub>C</sub>	0.184	0.328	0.343	0.526	0.176	0.363	0.073*	0.387
8ex BM25 <sub>C+R</sub>	0.029*	0.152	0.020*	0.092	0.022*	0.348	0.024*	0.440
12ex BM25 <sub>R</sub>	0.069*	0.338	0.153	0.213	0.110*	0.250	0.026*	0.401
12ex BM25 <sub>C</sub>	0.285	0.544	0.325	0.678	0.208	0.330	0.042*	0.365
12ex BM25 <sub>C+R</sub>	0.035*	0.168	0.088*	0.086*	0.100*	0.407	0.092*	0.343

Table 3.8: Spearman correlation of model predictions with human ratings for TNLGv2 6.7B model with algorithmically chosen examples. TU, PU, PZ, DZ, CG, DGU, DGR, EG, FT and FD are abbreviations for TopicalChat-USR, PersonaChat-USR [121], PersonaChat-Zhao [204], DailyDialog-Zhao [204], ConvAI2-GRADE [71], DailyDialog-Gupta [59], DailyDialog-GRADE [71], Empathetic-GRADE [71], FED-Turn and FED-Dial [120].

that response selection will work the best when choosing examples. PersonaChat has conversations that are persona-conditioned, so the quality of the conversation should take into account the entire conversation for each persona. It performs better with examples chosen for context and response or with just context.

FED is split into turn- and dialog-level annotations, thus, for turn-level annotations choosing examples based on responses should work best, and for dialog-level annotations choosing examples based on either the context or the context and response should perform the best. Choosing examples with context and response performs the best for EG, but randomly choosing examples outperforms that result. It may be that with emotionally grounded conversations, the model needs more, or more diverse examples due to the different ways emotion can be expressed.

In general, choosing examples algorithmically improves performance over randomly choosing examples. This is consistent with previous experiments above. However, randomly-chosen examples perform better on the DGR and EG datasets on the 530B TNLGv2 model. This may be because these two datasets were rated for *coherence*. Algorithmically, choosing examples based on context and response performs the best on EG, as was seen for coherence in FED in Section 3.2.4.

## Comparisons Across LLMs

Table 3.10 compares the evaluation results across various LLMs. Due to model input length restrictions, the following experiments were carried out using 4 in-context examples or in a zero-shot setting. BM25 is only used with the context as the example selection strategy, since it performed well with the TNLGv2 models.

In the few-shot setting, models that were not fine-tuned or trained with prompting (BLOOM, OPT) did not have consistent results across the datasets. However, those that

Model	TU	DZ	PU	DGU	DGR	FT	EG	FD
<i>Experiments with Random Examples</i>								
4ex	0.112 ± 0.03	0.428 ± 0.01	0.403 ± 0.02	0.542 ± 0.00	0.338 ± 0.01	0.318 ± 0.02	0.248 ± 0.04	0.290 ± 0.05
8ex	0.169 ± 0.03	0.430 ± 0.03	0.331 ± 0.03	0.570 ± 0.01	<b>0.429</b> ± 0.05	0.337 ± 0.01	0.200 ± 0.04	0.339 ± 0.18
12ex	0.148 ± 0.03	0.453 ± 0.02	0.384 ± 0.02	0.565 ± 0.01	0.410 ± 0.06	0.412 ± 0.03	0.160 ± 0.02	0.351 ± 0.08
<i>Experiments with Algorithmically Retrieved Examples</i>								
4ex BM25 <sub>R</sub>	0.247	0.424	0.252	0.482	0.342	0.364	0.144	0.264
4ex BM25 <sub>C</sub>	0.129	0.424	0.339	0.510	0.370	0.172	0.192	0.549
4ex BM25 <sub>C+R</sub>	0.213	0.441	0.432	0.479	0.371	0.137	0.211	0.479
8ex BM25 <sub>R</sub>	0.309	0.487	0.275	0.536	0.304	<b>0.426</b>	0.121	0.419
8ex BM25 <sub>C</sub>	0.227	0.564	0.460	0.627	0.387	0.323	0.123	0.518
8ex BM25 <sub>C+R</sub>	0.185	0.458	0.439	0.526	0.308	0.377	0.171	0.530
12ex BM25 <sub>R</sub>	0.300	0.474	0.358	0.570	0.337	0.393	0.095*	0.414
12ex BM25 <sub>C</sub>	0.278	<b>0.688</b>	0.449	<b>0.674</b>	0.397	0.377	0.106*	0.492
12ex BM25 <sub>C+R</sub>	0.202	0.491	0.452	0.465	0.349	0.358	0.148	0.493
Best of DSTC10 baselines	<b>0.319</b>	0.532	<b>0.493</b>	0.596	0.363	0.247	<b>0.395</b>	<b>0.555</b>

Table 3.9: Spearman correlation of model predictions for overall quality with human ratings for TNLGv2 530B model with algorithmically chosen examples. TU, PU, PZ, DZ, CG, DGU, DGR, EG, FT and FD are abbreviations for TopicalChat-USR, PersonaChat-USR [121], PersonaChat-Zhao [204], DailyDialog-Zhao [204], ConvAI2-GRADE [71], DailyDialog-Gupta [59], DailyDialog-GRADE [71], Empathetic-GRADE [71], FED-Turn and FED-Dial [120].

Model	TU	DZ	PU	DGU	DGR	FT	EG	FD
<i>Few-shot in-context Experiments</i>								
BLOOM-7B-4ex	0.027*	0.075	0.123	0.127	0.131	0.117	0.012	0.289
OPT-6.7B-4ex	0.115	0.258	0.444	0.228	0.091*	0.486	0.044*	<b>0.657</b>
TNLG-6.7B-4ex	0.124	0.198	0.237	0.209	0.214	0.296	0.057*	0.314
TNLG-530B-4ex	0.129	0.424	0.339	0.510	0.370	0.172	0.192	0.549
Flan-T5-3B-4ex	0.447	0.657	0.578	0.714	0.379	0.442	0.396	0.492
InstructGPT-175B-4ex	<b>0.616</b>	<b>0.716</b>	<b>0.687</b>	<b>0.746</b>	<b>0.472</b>	<b>0.506</b>	0.305	0.412
<i>Zero-shot Experiments</i>								
Flan-T5-3B-0ex	0.357	0.599	0.533	0.677	0.351	0.380	0.418	0.444
InstructDial-3B-0ex	0.446	0.601	0.376	0.634	0.286	0.263	<b>0.475</b>	0.228
Best of DSTC10 baselines	0.319	0.532	0.493	0.596	0.363	0.247	0.395	0.555
Best TNLGv2 value	0.309	0.688	0.460	0.678	0.429	0.426	0.248	0.549

Table 3.10: Spearman correlation of model predictions for overall quality with human ratings with 4 examples chosen with BM25 using context. Macro average scores are also shown.

were fine-tuned or prompted (Flan-T5, InstructGPT, InstructDial) had results that were close to or surpassed the previous best DSTC10 baselines. InstructGPT performed the best. Even in the zero-shot setting, Flan-T5 outperforms the baseline in 6 of the datasets, and InstructDial in 5.

These results clearly show that for dialog evaluation, it is insufficient to simply train on large amounts of general internet data. Specialized approaches such as instruction tuning on multiple tasks improve the generalization capabilities of models in zero- and few-shot settings. It is not surprising that InstructGPT performs the best since it fine-tunes a very large language model with instructions.

### 3.2.5 Conclusion

LLMs have the potential to significantly contribute to dialog evaluation. Current LLMs perform well for this task in a few-shot setting. However, this performance varies greatly depending on the content of and number of examples in the prompt. Models prefer more similar examples for metrics that they struggle to evaluate, while preferring examples with more diverse ratings for metrics that they can evaluate well. Very large language models also still afford performance gains, especially for overall quality evaluation at the turn and dialog level. Even though large language models perform better at dialog-level fine-grained metrics, there are still previously shown issues with how these models understand social situations and use context that may hinder further improvement if not addressed.

Performance is also affected by the model’s training data. Smaller language models that are fine-tuned on instructions, trained on dialog data, and/or trained on multiple dialog tasks outperform larger language models. These smaller models also perform more consistently over different domains. This indicates that LLMs should have more diverse pre-training data in order to be able to handle a larger variety of tasks in few or zero-shot settings.

More work needs to be done on understanding how a large language model models different types of tasks. In-context example selection and example wording still remains unstable across large language models in many tasks, and the performance variation over different dialog domains in this paper demonstrates that as well.

Presently, the LLMs explored in this paper have their own strengths. Smaller models such as BLOOM and OPT could share more training data similarity with dialog tasks based on their objective. TNLGv2 530B provides a very large language model that has shown improvement in dialog evaluation along with other NLP tasks. Flan-T5 and InstructDial show the efficacy of fine-tuning a LLM on dialog tasks, and InstructGPT shows the importance of training a model to better recognize prompts. The evaluations of these models provide suggestions for the characteristics of the best LLMs to use for dialog evaluation. Future work in using LLMs for other NLP tasks can benefit from such comprehensive analyses. Once a better understanding of LLMs is realized, the capabilities of large language models for zero- and few-shot tasks will increase greatly.

## 3.3 Quantifying the Statistical Effect of Rubric Modifications on Human-Autorater Agreement

### 3.4 Introduction

Autoraters, or LLM-as-judges, have been used as an alternative to human annotation due to their scalability, cost, and time effectiveness. Their effectiveness is typically validated through agreement with human annotation. Borrowing from education literature, a *rubric* is defined as having “*coherent sets of criteria*” and “*descriptions of levels of performance for these criteria*” [17]. This describes the scoring guidelines and instructions provided to any rater, whether human or automated, which is also referred to as part of a prompt

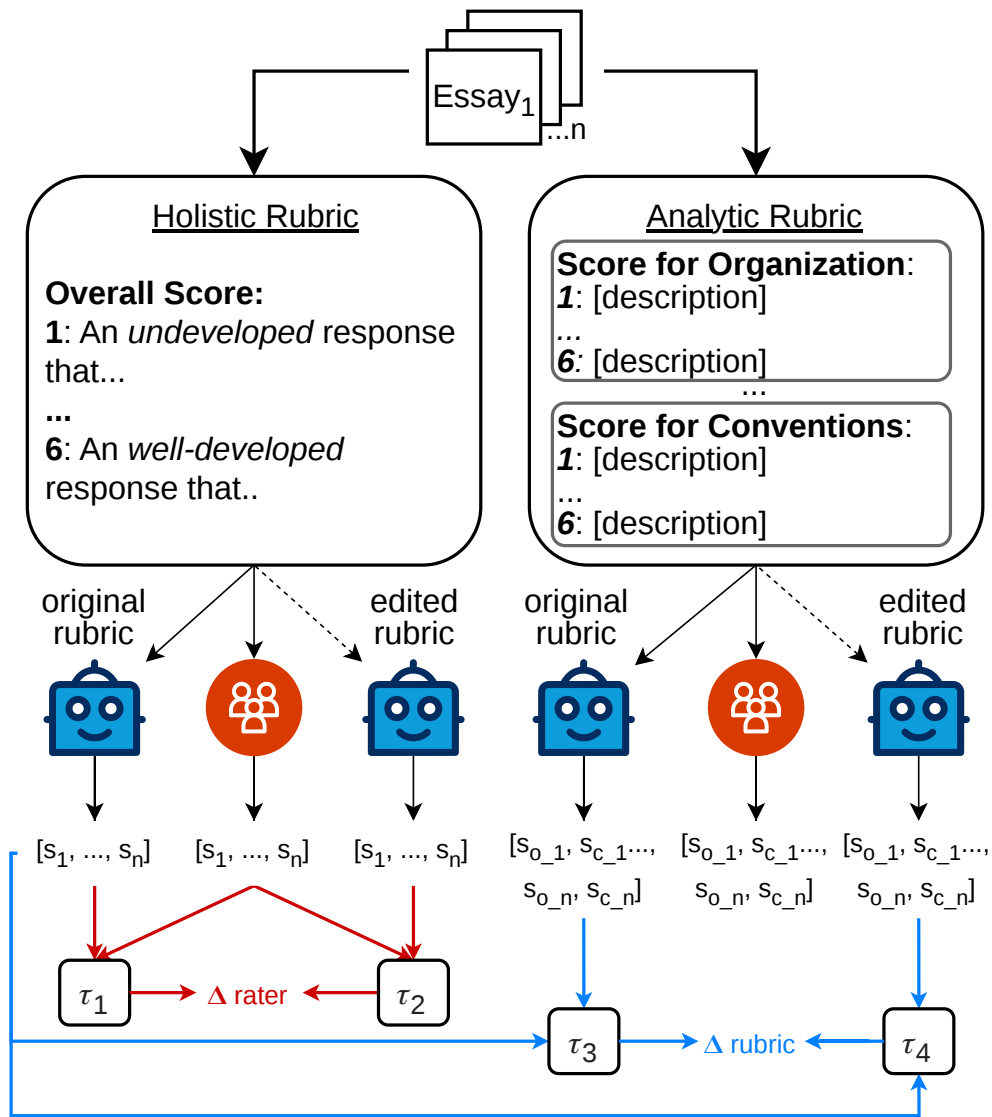


Figure 3.2: This diagram provides a walkthrough of the experimental setup as shown for automatic essay scoring. It represents comparisons made between human-atorater agreements  $\tau$  across holistic rubrics (left), in which all criteria are applied together in a single overall judgment, or analytic rubrics (right), in which criteria are evaluated separately, resulting in multiple scores. The original rubrics are given to a human and an autorater, while edited rubrics are given to autoraters only. Arrows in bold between  $\tau_1$ ,  $\tau_2$  and  $\tau_3$ ,  $\tau_4$  represent comparisons for which statistical significance can be calculated.  $\Delta rater$  represents comparisons where the type of rater is changed while the type of rubric remains constant, while  $\Delta rubric$  represents comparisons where the type of rubric is changed and the type of rater remains constant.

as described in autorater literature. Ideally, both human raters and autoraters would receive equivalent evaluation rubrics that accurately measure the same construct with reliable certainty. However, equivalence does not necessitate identical presentation. Wu and Quinn [191] show that expert and non-expert human raters may require different levels of instruction specificity, for instance, specifying tools and providing concrete examples improves accuracy specifically when raters lack task-relevant knowledge.

In addition to identical presentation across different types of raters, humans and autoraters are sensitive to variations in instruction presentations within the same type of rater. For humans, this sensitivity includes the interpretation of the task during crowdsourcing [86] and instruction specification, where increasing instruction specification increases accuracy on the task [191]. For autoraters, position bias (where the position of the evaluated text within the prompt will influence the autorater evaluation) and verbosity bias (where autoraters prefer more verbose texts) are only a few of the known sensitivities [207]. Autoraters are also sensitive to rubric variations such as formatting choices [158] and example ordering [109]. Another example of instruction presentation involves decomposing criteria into multiple sub-criteria. Previous work has successfully used decomposition to improve LLM performance, whether by having autoraters decompose evaluation criteria into sub-tasks [153] or by having humans decompose complex questions into simpler sub-questions for models [136].

Understanding whether rubric modifications produce statistically significant shifts in agreement is essential for practitioners who aim to deploy autoraters as evaluation tools. This work studies human-autorater agreement in two domains: automatic essay scoring (AES) and instruction-following (IF). We examine statistically how rubric presentation and broader rubric modifications affect human-autorater agreement on subjective evaluation tasks, as well as empirically examining whether decomposing general holistic judgments that ask for a single high-level criteria into more granular sub-criteria, originally designed for human evaluators, can be an effective approach for improving human-judge agreement. The findings indicate that instructions optimized for autoraters tend to improve agreement with human ratings when autoraters receive machine-optimized instructions and humans receive the original set of instructions. Conversely, giving autoraters simpler prompts does not guarantee higher agreement with human ratings. These results indicate that 1) rubric edits providing representative examples along with contextual information increased human-autorater agreement as well as autorater self-agreement, 2) higher criterion complexity and conservative aggregation methods tended to decrease human-autorater agreement, 3) reducing confirmation bias tends to significantly increase human-autorater agreement and 4) high human inter-rater agreement leads to significantly higher human-autorater agreement. These findings from the automatic essay scoring and instruction-following evaluation domains suggest that practitioners should carefully analyze domain-specific performance and modify rubrics to move towards high human-autorater agreement.

## 3.5 Related Work

### 3.5.1 Rubrics in Context

Expanding upon Section 3.4, rubrics consist of both the criteria, which are the components of the overall evaluation, and the descriptions for the criteria.

**Decomposition level** describes criteria presentation, and refers to whether prompts are *holistic*, where “all criteria are [applied] at the same time”, or *analytic*, where “work [is described] on each criterion separately”, [17].

**Generality level** details the descriptions for the criteria, and refers to whether prompts are *general* or *task-specific* (i.e., evaluation prompts that [can\cannot] also be used for other tasks) [17]. Autorater evaluation prompts can be viewed through this lens - for example, a holistic autorater evaluation prompt may ask for a single overall judgment whereas an analytic prompt would decompose the evaluation into criterion to be evaluated separately.

**Prompt complexity**, as used in this work, refers to the cognitive demands placed on a rater during evaluation. Prior work on task complexity and cognitive demand has identified the number of *paths*, or components, that must be considered simultaneously [22], the degree of element interactivity (i.e., the extent to which components must be processed together rather than independently) [167] and ambiguity resulting in communication failure [22] as characteristics of cognitive load. Thus in our work, we consider the number of criteria evaluated, the degree to which score-level descriptions are interrelated, and the extent to which the rater must resolve ambiguity across sub-criteria.

### 3.5.2 LLM-as-judges

The autorater (LLM-as-a-judge) paradigm has received substantial recent attention. Several past works demonstrated that LLMs can produce evaluations consistent with human experts. [30, 99]. However, research also shown that autoraters are sensitive to prompt variations, with different instructions leading to substantially different performances [123] and varying quality, necessitating statistical procedures to justify replacing human raters [20]. Several recent works have studied how prompt modifications affect autorater performance. [158] demonstrated that prompt formatting choices (e.g., separator characters, whitespace) can significantly affect task accuracy, though these effects weakly correlate across models. [109] showed that example ordering in few-shot prompts substantially impacts performance on classification tasks. Our work extends this literature by using statistical procedures where possible to test how rubric modification choices, such as decomposition level, example selection, and aggregation methods affect agreement with human judgment on tasks where even trained human raters often disagree. This work also empirically examines whether simpler rubrics, often assumed to reduce cognitive load, actually improve human-autorater agreement, in domains where there is no single correct answer, contrasting with the classification and multiple-choice tasks used in prior work.

### 3.5.3 Automatic Essay Scoring (AES) and Instruction-Following (IF)

Recent work has explored using autoraters for AES, employing various personas, including a “virtual evaluator with expertise in English composition” [192], a “helpful pattern-following assistant” [116], and an “English essay writing test evaluation committee” or “English teacher” [97]. However, these personas may be misaligned with the original human raters; for example, scoring guidelines for a portion of the dataset from Hamner et al. [64] explicitly state that raters should not be teachers. The number of few-shot examples provided to the autorater also varies, Xiao et al. [192] select the closest three examples with calculated embeddings, while Kundu and Barbosa [93] chose one essay that scored highly and one essay that scored poorly.

While most studies utilize the rubrics given by the AES datasets, Lee et al. [97] automatically decomposed the original essay scoring rubric into sub-criteria and performed a modified average aggregation on the sub-criteria, which on average performs better than using a single score from zero-shot prompting. However, Xiao et al. [192] demonstrated significant improvements by fine-tuning GPT-3.5-turbo and Llama3-8B compared to using GPT-4 with rubrics and few-shot examples. These previous works also may request explanations for scores from autoraters to mimic CoT, aiming for more accurate explanations and ratings. Studies on analytic rubrics primarily focus on cross-prompt scoring [28]. The experimental setup in this work follows the rigor of prior work by studying multiple autoraters, using given rubrics, and performing various rubric edits. However, the hypotheses examined extend prior findings on edited rubrics and further investigate rubric components across various essay scoring rubrics.

Research on instruction-following in LLMs has led to two primary approaches for improving and evaluating LLMs, instruction tuning and alignment tuning. Prior work has found that larger models tend to follow instructions more accurately, though this relationship is not strictly linear [130]. Honovich et al. [68] proposed allowing LLMs to write instructions based on only seeing examples of a task, although this is still less accurate than human-written instructions.

## 3.6 Experiment Setup

Our hypotheses focus on the criteria presentation of rubrics for autoraters: (1) **Edited** prompts will improve autorater alignment with human ratings over the original prompts (humans and autoraters may require different prompts due to the way that they process information); and (2) **Analytic** rubrics will improve autorater alignment with human ratings over **holistic** rubrics (the decomposition of holistic constructs into discrete components within analytic rubrics is expected to improve autorater alignment by simplifying the evaluation task).

### 3.6.1 Experiments

This work examines four different scores: human ratings on holistic ( $H_H$ ) and analytic prompts ( $H_A$ ), and autorater ratings on holistic ( $LLM_H$ ) and analytic prompts ( $LLM_A$ ). These scores facilitate two comparisons:  $\Delta\mathbf{Rater}$ , which assesses human-autorater agreement when using the same type of rubric,  $\Delta\mathbf{Rubric}$ , which investigates the impact of varying the type rubric while keeping the type of rater constant, revealing how different rubrics influence ratings even when measuring the same criteria. These comparisons are depicted in Figure 3.2. Additional analysis on  $\Delta\mathbf{Rater}+\mathbf{Rubric}$ , varying both the type of rubric and type of rater, is explored in Appendix .3.1 for completion.

Additionally, this study compares two approaches: presenting the autorater with the original human rubric versus a modified rubric, or **edited rubric**, designed to enhance agreement. Modifications include adding additional context (if available), incorporating examples, and reducing positional bias. Prior work [116] studied adding a rubric and then adding an example into the prompt, and found that ChatGPT<sup>3</sup> benefited from having rubrics and examples, whereas Llama-2-13b-chat-hf<sup>4</sup> did not for some cases. This work uses several modification conditions. First, the original holistic prompts for the AES task include a large number of examples (10-18) - edited holistic prompts reduces the examples to a representative set of three examples (**3ex**; high, medium, and low scores). This serves two purposes: (1) it creates a more direct comparison with the **edited** analytic prompts, which also use three examples, and (2) it avoids an excessively long prompt for the model. Second, the analytic rubric prompt formats of presenting all sub-criteria in a single API call (batch), in separate API calls to mitigate potential positional bias (separate), and a combination of enhancements such as incorporating additional context, the "3ex" set (as the original analytic rubrics had no examples), and the "separate" API call strategy (edited) are tested. Prompts are provided in Appendix .4.

## 3.7 Methods and Materials

### 3.7.1 Datasets

Each hypothesis is tested across the domains of AES and IF to determine whether any observed effect is context-dependent. Human annotations are provided by each dataset.

**Automatic Essay Scoring (AES).** Automatic essay scoring is a well-established field, with the Automated Student Assessment Prize (ASAP) as a primary dataset [64]. ASAP contains over 21,000 essays over 8 essay prompts, each with a holistic rubric used by trained human raters. To provide further diversity within AES, this study uses essay prompts 1, 4, and 6, which represent diverse essay prompt types (argumentative, source-dependent, source-dependent) and rubric types (general, general, task-specific) shown in Appendix

<sup>3</sup><https://platform.openai.com/docs/models>

<sup>4</sup><https://huggingface.co/meta-llama/Llama-2-13b-chat-hf>

Table 5. The essay prompts also provide single holistic scores, and integer scores from 1 to 5 (essay prompt 6’s scores are condensed to integers).

The ASAP++ dataset [117] provides corresponding analytic rubrics for these essay prompts. Except for the holistic rubric for essay prompt 6, all other rubrics are general. It is important to note while both ASAP and ASAP++ evaluate essay quality, the specific sub-criteria for essay quality differ (ex. “audience awareness” is mentioned in ASAP but not ASAP++ and “conventions” is mentioned in ASAP++ but ASAP explicitly asks raters to not consider this). This discrepancy may lead to imperfect agreement between ASAP and combined ASAP++ ratings. Additionally, the ASAP++ analytic rubrics do not include examples, so three examples per essay prompt were selected from ASAP++ and excluded from analysis. These examples represent high, medium, and low scores across all analytic sub-criteria (ex. a high scoring essay contains the highest scores for all analytic sub-criteria).

**Instruction-following (IF).** Compared to AES, instruction-following work has decomposed instructions into a series of questions that isolate individual criteria, but with no evaluation differences in the rubric. The InfoBench [141] dataset consists of both easy and hard instructions, along with outputs from five popular LLMs. Outputs are holistically scored on a scale of one to five with a general rubric, and analytically scored with binary yes/no responses to decomposed instructions with task-specific rubrics. In the dataset, each output was annotated with both scores by three expert annotators using provided rubrics. Expert annotators annotated 50 instructions with 5 generations each. The same example selection methodology used for AES was used to select input, output, instruction tuples from the remaining dataset that were not annotated by the expert annotators, since neither rubric in IF provides examples. The decomposed automatic instruction annotations released by the InfoBench authors served as a starting point to explore the tuples. The resulting example outputs were generated from three models: GPT-3.5-turbo<sup>3</sup>, Alpaca-7b [168], and GPT-4<sup>3</sup>[129]), and represent high, medium, and low instruction-following ratios (100%, 50%, 0%). This is translated into holistic scores of 5, 3, and 1, respectively, which were validated by the authors.

We select AES and IF for complementary purposes. The AES dataset provides validity through its multiple rubric variations (examples) all measuring the same underlying construct, essay quality, allowing us to observe the different effects of different operationalizations of the same evaluation task. Although using ASAP and ASAP++ simultaneously may result in a noisy comparison, it provides preliminary evidence for the behavior. Conversely the InfoBench dataset offers a direct 1:1 comparison holistic and analytic rubrics that measure identical criteria. Together, these datasets enable us to distinguish between effects that are domain specific versus those that generalize across evaluation contexts/tasks.

### 3.7.2 Autoraters

Experiments were conducted using gpt-4o-2024-11-20<sup>3</sup>[128] as the autorater, which has demonstrated high alignment with human evaluations across various tasks [25, 209], and Llama-3.1-70B-Instruct [58].

Autorater ratings are calculated using a probability weighting scheme similar to Huynh et al. [74], given by the equation:  $r = \sum_{i=0}^n \frac{p_i}{\sum_{j=0}^n p_j} * s_i$ , where  $r$  is the final rating from the autorater,  $n$  represents the number of possible scores given by the rubric,  $p_i$  represents the exponential of the log probability score given by the autorater, and  $s_i$  represents the integer score outputted by the autorater that corresponds to the log probability. In IF for analytic prompts,  $s_i$  is represented by 1 for “yes” and 0 for “no”.

### 3.7.3 Agreement Calculation

While direct numerical comparison (e.g., correlation between raw scores) is often calculated through agreement metrics such as Cohen’s  $\kappa$ , autoraters and human have been shown to use the same scales differently, with autoraters often exhibiting compressed or shifted score distributions relative to humans [93]. Preference-based evaluation (e.g., A/B) is more robust to such shifts, as it captures ordinal relationships rather than numerical ones; however, such methods do not scale with large datasets. Addressing both issues, we use Kendall’s  $\tau$  with tie consideration (calculated using the SciPy [176] implementation of [88]), which operates on the existing numerical scores from the data while evaluating the scores through pairwise ordinal comparisons.

Each domain employs a different method for aggregating analytic rubric scores when comparing them to holistic rubric scores to calculate concordant pairs for Kendall’s  $\tau$ . In AES, Pareto dominance is used, where essay A is considered better than essay B if all sub-criteria scores of A are at least tied to those of B, with at least one sub-criteria score from A being higher. Pareto dominance is used as a conservative aggregation method to ensure that one essay would truly be better than another without knowing about how the sub-criteria were factored into a holistic rating. In IF, response A is considered to follow instructions better than response B if A has a higher ratio of correctly followed instructions.

Pareto dominance is chosen as a conservative aggregation method precisely because the sub-criteria in ASAP and ASAP++ do not perfectly overlap, as noted in 3.7.1. Rather than imposing an arbitrary weighting scheme or score cutoff to determine which essay is better, Pareto dominance avoids assuming any particular trade-off between criteria, thus providing a reliable lower bound on the present agreement.

The confidence interval around the correlations is calculated using bootstrapping [44] with 1000 samples. The difference between two conditions is calculated, and the 95% confidence interval is determined by sampling the 25th and 975th sorted values. For comparisons involving three conditions, the confidence interval is adjusted using Bonferroni correction, with the interval bounds set as the average of the 8th and 9th values and the average of the 991st and 992nd values. The compared conditions are considered significantly different if 0 does not fall within the interval.

		GPT-4o					Llama				
		$\Delta$ rater									
		$H_H \rightarrow LLM_H$		$H_A \rightarrow LLM_A$			$H_H \rightarrow LLM_H$		$H_A \rightarrow LLM_A$		
P	C.	Full	3ex	Sep.	Bat.	Edited	Full	3ex	Sep.	Bat.	Edited
1	H.	0.437 <sup>†</sup>	0.387	-	-	-	0.576	0.571	-	-	-
	Ide.	-	-	0.474 <sup>*</sup>	0.464	0.552 <sup>s,b†</sup>	-	-	0.552	0.559	0.552
	Or.	-	-	0.497 <sup>*</sup>	0.448	0.544 <sup>s,b†</sup>	-	-	0.547 <sup>*</sup>	0.524	0.529 <sup>s↓</sup>
	WC	-	-	0.452 <sup>*</sup>	0.439	0.554 <sup>s,b†</sup>	-	-	0.547 <sup>*</sup>	0.527	0.507 <sup>s↓</sup>
	SF	-	-	0.459	0.455	0.553 <sup>s,b†</sup>	-	-	0.559 <sup>*</sup>	0.512	0.523 <sup>s↓</sup>
	Cv.	-	-	0.362	0.388 <sup>*</sup>	0.472 <sup>s,b†</sup>	-	-	0.491	0.477	0.487
4	H.	0.695 <sup>†</sup>	0.687	-	-	-	0.700	0.699	-	-	-
	Ct.	-	-	0.673	0.696 <sup>*</sup>	0.698 <sup>s†</sup>	-	-	0.698 <sup>*</sup>	0.693	0.701
	PA	-	-	0.662	0.677 <sup>*</sup>	0.680 <sup>s†</sup>	-	-	0.683 <sup>*</sup>	0.665	0.682 <sup>b†</sup>
	La.	-	-	0.627 <sup>*</sup>	0.595	0.639 <sup>s,b†</sup>	-	-	0.630 <sup>*</sup>	0.573	0.621 <sup>b†</sup>
	Na.	-	-	0.669 <sup>*</sup>	0.652	0.682 <sup>s,b†</sup>	-	-	0.674 <sup>*</sup>	0.642	0.669 <sup>b†</sup>
6	H.	0.629	0.644 <sup>†</sup>	-	-	-	0.666	0.680 <sup>†</sup>	-	-	-
	Ct.	-	-	0.610	0.652 <sup>*</sup>	0.676 <sup>s,b†</sup>	-	-	0.680 <sup>*</sup>	0.648	0.694 <sup>s,b†</sup>
	PA	-	-	0.605	0.608	0.666 <sup>s,b†</sup>	-	-	0.652 <sup>*</sup>	0.601	0.668 <sup>s,b†</sup>
	La.	-	-	0.524 <sup>*</sup>	0.510	0.546 <sup>s,b†</sup>	-	-	0.542 <sup>*</sup>	0.496	0.547 <sup>b†</sup>
	Na.	-	-	0.562	0.557	0.590 <sup>s,b†</sup>	-	-	0.579 <sup>*</sup>	0.547	0.589 <sup>b†</sup>
		$\Delta$ rubric									
		$H_H \rightarrow H_A$		$LLM_H \rightarrow LLM_A$			$H_H \rightarrow H_A$		$LLM_H \rightarrow LLM_A$		
P	C.	Full	3ex	Sep.	Bat.	Edited	Full	3ex	Sep.	Bat.	Edited
1	full	0.591	-	0.756	0.750	0.839 <sup>s,b†</sup>	-	-	0.685 <sup>*</sup>	0.671	0.711 <sup>s,b†</sup>
	3ex	-	-	0.792	0.789	0.838 <sup>s,b†</sup>	-	-	0.736 <sup>*</sup>	0.713	0.768 <sup>s,b†</sup>
4	full	0.656	-	0.838 <sup>*</sup>	0.817	0.876 <sup>s,b†</sup>	-	-	0.854 <sup>*</sup>	0.780	0.869 <sup>s,b†</sup>
	3ex	-	-	0.847 <sup>*</sup>	0.823	0.880 <sup>s,b†</sup>	-	-	0.861 <sup>*</sup>	0.784	0.877 <sup>s,b†</sup>
6	full	0.681	-	0.781 <sup>*</sup>	0.773	0.842 <sup>s,b†</sup>	-	-	0.803 <sup>*</sup>	0.748	0.864 <sup>s,b†</sup>
	3ex	-	-	0.804	0.801	0.857 <sup>s,b†</sup>	-	-	0.813 <sup>*</sup>	0.751	0.868 <sup>s,b†</sup>

Table 3.11: Kendall’s  $\tau$  results on AES with GPT-4o and Llama for  $\Delta$ rater. P. indicates the essay prompt, and C. indicates what ratings are being compared, with ideas, organization, word choice, sentence fluency, and conventions compared for prompt 1, and content, prompt adherence, language, and narrativity compared for prompts 4 and 6.  $\tau$  is calculated with singular numerical values for  $\Delta$ rater and calculated through Pareto dominance comparison for preferences for  $\Delta$ rubric. Significance tests between separate (sep.), batch (bat.), and edited prompts are performed, where <sup>s</sup> and <sup>b</sup> in the edited prompt column represents significant differences with separate and batch prompts respectively. <sup>†</sup> is indicated next to comparisons that are significantly larger within holistic prompts. <sup>\*</sup> is indicated next to comparisons that are significantly larger between separate and batch comparisons. <sup>↑</sup> and <sup>↓</sup> represent that the  $\tau$  value for edited prompts is significantly larger or smaller respectively with the separate <sup>s</sup> or batch <sup>b</sup> prompts’  $\tau$ . The lack of any dagger, star, or arrow denotes no statistical significance. H is a shortened form for Human.

		GPT-4o					Llama				
$\Delta_{\text{rater}}$	Holistic		Analytic			Holistic		Analytic			
	0ex	3ex	Sep.	Bat.	Edited	0ex	3ex	Sep.	Bat.	Edited	
		0.536	0.585	0.464*	0.167	0.471 <sup>b†</sup>	0.470	0.578	0.445*	0.166	0.426 <sup>b†</sup>
$\Delta_{\text{rubric}}$	Ex.	$H_H \rightarrow H_A$		$LLM_H \rightarrow LLM_A$			$H_H \rightarrow H_A$		$LLM_H \rightarrow LLM_A$		
	0ex	0.534	-	0.640*	0.455	0.640 <sup>b†</sup>	0.551	-	0.531*	0.285	0.551 <sup>b†</sup>
	3ex	-	-	0.623*	0.484	0.617 <sup>b†</sup>	-	-	0.566*	0.326	0.623 <sup>b†</sup>

Table 3.12: Kendall’s  $\tau$  results on IF with GPT-4o and Llama.  $\Delta_{\text{rubric}}$  is calculated through instruction following ratio comparison for preferences. All other calculations and significance follow the methodology of Table 3.11.

## 3.8 Results

### 3.8.1 Edited Rubrics

When editing rubrics for autoraters to improve human-autorater agreement, it is important to provide examples and context and remove confirmation bias from analytic rubrics.

**Using GPT with edited analytic AES rubrics mostly significantly improves agreement with humans.** In Table 3.11 under  $\Delta_{\text{rater}}$  under the “edited” column, adding examples and context in the edited rubric always improved human-autorater agreement, significantly in the majority of cases, when using GPT, indicated by  $s, b \uparrow$ . These improvements range from 0.696 (batch prompt 4 content) to 0.698 (edited prompt 4 content) for the smaller, non-significant improvements, to 0.439 (batch prompt 1 word choice) to 0.554 (edited prompt 1 word choice), for the larger, significant improvements. While the analytic rubrics lacked explanations for the examples that the holistic AES rubrics provided, this statistically significant improvement suggests that the examples and context still provide necessary grounding for autorater scores.

**Edited analytic AES rubrics significantly improves both models’ self-alignment but not with IF rubrics.** In addition to improving human-autorater alignment, adding examples and context in the edited rubric always significantly improves GPT and Llama’s self-agreement with their scores on the original holistic rubric, although these two rubrics have inherent differences, seen in Table 3.11 under  $\Delta_{\text{rubric}}$ , indicated by  $s, b \uparrow$ . Interestingly, these agreements surpass human-human agreement with the two types of unedited rubrics (0.591, 0.656, 0.681 compared to 0.750, 0.817, 0.773 for GPT, and 0.671, 0.780, and 0.748 for Llama). This could be because the same autorater is used with both rubrics, whereas human ratings are from different individuals. This suggests that while the prompts superficially evaluate different criteria, the autorater converges to a consistent understanding of essay quality across both rubrics, which aligns with the overall rating objective, a convergence not observed to the same degree in human ratings.

However, adding examples of human analytic scores does not necessarily increase either autoraters’ self-alignment on IF. Adding examples decreased  $\tau$  with GPT-4o ( $\tau$  from 0.640 to 0.640, and 0.623 to 0.617 between separate and edited rubrics in Table 3.12), but increased  $\tau$  with Llama (0.531 to 0.551 and 0.566 to 0.623 respectively). This suggests

that autoraters in this case do not achieve a unified understanding of instruction following across prompts. Despite this, the ratio aggregation method aligns more closely with the autorater’s internal reasoning about output instruction adherence. With edited prompts, GPT-4o’s alignment (but not Llama’s) exceeded human alignment on both prompts, even though the human expert annotators across both rubrics remained consistent.

**Reducing confirmation bias in all analytic rubrics provides mostly significant improvement for both models.** Prior work has highlighted confirmation bias [36], in which conceived judgment is reinforced as the task progresses. This bias could affect analytic rubrics, leading raters to assign low scores across all criteria if they initially believe the piece of text is of poor quality. Lee et al. [97] has shown that using separate conversations to rate each sub-criteria tends to increase agreement over using a single conversation to rate all sub-criteria. However, this bias has not been statistically assessed for autoraters, and may affect autoraters differently than humans.

Across the analytic rubrics on both tasks, the majority of comparisons show separate rubrics significantly outperforming the batched rubrics. With Llama on AES and both models on IF, edited rubrics do not consistently outperform separate rubrics. This suggests the separation of the individual criteria is more important than the examples for Llama. The scoring rubric itself may contribute to the inconsistency for GPT-4o between AES and IF. In AES, example essays answered the same writing prompt as the evaluated essay, and were accompanied by the same scoring rubric during evaluation. Conversely, in IF, the examples did not correspond to the same decomposed questions being scored. The rubrics in IF are task-specific and vary significantly across instructions.

In addition, there is a trend of  $\tau$  decreasing and then increasing based on criteria order in prompts 4 and 6, observed with both GPT-4o (0.696, 0.677, 0.595, 0.652 and 0.652, 0.608, 0.510, 0.557 respectively) and Llama (0.693, 0.665, 0.573, 0.642 and 0.648, 0.601, 0.496, 0.547 respectively). This suggests potential differences in bias between humans and autoraters. This may also suggest that autoraters struggle with rating word choice or language. Qualitative analysis of model outputs for instruction following reveal that batch ratings consistently demonstrated a bias with probability distributions heavily skewed towards “yes” responses when answering yes/no decomposed questions. This contrasts with the separate API call approach, which more closely resembled human response distributions, suggesting an underlying bias effect in batch prompts.

**Task-specific AES holistic rubrics show significant improvement when reducing examples, but general AES holistic rubrics do not.** Reducing the number of examples from the full set to three (3ex) tends to decrease  $\tau$  for GPT on AES for the general holistic rubrics in essay prompts 1 and 4 ( $\tau$  dropped from 0.437 to 0.387 and 0.695 to 0.687, respectively), and increase  $\tau$  for the task-specific holistic rubric in essay prompt 6 ( $\tau$  rose from 0.629 to 0.644). This could also be because the examples have explanations for essay prompts 1 and 4, but not for 6. The score explanations could be more influential than the scores alone when examples were originally provided, but this could also be due to the type of rubric.

**Adding examples for IF holistic rubrics show improvement.** Although there is an increase in  $\tau$  when adding examples to the holistic rubric in IF for both autoraters (0.536 to 0.585 and 0.470 to 0.578 for GPT and Llama respectively), it is not statistically

significant.

### 3.8.2 Decomposition Level

Analytic rubrics do not consistently outperform holistic rubrics in aligning autoraters with human judgments due to prompt complexity or aggregation methods.

**Prompt complexity moderates effects.** In Table 3.11 for AES, under the  $\Delta$ rater condition for essay prompt 1, almost all analytic batch prompts had higher correlation with humans ( $\tau$  of 0.464, 0.455, 0.448, 0.439, 0.388) than the full holistic prompt ( $\tau$  of 0.437). Conversely in essay prompts 4 and 6, the full holistic prompt outperformed most analytic batch prompts ( $\tau$  of 0.695 compared to 0.696, 0.677, 0.652, 0.595 and  $\tau$  of 0.629 compared to 0.652, 0.608, 0.557, 0.510). This discrepancy may stem from the complexity of the holistic prompts. Essay prompt 1’s holistic prompt is highly complex, involving multiple sub-criteria that contain complex decisions, whereas, essay prompts 4 and 6 have less complex prompts. Introducing analytic rubrics may increase evaluation complexity, which leads to lower  $\tau$ .

**Aggregation methods influence agreement.** Another factor is the prompt’s output. The Pareto dominance aggregation method, a conservative estimate of essay comparison, is highly sensitive to disagreements in any single sub-criteria. In IF, both GPT-4o and Llama performed worse with analytic prompts than with holistic prompts ( $\tau$  of 0.167 to 0.536 and 0.166 to 0.470). This is surprising, given that IF analytic prompts are considered task-specific, which should provide more detailed information about the task, potentially leading to higher agreement. This may be due to the aggregation method used—the ratio of “yes” to “no” responses—which does not account for the varying weights of sub-criteria in the overall holistic evaluation. Therefore, holistic preferences, being more straightforward to calculate, may yield higher performance. Higher human-autorater agreement is not necessarily achieved by decomposing a holistic rubric into several analytic parts, rather it is more important to understand the complexity of each evaluation measure as well as the aggregation methods used.

### 3.8.3 Agreement Level

Both datasets are stratified by human agreement level since human inter-rater agreement may influence human-autorater agreement.

**High human inter-rater agreement is important.** On IF holistic rubrics, human-autorater agreement with the consolidated score was significantly higher when all three human annotators agreed under all conditions, than if only two or no annotators agreed with each other. This held for IF analytic rubrics for both the separate and edited conditions, although the batch condition had an increase, it was not significant. For AES holistic rubrics, a significant increase was only observed for prompt 4, while a non-significant increase was observed for prompt 6. Detailed analyses can be found in Appendix .3.2.

## 3.9 Conclusion

This work highlights the importance of understanding the domain, autorater, and rubric during rubric creation. First, adding examples and context significantly improves human-autorater agreement in addition to autorater self-agreement, but this is dependent on the domain and autorater. Second, rubrics that reduce confirmation bias tend to provide significant improvement for human-autorater agreement. Third, rubric complexity and aggregation methods across holistic and analytic rubrics influence human-autorater agreement. Lastly, higher human inter-rater agreement contributes positively to human-autorater agreement. Practitioners aiming to use autoraters should carefully curate human annotation data and design rubrics that appropriately address the differences across domains and autoraters. Future work should explore a wider range of domains, autoraters, and rubrics to develop more comprehensive recommendations.



# Chapter 4

## Autoraters for Simulated Personalized Evaluation

### 4.1 Introduction

In Chapter 3, we began to understand how to use autoraters for dialog evaluation in order to achieve high alignment with humans. Now in this chapter, we move toward adding an additional element to the autorater’s prompt, the injection of additional information about the user, to create personalized autoraters. This chapter focuses on attitude injection, which would ask a rater to assume an attitude or emotion, and then rate some text. It is difficult to perform this injection with humans; it is possible to ask someone to behave as someone else [39], but to simulate both an attitude as well as use the attitude to rate raises concerns due to there being no way to validate that the attitude and the rating were “simulated correctly”. Datasets for emotion recognition in a natural setting [41, 139] often have a large amount of certain emotions but not others. In order to collect additional evaluation data naturally, it would be difficult to obtain data for some attitudes, but it may also be unethical to manipulate a participant to feel a negative attitude.

Therefore, we choose to perform this injection in simulation as an indicator for what the potential of using attitude-based personalized autoraters is. It is also possible to ask LLMs to behave as humans for human experiments to success [4]. We find in the following work that simulated negative attitude ratings have more separation with each other and with simulated positive attitude ratings, which causes us to recommend focusing on negative attitudes for future work. This could make sense due to a large amount of dialog work focusing on recovering from errors. This chapter shows that there is potential for using personalized autoraters with attitude injection, and creates a foundation for additional work on personalized autoraters.

## 4.2 Sensitivity of Automatic Dialog Evaluation with Simulated Attitudes

### 4.2.1 Introduction

Users will have different goals, prior assumptions, and experiences when using dialog systems. Understanding user preferences [83], their receptiveness to persuasive strategies [160, 182], or algorithmic aversion [84] could lead to improved user experience. Therefore, current work creates and then uses representations of a user profile to customize system outputs [100, 193]. The system’s responses are then evaluated with either reference-based metrics (e.g. BLEU [133]), or reference-free metrics (e.g. USR and FED [120, 121]) which addresses the one-to-many problem in dialog evaluation. However, these metrics do not explicitly capture users’ varied preferences for a system response.

The personalization of metrics such as the coherence of a system or the interestingness of a system output is studied in this work for users with different profiles. For example, a user who is bored may find system responses less interesting than a user who is satisfied. This may be an example of the “affect heuristic”, which suggests that depending on the person and the task that the effect of the affect will differ [6, 162? ]. With this knowledge, if a system detects that a user is bored, it may present even more interesting responses to recapture that user’s attention. However, it may not change its behavior for a user who is already satisfied. By taking into account these differences, metrics can be structured to more accurately reflect the performance of a system over a diverse group of users, as found in production environments. Systems have used user modeling to adapt their outputs if necessary [21, 77, 102]. This would then inform developers with ways to strengthen the ability of a system to perform well over all users, and improve overall user satisfaction.

This work uses LLM simulated users to generate and evaluate conversations as an initial experiment into understanding the sensitivity of automatic personalized evaluation. The study consists of three models. First, a model is prompted with an attitude, which is chosen from a generated set of diverse user attitudes. A second model interacts with the simulated user for open-domain chit-chat and task-oriented conversations. Finally, a third model is prompted with the user’s attitude to rate the resulting conversation. The results show that ratings from attitudes grouped by positive and negative sentiments show more inter-sentiment than intra-sentiment differences, suggesting that sentiment may be enough to indicate personalization sensitivity. In addition, attitudes with negative sentiment tend to rate metrics lower than metrics rated with no attitude at all. Further experiments show that whether users think a system is likeable can be a good place to begin building personalized metrics that can be tailored to specific users and correlate with overall user satisfaction.

## 4.2.2 Related Work

### User-Centric Evaluation

There has been substantial work in user-centric evaluation - focusing on evaluating the experience of the user instead of other offline metrics - for dialog systems and LLMs including recommender systems [82, 89, 210] and in benchmark datasets [180]. Work has shown that metrics which are made with users in mind may or may not differ from conventional metrics that systems are evaluated with [13, 37]. However, these metrics and their importance may vary between domains and users, which is shown in this work through both open-domain and task-oriented conversations.

A user’s attributes can affect how a user rates a conversation, occurring frequently in the domain of conversation safety. Aroyo et al. [8] and Sap et al. [157] study demographics, and Wang et al. [179] compares crowd-worker annotations to expert annotations for safety. Differences in each rater’s background may also contribute to different ratings even if they share the same set of demographic information [200]. Prior work has shown that contextual information such as the objects being compared affect the attributes used during similarity rating [119], which is another aspect of personalized evaluation. Work that does consider demographic information when providing safety ratings consolidates these ratings into “safe” and “unsafe” ratings, which removes some of the personalization of evaluation [135]. Attitudes toward systems such as Alexa can also have a range of differences across demographics [49]. Even when asking the same question to a question-answering system, users may have slightly different ideas of relevance [169]. Previous work in cognitive science has looked at how differences in domain knowledge impact how people reason about physics problems [29]. Additional work has started to study user-specific classification [189]. This work begins to explore user attitudes toward the system as one factor in a user’s background for personalized evaluation.

Recent work has used personas for rating in datasets relating to toxicity detection, politeness, and irony detection through a combination of Twitter, email, and Reddit. [69] notes that there are cases where persona injection slightly improves correlation with humans. Although this work looks at different domains, persona injection is promising for automatic ratings. Other work [43] notes that “LLM-as-a-personalized-judge” and third-party human annotators have similar performance on a question-answering dataset.

### Emotion-Enhanced States

Additionally, there has been extensive work in psychology and behavioral economics that suggest a difference in behavior between a normal state and a state enhanced by emotions. Specifically, [108] mentions that the utility of decisions is affected by “visceral factors” which includes emotions, and that immediate visceral factors can cause differences “between the actual and desired value on a particular [activity]”. Supporting empirical results include the effect after shown negative affect video clips [98], the effect of positive affect on beverage consumption when shown smiling faces [187], and the effect of emotion state on various aspects of typing [96]. However, [187] also noted that the ratings that participants gave to the beverages were not affected by the emotion felt. Other works have discovered a

difference specifically in rating behavior such as rating faces for similarity [62]. Therefore, it is important to understand the potential of such behavior difference when injected into autoraters since current autoraters lack personalization.

### **Modeling (dynamic) users**

Simulated users need to both exhibit patterns in behavior such that conclusions can be drawn about the population and exhibit variations in behavior to be representative of a diverse population - the trade-off between bias and variance [24]. Platforms have been built [75, 110, 150] to try and simulate a diverse set of users by creating changeable parameters that would alter user simulation behavior. User preferences and interests may also change as the conversation progresses based on information presented by a system [125, 198]. Previous work [70] uses a wide but manageable range of simulated personas and also varies the user persona within a conversation, which this work follows.

### **4.2.3 Studying Personalized Evaluation in Dialog**

Prior work in Section 4.2.2 has shown that user attitude toward a conversation affects how the user views the system, but previous automatic metrics do not consider user attitude, and it is important to understand how different users are satisfied. This work studies how users with various user profiles converse and rate dialog systems in simulation.

### **Research Questions**

This study will show: (1) if automatic evaluation is sensitive to simulated attitudes by studying the quantitative differences between a simulated rater with and without an attitude, (2) which attitudes' ratings are separable through studying the quantitative differences between simulated raters with different attitudes on the same conversation, (3) how these factors interact in considering user satisfaction through correlating the metrics with automatically evaluated user satisfaction to discover the potential underlying components that influence user satisfaction for different users.

### **Framework**

Figure 4.1 shows the framework with three models: the user simulator (GPT-4o<sup>1</sup>), the system (GPT-4o), and the rater (either GPT-4o or Llama-2-70b-chat-hf [172]). Each conversation starts with a seed and an initial attitude for the user simulator. The task-oriented and open-domain conversations are both seeded with 50 utterances. The open domain conversations are seeded with Topical-Chat [55] train.json. Using the first utterance in each conversation, any conversations starters that are labeled as using the knowledge sources were the only ones considered to avoid any generic starters. Seeds are chosen in a semi-random method that satisfy the following conditions: does not ask for an opinion, does not make the other conversation participant sound like a human (ex. "did you watch the game

<sup>1</sup><https://openai.com/index/gpt-4o-system-card/>

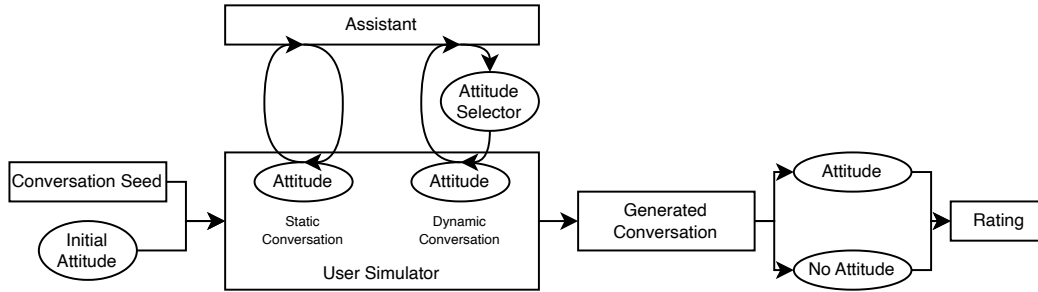


Figure 4.1: Task-oriented and open-domain conversations are seeded for the simulated user and their initial attitude. Static conversations are generated with the user simulator using the initial attitude throughout the entire conversation, while dynamic conversations are generated with an attitude selector, which chooses an attitude based on the conversation history at each turn for the user simulator. The generated conversation is then rated both on the dialog and turn level, with and without an attitude.

last night”), contains full sentences, seemingly not using context from the article associated with the conversation, and not containing any potentially sensitive topics. Seeds are also chosen such that there is a spread of topics. The task oriented conversations are seeded by asking GPT-4o to give a list of 50 tasks that someone may want to complete with an assistant (insert exact prompt) with seed 42. Temperature 0 is not used since the task list became repetitive. This task is passed as a goal for the user simulator to keep in mind for the conversation generation. The attitude represents the attitude of the simulated user toward the system, and consists of 12 attitudes, 6 positive: amusement, curiosity, relief, satisfaction, trust, engagement, and 6 negative: annoyance, boredom, confusion, disappointment, frustration, and skepticism, following the methodology of [70]. These attitudes were automatically chosen from the intersection of attitudes users may have toward open-domain and task-oriented systems, and defined based on whether the conversation was open-domain or task-oriented described below. Next, the user simulator takes the seed and the initial attitude and has a conversation with the system. There are two types of conversations: static conversations, where the user simulator keeps one attitude throughout the conversation, and dynamic conversations, where a model chooses an attitude at every turn based the conversation history. The resulting generated conversation is then rated on both the turn and dialog level, with the rater model either assuming an attitude, or not assuming an attitude.

The attitudes for the user simulation are created through the same approach as [70] using GPT-4o. The system prompt is “You are a helpful assistant”, which is the example given in the OpenAI documentation as the default prompt. GPT-4o is asked to “List twenty feelings that a user could have when having a conversation with a open-domain chitchat/task-oriented dialog system.” The prompt is run three times, each with seed 42 and temperature 0 for the highest probability of reproducibility. The three runs are combined, where attitudes that occurred in more than 2 runs and attitudes that occurred between both the task oriented and open domain results are selected, resulting in 12 attitudes.

With the same seed 42 and temperature 0, the definitions of these attitudes are created with the prompt “Give a concise definition of a person who has a feeling of while having a conversation with an open-domain chitchat/task-oriented dialog system.”. The definitions are edited slightly to remove the text that repeats the prompt at the beginning, and any third person pronouns are changed to first person pronouns. All other aspects of the definitions remain the same.

## 4.2.4 Methodology

### Prompts

**System Prompt.** The system is prompted, in Appendix .5.4, to emulate a helpful assistant, and respond concisely in the first person to reduce the verbosity of the system. Each conversation can then have more turns for the same number of tokens, which would be better for analyzing both dialog- and turn-level metrics.

**Simulated User Prompt.** The prompts used are derived from [70]. The prompt varies slightly based on the model and on the type of simulated user - a static user who does not change their attitude throughout the conversation, and a dynamic user who has the opportunity to change their attitude at each turn in the conversation. Both static and dynamic simulated user prompts instruct the LLM to have at least 5 turns in the conversation but no more than 10, return a special token similar to [70] when the simulated user feels like the conversation has come to natural end, and be concise in the responses. Conversations longer than 10 back and forth turns are truncated. For the static user, the simulated user is asked to exaggerate the given attitude for each turn, and given a definition for the attitude. The definition was created by asking GPT-4o to “give a concise definition of a person who has a feeling of [attitude] while having a conversation with [an open-domain chitchat/a task-oriented] dialog system.” Detailed prompts are found in Appendices .5.5, .5.6, and .5.7. For the dynamic user, the simulated user is asked to exaggerate the given attitude for each turn and given the same definition for the attitude in the first turn. For each subsequent turn, following [70], the simulated user can choose any of the 12 attitudes provided as how the simulated user feels for the next turn given the dialog history to facilitate conversation analysis.

The dynamic conversations are analyzed for attitude changes throughout the conversation to determine what attitudes GPT-4o tended to select for the next turn. In both open-domain and task-oriented conversations, 0 out of 300 conversations that started with a positive attitude ended with a negative attitude. However, 149 out of 300 open-domain conversations and 239 out of 300 task-oriented conversations started with a negative attitude and ended with a positive attitude. This indicates that either GPT-4o as a user simulator tends to select positive attitudes for replies, or that GPT-4o as a dialog system provides responses that satisfy itself as a user simulator. The difference in the number of conversations that ended positively could be that if the user simulator has an explicit goal that the dialog system can complete (as it is prompted to act as if it can), task-oriented conversations should tend to end positively. Both open-domain and task-oriented conversations tend to select the “engagement” attitude most frequently, with open-domain

conversations selecting it around 4 times more than the next attitude (curiosity) selected. Explicitly creating more negative conversations is a consideration for future work.

In real conversation, users may have more nuanced attitudes toward a conversation, or may take a mostly neutral stance as seen in many conversational emotion datasets [139]. However, by not including a neutral attitude, the resulting experiments are guaranteed to not represent the neutral class, but instead provide insight into the classes of interest that deviate from the norm. The exaggerated attitudes throughout the conversation provide an upper bound for this analysis as well.

## User Rating

GPT-4o is used to rate the generated conversations on the 10 dialog-level and 8 turn-level qualities found in Mehri and Eskenazi [120]. Each metric is rated separately by the model from 1 to 5 to eliminate ordering effect [109] in the rating prompt, seen in Appendix .5.7. The user rating uses a seed of 42, temperature 0, maximum returned tokens of 7, and top log probabilities of 20 for reproducibility purposes.

To retrieve the numerical rating from the LLM, the prompt specifies a return format of “rating: number” to standardize the responses, hence the maximum returned tokens of 7 to further guarantee the format. The token corresponding to the number is extracted, and used only if it is in the range of 1 through 5. The top three ratings along with the corresponding log probabilities are used to calculate the weight of the numerical from Section 4.2.4.

Definitions are created with GPT-4o using seed 42 and temperature 0 with the prompt “Give a brief definition of what it means for a [open-domain chitchat dialog system,task-oriented dialog system] to [be , have a response].”

There are two types of ratings performed - ratings with no attitude given, and ratings with a attitude given. The ratings with no attitude given are similar to [74], where ratings with no attitude given are assumed to be similar to an “average” rating across the population. For the attitude ratings, the LLM is prompted with the attitude and asked to rate as if it was feeling the attitude in the current point in the conversation. For the static ratings, the LLM is prompted with the same attitude. For the dynamic ratings, the LLM is prompted with the attitude that the user simulation model chose for the next user turn.

The expected rating is calculated based on the top three, when available, log probabilities, similar to [74]:  $\mathbb{E}[r] = \frac{\sum_{i=1}^3 p_i r_i}{\sum_{i=1}^3 p_i}$ , where  $p_i$  is the probability derived from the log-probability and  $r_i$  is the rating.

## 4.2.5 Results

SOTOPIA-EVAL’s “believability” metric is used [209] with GPT-4o, with the mean believability of the simulated user at 9.01 and the system at 8.92 out of 10, which shows that the conversations are realistic. To evaluate the believability of the simulated conversation, SOTOPIA-EVAL’s “believability” metric is used [209]. SOTOPIA’s character profiles are slightly different from the given attitudes, but the metric measures both naturalness and consistency on a scale from 0 to 10. A high score on this metric would suggest that the

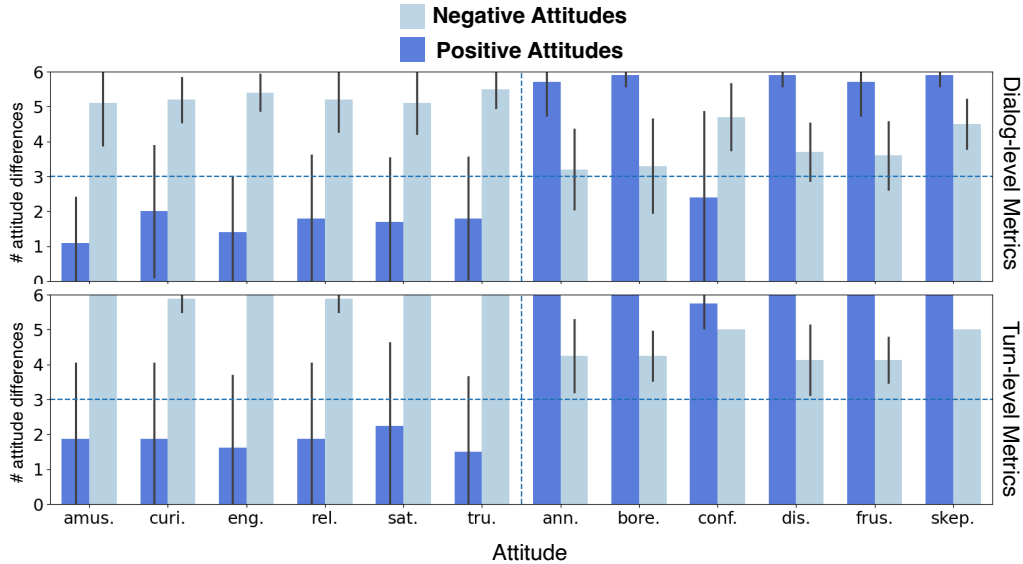


Figure 4.2: All attitudes rating the same set of open-domain conversations over all dialog- and turn-level metrics. The number of attitudes (positive or negative) that have significantly different ratings from the current attitude for each metric are counted. Each number represented in the graph corresponds to the counts for one of the metrics. The darker colored bars represent comparisons between that attitude and all positive attitudes, while the lighter colored bars represent comparisons between that attitude and all negative attitudes. The x-axis represents each attitude, and the y-axis represents how many metric comparisons were statistically significant.

simulated conversations are believable and consistent with the given attitude. The metric is used with gpt4o-08-06, to remain consistent with the current work with seed 42 and temperature 0. All open-domain and task-oriented generated conversations are analyzed. The mean believability for the simulated user is 9.0075 with standard deviations ranging from 0 to 0.463, minimum score of 8 and maximum of 10, while the mean believability for the system is 8.92 with standard deviations ranging from 0 to 0.890, minimum score of 5 and maximum of 10. This shows the believability of the conversations and attitudes. Prompts can be found in Appendix .5.9.

### Attitudes Rating The Same Conversation

All attitudes are used to rate the same set of conversations on both the dialog and turn level to fairly compare ratings across attitudes. These conversations are created by not setting an attitude for the simulated user in the conversation, and both open-domain and task-oriented conversations are generated using a similar methodology as Figure 4.1.

Within each metric, a set of paired t-tests are performed across all combinations of attitudes with a Bonferroni correction of 66. This tests whether attitude ratings are significantly different from each other when paired by ratings on each specific conversation.

	amusement		curiosity		engag.		relief		satis.		trust		annoyance		boredom		confusion		disapp.		frustration		skepticism			
	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg		
OD dialog																										
coherence	0	5	0	5	0	5	0	5	0	5	0	5	6	2	6	2	<b>0*</b>	<b>5*</b>	6	3	6	3	6	3	6	<b>5*</b>
error recovery	0	5	1	5	3	5	1	5	1	5	2	6	6	3	6	2	1	<b>5*</b>	6	3	6	4	6	4	6	<b>5*</b>
consistency	0	2	0	5	0	5	0	3	0	3	0	5	3	1	5	1	<b>0*</b>	2	6	3	3	2	6	3	6	<b>5*</b>
diversity	2	6	4	5	4	6	4	6	4	6	4	6	6	3	6	4	5	<b>5*</b>	6	3	6	3	6	3	6	4
topic depth	3	6	3	6	3	6	3	5	3	5	3	5	6	4	6	4	3	<b>5*</b>	6	<b>5*</b>	6	3	6	3	6	<b>5*</b>
likeability	1	6	4	6	0	6	1	6	1	6	1	6	6	4	6	<b>5*</b>	6	<b>5*</b>	6	4	6	4	6	4	6	4
understandingness	0	6	0	6	0	6	0	6	0	6	0	6	6	4	6	<b>5*</b>	6	<b>5*</b>	6	4	6	4	6	<b>5*</b>	6	<b>5*</b>
flexibility	2	5	2	5	2	5	<b>5*</b>	6	4	5	1	5	6	<b>5*</b>	6	3	1	<b>5*</b>	6	3	6	<b>5*</b>	6	3	6	3
informativeness	0	5	1	5	0	5	1	5	0	5	2	5	6	3	6	3	<b>0*</b>	<b>5*</b>	6	<b>5*</b>	6	3	6	3	6	<b>5*</b>
inquisitiveness	3	5	<b>5*</b>	4	2	5	3	5	4	5	<b>5*</b>	6	6	3	6	4	2	<b>5*</b>	5	4	6	4	5	4	6	4
TO dialog																										
coherence	0	6	0	6	0	6	0	6	0	6	0	6	6	3	6	4	6	3	6	4	6	3	6	3	6	<b>5*</b>
error recovery	1	2	<b>5*</b>	4	1	3	1	3	1	2	1	2	1	2	2	2	1	2	6	4	<b>0*</b>	2	6	4	6	4
consistency	0	4	0	4	0	4	0	4	0	4	0	4	1	2	6	2	5	2	6	<b>5*</b>	<b>0*</b>	2	6	3	6	<b>5*</b>
diversity	4	6	4	6	3	6	4	5	2	5	3	6	6	3	6	<b>5*</b>	4	<b>5*</b>	6	3	6	3	6	3	6	<b>5*</b>
topic depth	<b>5*</b>	2	<b>5*</b>	4	2	5	2	5	2	5	2	5	5	2	6	3	1	<b>5*</b>	4	4	5	2	5	2	5	2
likeability	4	6	4	6	2	6	2	6	2	6	2	6	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>
understandingness	0	6	0	6	0	6	0	6	0	6	0	6	6	<b>5*</b>	6	3	6	3	6	3	6	3	6	<b>5*</b>	6	<b>5*</b>
flexibility	3	4	<b>5*</b>	6	4	4	3	5	4	5	3	4	3	4	6	<b>5*</b>	3	<b>5*</b>	6	<b>5*</b>	4	4	6	5	4	6
informativeness	2	6	<b>5*</b>	2	<b>5*</b>	6	2	6	2	6	2	6	5	3	5	2	5	3	6	<b>5*</b>	5	2	6	5	4	6
inquisitiveness	4	4	<b>5*</b>	3	2	3	2	3	2	3	3	4	1	3	5	3	2	4	5	3	2	4	5	3	6	5
OD turn																										
interestingness	4	6	4	6	4	6	<b>5*</b>	6	<b>5*</b>	6	4	6	6	4	6	<b>5*</b>	6	<b>5*</b>	6	3	6	4	6	4	6	<b>5*</b>
engagingness	4	6	<b>5*</b>	6	3	6	4	6	<b>5*</b>	6	3	6	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>
specificity	<b>5*</b>	6	4	5	<b>5*</b>	6	4	5	<b>5*</b>	6	<b>5*</b>	6	6	<b>5*</b>	6	4	4	<b>5*</b>	6	<b>5*</b>	6	4	6	5	6	<b>5*</b>
relevance	2	6	2	6	0	6	2	6	2	6	0	6	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>
correctness	0	6	0	6	1	6	0	6	1	6	0	6	6	4	6	3	6	<b>5*</b>	6	3	6	4	6	5	6	<b>5*</b>
sem. appropr.	0	6	0	6	0	6	0	6	0	6	0	6	6	2	6	4	6	<b>5*</b>	6	3	6	3	6	3	6	<b>5*</b>
understandability	0	6	0	6	0	6	0	6	0	6	0	6	6	<b>5*</b>	6	4	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	4	6	<b>5*</b>
fluency	0	6	0	6	0	6	0	6	0	6	0	6	6	4	6	4	6	<b>5*</b>	6	4	6	4	6	4	6	<b>5*</b>
TO turn																										
interestingness	3	5	4	6	<b>5*</b>	6	4	6	<b>5*</b>	6	<b>5*</b>	6	6	4	6	<b>5*</b>	5	<b>5*</b>	6	3	6	4	6	4	6	<b>5*</b>
engagingness	4	5	4	5	<b>5*</b>	6	4	6	4	6	<b>5*</b>	6	6	<b>5*</b>	6	<b>5*</b>	4	<b>5*</b>	6	4	6	4	6	4	6	<b>5*</b>
specificity	<b>5*</b>	6	<b>5*</b>	6	2	6	3	6	4	6	3	6	6	4	6	4	6	4	6	4	6	3	6	3	6	<b>5*</b>
relevance	1	6	<b>5*</b>	6	1	6	1	6	1	6	1	6	6	3	6	<b>5*</b>	6	<b>5*</b>	6	3	6	3	6	3	6	<b>5*</b>
correctness	1	6	<b>5*</b>	5	1	6	1	6	1	6	1	6	6	3	5	<b>5*</b>	6	<b>5*</b>	6	3	6	3	6	3	6	<b>5*</b>
sem. appropr.	1	6	3	6	1	6	0	6	1	6	0	6	6	3	6	<b>5*</b>	6	<b>5*</b>	6	3	6	3	6	3	6	<b>5*</b>
understandability	0	6	0	6	0	6	0	6	0	6	0	6	6	4	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	4	6	<b>5*</b>
fluency	0	6	0	6	0	6	0	6	0	6	0	6	6	2	6	<b>5*</b>	6	2	6	2	6	2	6	2	6	<b>5*</b>

Table 4.1: Statistical significance for attitudes rating the same conversations, created with no attitude indicators and rated by GPT-4o. OD represents open-domain, TO represents task-oriented conversations. The positive column represents significant differences between the given attitude’s ratings and other positive attitudes’ ratings for each metric. The negative column represents significant differences between the given attitude’s ratings and other negative attitudes’ ratings for each metric. Bold and starred entries represent entries that deviate greatly from the observation that positive attitudes are different from all other negative attitudes and vice versa. The four experimental conditions are outlined on the left, along with the list of metrics under each condition.

Then, the results are aggregated for each attitude by counting how many positive attitudes and how many negative attitudes were significantly different across each metric. The open-domain conversations results are shown in Figure 4.2. Each data point in the bar graph represents the number of counted differences for one metric. Full results can be seen in Table 4.1.

If the ratings from each attitude for each metric can be grouped distinctly, then the metrics can be used as a basis for a personalized metric. This can be calculated through intra-class correlation (ICC) with a single-rater, two-way mixed effects model using consistency agreement (ICC3 from Vallat [174] version 0.5.4) and interpreted with guidelines from [92]. High agreement suggests that a linear transformation exists; however, low agreement can exist for data that is not varied [92]. All ICC scores are significant and calculated with a 95% confidence interval. All metrics on open-domain conversations have at least

moderate agreement except for error recovery. The conversations were not explicitly generated with errors, which may contribute to the low agreement. However, 7 metrics have poor agreement on task-oriented conversations: coherence, error recovery, consistency, likeability, understandingness, understandability, and fluency. Most of these metrics suffer from the ceiling effect in both the open-domain and task-oriented conversations, where the ratings tend toward 5. This indicates that attitudes rating task-oriented conversations may be less distinguishable.

Some metrics are less suitable for personalization. In open-domain conversations, six metrics (coherence, understandingness, flexibility, semantic appropriateness, understandability, fluency) have moderate agreement. These metrics have some of the lowest average ranges (.42 - 1.65) and average standard deviations (.14 - .93), which may result in lower agreement. Five out of six metrics also experience the ceiling effect for ratings from positive attitudes seen in Table 4.2 with low average ranges (.15 - 1.46) and average standard deviations (.02 - .28), while the average rating is around 5. The metrics with excellent agreement (topic depth, informativeness, inquisitiveness, interestingness, engagingness, specificity) have high average ranges (2.38 - 3.64) and average standard deviations (1.37 - 2.25) within all attitude ratings for a metric. For task-oriented conversations, the metrics with poor agreement with the exception of error recovery have the lowest average ranges (.3 - .73) and standard deviations (0.08 - 0.32). Coherence, consistency, understandingness, understandability and fluency all have the ceiling effect as well in Table 4.3. For the four metrics with excellent agreement: topic depth, inquisitiveness, interestingness, and specificity, they have some of the highest average ranges (2.11 - 2.64) and average standard deviations (1.18 - 1.71).

The overlap between metrics that have both excellent ICC and varied metric ratings for both open-domain and task-oriented conversations (topic depth, inquisitiveness, interestingness and specificity) suggests that these metrics are potentially good candidates to use for personalized metrics as long as they can be distinguished from each other.

**$\Delta$ positive attitudes,  $\Delta$ negative attitudes.** Across all figures, in Figure 4.2 for open-domain conversations, positive attitude ratings have significant differences compared to other positive attitude ratings and negative attitude ratings have significant differences compared to other negative attitude ratings as the average differences are all above 0. Negative attitude ratings tend to have more differences with each other than positive attitude ratings, which means that they are slightly more separable from each other. Task-oriented conversation ratings tend to exhibit similar tendencies.

For task-oriented conversations in Figure 4.3, it is interesting to note that for dialog-level metrics, positive attitude ratings had on average more differences with each other than open-domain conversations for amusement, curiosity, and engagement, while negative attitude ratings had a lower amount of differences with each other for confusion, frustration, and skepticism. On turn-level metrics, positive attitude ratings had on average more differences with each other than open-domain conversations for curiosity, engagement, and trust, while negative attitude ratings had a lower amount of differences with each other for annoyance, confusion, disappointment, and frustration. It is also of note that the curiosity

Dialog-level Metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	5.0	4.99	4.99	4.99	5.0	5.0	4.83+	4.81+	4.97-	4.79+	4.84+	4.51+-
error recovery	4.57	4.52	4.49	4.65	4.63	4.63-	3.6+	3.71+	4.47-	3.64+	3.95+	2.36+-
consistency	4.99	4.98	4.98	4.99	4.99	4.98	4.91	4.94	4.93	4.88+	4.91	4.73+-
diversity	3.95-	3.86	3.95-	4.19-	4.18-	4.04-	3.25+	3.32+	3.82-	3.13+	3.19+	3.07+
topic depth	4.05-	4.07-	4.06-	4.17	4.19	4.19	3.71+	3.62+	4.21-	3.47+-	3.65+	3.32+-
likeability	4.92-	4.88-	4.91-	4.95-	4.98-	4.91-	4.07+	3.63+-	4.79+-	3.84+	4.12+	3.76+
understandingness	4.99-	4.97-	4.99-	4.99-	5.0-	4.99-	4.5+	4.37+-	4.92+-	4.49+	4.68+-	4.16+-
flexibility	4.84	4.78	4.85	4.94+-	4.9	4.83	4.41+-	4.17+	4.84-	4.2+	4.5+-	4.09+
informativeness	4.42	4.36	4.42	4.39	4.43	4.46	3.87+	3.85+	4.39-	3.7+-	3.86+	3.59+-
inquisitiveness	4.1	3.49+	4.08	4.14	3.99	4.25+-	3.72+	3.82+	4.04-	3.44	3.65+	3.48

Turn-level Metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
interestingness	4.0-	3.82-	3.97-	3.91+-	4.16+	3.85-	3.02+	2.55+	3.69+-	3.06+	3.08+	2.89+-
engagingness	4.69-	4.54+-	4.66-	4.64-	4.79+-	4.68-	3.92+-	3.09+-	4.47+-	3.78+-	3.88+-	3.7+-
specificity	4.49+-	4.34	4.41+-	4.37	4.53+-	4.45+-	3.88+-	3.93+	4.35-	3.84+-	3.96+	3.69+-
relevance	4.97-	4.95-	4.96-	4.95-	4.98-	4.97-	4.51+-	4.42+-	4.88+-	4.68+-	4.57+-	4.23+-
correctness	4.99-	4.98-	4.98-	4.98-	4.99-	4.98-	4.79+	4.79+	4.92+-	4.76+	4.73+	4.37+-
sem. appropri.	4.99-	4.98-	4.98-	4.98-	4.99-	4.99-	4.81+	4.77+	4.96+-	4.83+	4.83+	4.51+-
understandability	5.0-	5.0-	5.0-	5.0-	5.0-	5.0-	4.86+-	4.9+	4.96+-	4.84+-	4.9+	4.67+-
fluency	5.0-	4.99-	4.99-	4.99-	5.0-	4.99-	4.88+	4.92+	4.97+-	4.9+	4.88+	4.7+-

Table 4.2: All attitude rating the same set of open-domain conversations. “+” signs indicate that the set of ratings for the attitude on the metric, when compared with each set of ratings for all positive attitudes, was found significantly different in every case. “-” signs indicate similarly for the set of ratings when compared with each set of ratings for all negative attitudes.

attitude ratings for task-oriented conversations are on average different from 3.3 and 3.25 other positive attitudes on the dialog and turn level - higher than the other average differences that are around 2.3 or below for both open-domain and task-oriented conversations. This shows that there are some differences between the open-domain and task-oriented conversations, where positive attitude ratings may be more easily distinguished across the metrics for task-oriented conversations, and negative attitude ratings may be more easily distinguished for open-domain conversations.

Positive and negative attitudes being distinguishable from each other exists for task-oriented conversations as well; however, there are some differences in the intensity of the trend presented in Table 4.1 and figure 4.3. At the dialog level, negative attitude ratings when compared with all positive attitude ratings have on average less differences than the open-domain conversations (except for confusion), and for positive attitude ratings when compared with all negative ratings, the same trend occurs. At the turn level, positive attitude ratings, when compared with negative attitude ratings, and vice versa, all have average differences of above 5.7, which means that they are able to be distinguished from a majority of the opposing sentiment attitudes’ ratings.

The means of positive attitude ratings for dialog-level metrics that had significant differences with each negative ratings are high - for open-domain conversations, between 5.1 and 5.5, and for task-oriented conversations, between 4.6 and 4.9. The turn-level comparisons for both types of conversations have a small range as well, between 5.75 and 6. There are some surprising results when looking at the difference between each negative attitude ratings and all positive attitude ratings. All of the negative attitudes have a large amount of differences except for confusion.

Dialog-level Metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	5.0-	5.0-	5.0-	5.0-	5.0-	5.0-	4.9+	4.85+	4.93+	4.81+	4.92+	4.11+-
error recovery	3.88	3.42+	3.87	3.95	3.74	3.76	3.81	3.63	3.79	2.67+	3.55	2.77+
consistency	5.0	4.99	5.0	5.0	5.0	5.0	4.96	4.97+	4.98	4.92+-	4.96	4.28+-
diversity	3.14-	3.11-	3.27-	3.38	3.34	3.28-	2.9+	2.51+-	3.37-	2.84+	2.85+	2.62+-
topic depth	1.98+	2.04+	2.11	2.11	2.11	2.11	1.9	1.88+	2.09-	2.0	1.96	1.93
likeability	4.82-	4.83-	4.97-	4.98-	4.99-	4.99-	4.24+-	3.36+-	4.54+-	3.94+-	4.41+-	3.51+-
understandingness	4.99-	4.98-	4.99-	4.99-	5.0-	5.0-	4.79+-	4.69+	4.68+	4.66+	4.85+-	3.96+-
flexibility	4.2	3.78+-	4.11	4.33	4.38	4.27	4.2	3.37+-	4.38-	3.96+-	4.12	3.6+-
informativeness	4.85-	4.53+	4.75+-	4.83-	4.85-	4.84-	4.48	4.51	4.62	4.24+-	4.49	3.71+-
inquisitiveness	3.79	3.42+	3.83	3.89	3.9	3.94	3.86	3.51	3.92	3.47	3.81	3.47

Turn-level Metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
interestingness	2.77	2.65-	3.1+-	2.91-	2.99+-	3.05+-	2.46+	2.25+-	2.86-	2.45+	2.42+	2.35+-
engagingness	4.0	3.99	4.43+-	4.3-	4.29-	4.38+-	3.64+-	3.14+-	4.03-	3.58+	3.6+	3.36+-
specificity	4.83+-	4.66+-	4.77-	4.78-	4.75-	4.79-	4.43+	4.62+	4.58+	4.38+	4.42+	4.03+-
relevance	4.98-	4.96+-	4.98-	4.98-	4.98-	4.98-	4.82+	4.92+-	4.89+-	4.8+	4.8+	4.29+-
correctness	4.91-	4.77+	4.9-	4.9-	4.88-	4.88-	4.44+	4.78-	4.6+-	4.42+	4.41+	3.91+-
sem. appropri.	4.99-	4.98-	4.99-	4.99-	4.99-	4.99-	4.83+	4.94+-	4.91+-	4.82+	4.81+	4.19+-
understandability	5.0-	4.99-	5.0-	5.0-	5.0-	5.0-	4.88+	4.96+-	4.56+-	4.91+-	4.88+	4.31+-
fluency	5.0-	5.0-	5.0-	5.0-	5.0-	5.0-	4.94+	4.96+-	4.93+	4.94+	4.94+	4.51+-

Table 4.3: All attitude rating the same set of task-oriented conversations. “+” signs indicate that the set of ratings for the attitude on the metric, when compared with each set of ratings for all positive attitudes, was found significantly different in every case. “-” signs indicate similarly for the set of ratings when compared with each set of ratings for all negative attitudes.

**$\Delta$ between positive and negative attitudes.** In Figure 4.2 for open-domain conversations, for both dialog and turn-level metrics, each set of positive attitude ratings is largely different from most negative attitude ratings, and each set of negative attitude ratings is largely different from most positive attitude ratings.

Attention to specific attitudes may be needed to accurately calculate most metrics due to subjectivity. However, it is surprising that relevance and understandability, metrics that are expected to have static definitions, have significant differences across attitudes. Both scores are high across all attitudes in Table 4.2, but have an average rating of almost 5 on positive attitudes, which may lead to noticeable differences with ratings from negative attitudes which are not 5. Similar trends occur for fluency, semantic appropriateness, correctness, and relevance on the turn level, and understandability on the dialog level.

This suggests that significant separation exists between the groups of positive and negative attitudes when rating the same conversation, with some separation within the groups. Therefore, given one conversation, it is important to understand both the sentiment of the user and user’s attitude because the ratings will differ. This also suggests that users with certain attitudes may require different responses to rate certain metrics highly.

### Attitude vs. No Attitude Rating

It is also important to compare automatic personalized metrics with automatic metrics which use no attitude (“no attitude ratings”), which is how many automatic dialog evaluations have been performed to date. Studies have shown a clear difference between no

Open-domain static conversations, dialog-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	(5.0, 5.0)	(5.0, 5.0)	(4.99, 4.99)	(5.0, 5.0)	(5.0, 5.0)	(4.98, 4.98)	<b>(4.66, 3.89)</b>	<b>(4.95, 4.32)</b>	<b>(5.0, 4.81)</b>	<b>(4.92, 4.29)</b>	<b>(4.88, 4.22)</b>	<b>(4.96, 4.15)</b>
error recovery	<b>(3.82, 4.42)</b>	<b>(4.64, 4.79)</b>	<b>(3.88, 4.46)</b>	<b>(4.12, 4.74)</b>	<b>(3.79, 4.5)</b>	<b>(3.59, 4.44)</b>	<b>(3.94, 3.23)</b>	<b>(3.69, 3.36)</b>	<b>(4.71, 4.53)</b>	<b>(4.2, 3.53)</b>	<b>(4.2, 3.54)</b>	<b>(4.4, 3.82)</b>
consistency	(5.0, 5.0)	(5.0, 5.0)	(4.98, 4.98)	(4.99, 5.0)	(5.0, 5.0)	(4.97, 4.98)	<b>(4.75, 4.29)</b>	<b>(4.99, 4.76)</b>	<b>(5.0, 4.91)</b>	<b>(4.97, 4.52)</b>	<b>(4.95, 4.5)</b>	<b>(4.97, 4.42)</b>
diversity	<b>(3.94, 4.13)</b>	(4.81, 4.82)	(4.2, 4.27)	<b>(3.89, 4.29)</b>	<b>(3.61, 3.88)</b>	<b>(3.52, 3.69)</b>	<b>(3.48, 2.9)</b>	<b>(3.68, 3.02)</b>	(4.0, 4.0)	<b>(3.72, 2.96)</b>	<b>(3.6, 2.94)</b>	<b>(3.83, 3.39)</b>
topic depth	(3.76, 3.7)	(4.89, 4.87)	(4.18, 4.17)	(4.08, 4.18)	<b>(3.68, 3.83)</b>	<b>(3.84, 3.94)</b>	<b>(3.2, 2.81)</b>	<b>(2.59, 2.31)</b>	(4.17, 4.1)	<b>(3.73, 2.92)</b>	<b>(3.54, 3.02)</b>	<b>(3.75, 2.97)</b>
likeability	(4.99, 4.99)	(5.0, 5.0)	(4.99, 5.0)	(4.99, 5.0)	(4.99, 5.0)	<b>(4.87, 4.93)</b>	<b>(3.33, 2.5)</b>	<b>(3.57, 2.48)</b>	<b>(4.77, 4.35)</b>	<b>(4.18, 2.96)</b>	<b>(3.62, 2.8)</b>	<b>(3.83, 2.75)</b>
understandingness	(5.0, 5.0)	(4.99, 4.99)	(4.99, 5.0)	(4.99, 5.0)	(5.0, 5.0)	(4.95, 4.98)	<b>(4.33, 3.22)</b>	<b>(4.58, 3.62)</b>	<b>(4.98, 4.67)</b>	<b>(4.8, 3.74)</b>	<b>(4.61, 3.67)</b>	<b>(4.76, 3.57)</b>
flexibility	<b>(4.59, 4.8)</b>	(4.99, 4.98)	(4.77, 4.85)	<b>(4.64, 4.98)</b>	<b>(4.45, 4.71)</b>	<b>(4.27, 4.48)</b>	<b>(4.25, 3.61)</b>	<b>(4.44, 3.62)</b>	<b>(4.91, 4.75)</b>	<b>(4.61, 3.64)</b>	<b>(4.49, 3.84)</b>	<b>(4.68, 3.82)</b>
informativeness	<b>(3.69, 3.8)</b>	(4.88, 4.88)	<b>(4.15, 4.27)</b>	(4.0, 4.02)	<b>(3.69, 3.87)</b>	<b>(3.84, 4.05)</b>	<b>(3.76, 3.09)</b>	<b>(3.55, 3.07)</b>	<b>(4.82, 4.57)</b>	<b>(3.99, 3.13)</b>	<b>(4.01, 3.33)</b>	<b>(4.4, 3.51)</b>
inquisitiveness	<b>(3.16, 3.4)</b>	<b>(4.69, 4.28)</b>	<b>(3.31, 3.66)</b>	<b>(3.2, 3.45)</b>	<b>(2.91, 3.08)</b>	<b>(2.95, 3.27)</b>	(2.74, 2.62)	<b>(3.48, 3.17)</b>	<b>(2.54, 3.07)</b>	<b>(3.02, 2.75)</b>	<b>(2.94, 2.74)</b>	(2.4, 2.34)

Open-domain dynamic conversations, dialog-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(4.98, 4.98)	(0, 0)	<b>(4.79, 4.07)</b>	<b>(4.94, 4.23)</b>	(0, 0)	(4.73, 3.75)	<b>(4.94, 4.35)</b>	<b>(4.97, 4.16)</b>
error recovery	<b>(4.26, 4.63)</b>	<b>(4.62, 4.78)</b>	<b>(4.51, 4.73)</b>	(4.43, 4.6)	<b>(4.49, 4.69)</b>	(0, 0)	<b>(4.08, 3.29)</b>	<b>(3.27, 3.02)</b>	(0, 0)	(3.61, 2.8)	<b>(4.4, 3.74)</b>	<b>(4.53, 3.71)</b>
consistency	(5.0, 5.0)	(5.0, 5.0)	(4.99, 4.99)	(5.0, 5.0)	(4.99, 4.99)	(0, 0)	<b>(4.9, 4.42)</b>	<b>(4.98, 4.73)</b>	(0, 0)	(4.93, 4.39)	<b>(4.98, 4.58)</b>	<b>(4.98, 4.42)</b>
diversity	(4.49, 4.62)	(4.54, 4.58)	(4.44, 4.49)	(3.71, 3.84)	<b>(4.27, 4.47)</b>	(0, 0)	<b>(3.61, 2.99)</b>	<b>(3.58, 2.86)</b>	(0, 0)	(3.37, 2.74)	<b>(3.61, 2.92)</b>	<b>(4.0, 3.41)</b>
topic depth	(3.93, 3.81)	(4.79, 4.78)	<b>(4.57, 4.53)</b>	(3.59, 3.87)	(4.42, 4.43)	(0, 0)	<b>(3.16, 2.71)</b>	<b>(2.54, 2.24)</b>	(0, 0)	(3.62, 2.81)	<b>(3.5, 2.89)</b>	<b>(4.0, 3.3)</b>
likeability	(5.0, 5.0)	(4.99, 4.99)	(4.98, 4.99)	(4.99, 5.0)	(4.91, 4.94)	(0, 0)	<b>(3.68, 2.67)</b>	<b>(3.61, 2.4)</b>	(0, 0)	(4.34, 2.84)	<b>(4.2, 3.22)</b>	<b>(4.17, 3.16)</b>
understandingness	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(4.97, 4.97)	(0, 0)	<b>(4.45, 3.39)</b>	<b>(4.51, 3.4)</b>	(0, 0)	(4.66, 3.39)	<b>(4.79, 3.86)</b>	<b>(4.85, 3.79)</b>
flexibility	(4.95, 4.96)	(4.98, 4.99)	(4.94, 4.96)	(4.24, 4.6)	<b>(4.88, 4.96)</b>	(0, 0)	<b>(4.48, 3.75)</b>	<b>(4.16, 3.38)</b>	(0, 0)	(4.74, 3.46)	<b>(4.71, 4.0)</b>	<b>(4.73, 3.91)</b>
informativeness	(4.28, 4.26)	(4.98, 4.96)	<b>(4.7, 4.66)</b>	(3.3, 3.37)	(4.67, 4.68)	(0, 0)	<b>(3.91, 3.19)</b>	<b>(3.46, 2.91)</b>	(0, 0)	(3.8, 2.9)	<b>(4.16, 3.4)</b>	<b>(4.55, 3.69)</b>
inquisitiveness	(3.2, 3.5)	(4.05, 4.03)	<b>(4.07, 4.32)</b>	(2.69, 2.84)	<b>(3.74, 3.85)</b>	(0, 0)	<b>(3.07, 2.88)</b>	<b>(3.29, 2.85)</b>	(0, 0)	(1.88, 2.04)	(2.85, 2.69)	(2.63, 2.62)

Task-oriented static conversations, dialog-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	(4.99, 4.98)	(5.0, 5.0)	(4.97, 4.99)	(4.99, 5.0)	(4.96, 4.98)	(5.0, 5.0)	<b>(4.96, 4.34)</b>	<b>(4.98, 4.53)</b>	<b>(4.99, 4.49)</b>	<b>(4.84, 4.2)</b>	<b>(4.91, 4.28)</b>	<b>(4.96, 3.97)</b>
error recovery	<b>(3.47, 4.23)</b>	(4.7, 4.7)	<b>(3.5, 4.34)</b>	<b>(3.27, 4.37)</b>	<b>(3.18, 4.32)</b>	<b>(3.28, 4.17)</b>	(3.83, 3.95)	<b>(3.25, 3.71)</b>	<b>(4.66, 4.33)</b>	(3.38, 3.22)	(3.76, 3.7)	(3.7, 3.53)
consistency	(5.0, 4.99)	(5.0, 5.0)	(4.96, 4.97)	(4.99, 5.0)	(4.97, 4.96)	(5.0, 5.0)	<b>(4.98, 4.65)</b>	<b>(4.99, 4.86)</b>	<b>(5.0, 4.85)</b>	<b>(4.93, 4.46)</b>	<b>(4.98, 4.61)</b>	<b>(4.98, 4.26)</b>
diversity	(4.48, 4.55)	<b>(4.87, 4.82)</b>	(4.09, 4.15)	<b>(3.62, 3.89)</b>	<b>(3.81, 4.03)</b>	<b>(3.54, 3.64)</b>	<b>(3.16, 2.82)</b>	<b>(2.98, 2.3)</b>	(3.79, 3.89)	<b>(3.14, 2.73)</b>	<b>(3.04, 2.67)</b>	<b>(3.32, 2.77)</b>
topic depth	<b>(1.95, 2.14)</b>	(4.64, 4.62)	<b>(2.25, 2.46)</b>	<b>(1.91, 2.28)</b>	<b>(2.04, 2.33)</b>	<b>(2.05, 2.41)</b>	(1.74, 1.88)	(1.56, 1.62)	<b>(2.61, 2.89)</b>	<b>(1.91, 2.05)</b>	<b>(1.82, 1.92)</b>	(2.14, 2.26)
likeability	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(4.96, 5.0)	(5.0, 5.0)	<b>(3.86, 3.33)</b>	<b>(4.13, 2.89)</b>	<b>(4.86, 4.11)</b>	<b>(3.9, 3.1)</b>	<b>(3.72, 3.18)</b>	<b>(4.17, 3.26)</b>
understandingness	(4.98, 4.97)	(5.0, 5.0)	(5.0, 5.0)	(4.99, 5.0)	(4.96, 5.0)	(5.0, 5.0)	<b>(4.85, 4.07)</b>	<b>(4.92, 4.24)</b>	<b>(4.97, 4.05)</b>	<b>(4.67, 3.82)</b>	<b>(4.71, 4.0)</b>	<b>(4.89, 3.81)</b>
flexibility	(4.92, 4.93)	(5.0, 4.98)	(4.9, 4.86)	<b>(4.71, 4.86)</b>	<b>(4.75, 4.93)</b>	<b>(4.63, 4.77)</b>	<b>(4.47, 3.94)</b>	<b>(4.01, 3.21)</b>	<b>(4.94, 4.55)</b>	<b>(4.26, 3.7)</b>	<b>(4.36, 3.8)</b>	<b>(4.69, 3.97)</b>
informativeness	<b>(4.55, 4.65)</b>	(4.98, 4.98)	(4.82, 4.86)	<b>(4.67, 4.87)</b>	<b>(4.68, 4.9)</b>	<b>(4.75, 4.86)</b>	<b>(4.47, 3.96)</b>	<b>(4.34, 4.04)</b>	<b>(4.81, 4.19)</b>	<b>(4.2, 3.62)</b>	<b>(4.33, 3.85)</b>	<b>(4.55, 3.78)</b>
inquisitiveness	(3.45, 3.53)	<b>(4.65, 4.51)</b>	(3.69, 3.73)	<b>(3.62, 3.84)</b>	<b>(3.37, 3.48)</b>	(3.6, 3.69)	(3.4, 3.52)	<b>(3.41, 3.04)</b>	(4.33, 4.25)	<b>(3.35, 3.16)</b>	(3.28, 3.36)	(3.39, 3.44)

Task-oriented dynamic conversations, dialog-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	(4.97, 4.99)	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(4.97, 4.98)	(4.98, 4.99)	<b>(4.94, 4.37)</b>	<b>(5.0, 4.53)</b>	(0, 0)	(4.99, 4.37)	<b>(4.75, 4.14)</b>	<b>(4.98, 3.84)</b>
error recovery	(3.61, 4.45)	(4.66, 4.59)	<b>(3.91, 4.52)</b>	<b>(3.8, 4.57)</b>	<b>(3.6, 4.36)</b>	<b>(3.44, 4.13)</b>	(3.57, 3.75)	<b>(3.12, 3.76)</b>	(0, 0)	(2.09, 3.14)	(4.08, 3.67)	(3.61, 3.26)
consistency	(4.97, 4.98)	(5.0, 4.98)	(5.0, 5.0)	(5.0, 5.0)	(4.98, 4.98)	(4.97, 4.99)	<b>(4.99, 4.61)</b>	<b>(5.0, 4.88)</b>	(0, 0)	(5.0, 4.8)	<b>(4.7, 4.43)</b>	<b>(5.0, 4.08)</b>
diversity	(4.22, 4.23)	<b>(4.63, 4.42)</b>	(4.22, 4.2)	<b>(3.31, 3.49)</b>	<b>(3.61, 3.74)</b>	(3.24, 3.35)	<b>(3.12, 2.68)</b>	<b>(2.98, 2.21)</b>	(0, 0)	(2.32, 2.24)	<b>(3.51, 2.99)</b>	<b>(3.27, 2.67)</b>
topic depth	(2.24, 2.35)	(4.17, 4.12)	<b>(3.19, 3.31)</b>	<b>(1.69, 2.02)</b>	<b>(2.23, 2.47)</b>	<b>(1.91, 2.17)</b>	(1.98, 1.88)	(1.69, 1.75)	(0, 0)	(1.01, 1.09)	(2.38, 2.34)	(2.15, 2.19)
likeability	(5.0, 5.0)	(5.0, 5.0)	(4.99, 4.99)	(4.8, 4.83)	<b>(4.95, 4.97)</b>	(4.79, 4.9)	<b>(3.81, 3.12)</b>	<b>(4.25, 2.88)</b>	(0, 0)	(3.94, 2.99)	<b>(4.04, 3.32)</b>	<b>(4.26, 3.19)</b>
understandingness	(4.99, 4.99)	(5.0, 4.99)	(5.0, 5.0)	(4.99, 4.99)	(4.98, 4.99)	(4.99, 4.99)	<b>(4.87, 4.03)</b>	<b>(4.95, 4.34)</b>	(0, 0)	(4.97, 3.93)	<b>(4.75, 3.89)</b>	<b>(4.95, 3.66)</b>
flexibility	(4.91, 4.86)	(4.96, 4.89)	<b>(4.88, 4.84)</b>	<b>(4.68, 4.78)</b>	<b>(4.73, 4.84)</b>	(4.57, 4.67)	<b>(4.32, 3.77)</b>	<b>(4.01, 3.12)</b>	(0, 0)	(3.12, 3.33)	<b>(4.38, 3.79)</b>	<b>(4.61, 3.65)</b>
informativeness	(4.69, 4.75)	(4.97, 4.94)	(4.86, 4.89)	(4.84, 4.86)	<b>(4.85, 4.91)</b>	(4.43, 4.59)	<b>(4.46, 3.94)</b>	<b>(4.53, 4.11)</b>	(0, 0)	(4.14, 3.31)	<b>(4.5, 3.92)</b>	<b>(4.48, 3.56)</b>
inquisitiveness	(3.67, 3.81)	<b>(4.52, 4.15)</b>	(3.92, 3.94)	(3.38, 3.38)	<b>(3.67, 3.72)</b>	(3.66, 3.76)	(3.18, 3.17)	<b>(3.45, 2.97)</b>	(0, 0)	(1.98, 2.02)	(3.42, 3.38)	<b>(3.81, 3.53)</b>

Table 4.4: Ratings for conversations for (no attitude, attitude) on GPT-4o for dialog-level metrics. Pairs of numbers in bold represent statistically significant differences when a paired t-test is conducted between no attitude ratings and attitude ratings for the same metric on the same set of conversations.

Open-domain static conversations, turn-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	amoyance	boredom	confusion	disapp.	frustration	skepticism
interestingness	(4.16, 4.4)	(4.45, 4.54)	(4.24, 4.46)	(3.98, 4.19)	(4.0, 4.39)	(3.82, 4.01)	(3.33, 2.76)	(3.16, 2.24)	(3.69, 3.67)	(3.62, 2.95)	(3.46, 2.87)	(3.45, 2.86)
engagingness	<b>(4.82, 4.85)</b>	<b>(4.85, 4.82)</b>	<b>(4.81, 4.86)</b>	<b>(4.76, 4.82)</b>	<b>(4.73, 4.89)</b>	<b>(4.65, 4.74)</b>	(4.19, 3.43)	(4.09, 2.68)	(4.51, 4.33)	(4.4, 3.52)	(4.27, 3.47)	(4.22, 3.49)
specificity	<b>(4.35, 4.52)</b>	(4.76, 4.76)	<b>(4.46, 4.53)</b>	<b>(4.27, 4.34)</b>	<b>(4.19, 4.43)</b>	<b>(4.21, 4.36)</b>	(4.1, 3.56)	(3.78, 3.34)	(4.6, 4.51)	(4.18, 3.61)	(4.16, 3.68)	(4.32, 3.71)
relevance	(5.0, 5.0)	<b>(4.99, 4.99)</b>	(4.97, 4.97)	(4.99, 4.99)	(4.99, 4.99)	(4.99, 4.99)	(4.97, 4.98)	(4.87, 4.0)	(4.9, 4.0)	<b>(4.99, 4.88)</b>	(4.96, 4.27)	(4.91, 4.12)
correctness	(5.0, 5.0)	(4.98, 4.99)	(4.97, 4.97)	(4.99, 4.99)	<b>(5.0, 5.0)</b>	(4.98, 4.98)	<b>(4.88, 4.25)</b>	<b>(4.96, 4.48)</b>	<b>(4.98, 4.92)</b>	<b>(4.94, 4.32)</b>	<b>(4.93, 4.26)</b>	<b>(4.95, 4.12)</b>
sem. appropri.	(5.0, 5.0)	(5.0, 5.0)	(4.98, 4.98)	(4.99, 4.99)	(5.0, 5.0)	(4.99, 4.99)	<b>(4.91, 4.31)</b>	<b>(4.97, 4.42)</b>	<b>(5.0, 4.94)</b>	<b>(4.98, 4.44)</b>	<b>(4.96, 4.39)</b>	<b>(4.99, 4.19)</b>
understandability	(5.0, 5.0)	(5.0, 5.0)	(4.98, 4.98)	(4.99, 5.0)	<b>(5.0, 5.0)</b>	(5.0, 5.0)	<b>(4.96, 4.54)</b>	<b>(5.0, 4.79)</b>	<b>(5.0, 4.89)</b>	<b>(4.98, 4.5)</b>	<b>(4.99, 4.62)</b>	<b>(4.99, 4.36)</b>
fluency	(5.0, 5.0)	(5.0, 5.0)	(4.98, 4.98)	(4.99, 4.99)	(5.0, 5.0)	(5.0, 5.0)	<b>(4.98, 4.71)</b>	<b>(5.0, 4.86)</b>	<b>(5.0, 4.96)</b>	<b>(4.98, 4.71)</b>	<b>(4.99, 4.7)</b>	<b>(5.0, 4.47)</b>
Open-domain dynamic conversations, turn-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	amoyance	boredom	confusion	disapp.	frustration	skepticism
interestingness	<b>(4.25, 4.48)</b>	<b>(4.05, 4.24)</b>	<b>(4.15, 4.42)</b>	<b>(3.58, 3.85)</b>	<b>(3.84, 4.3)</b>	(2.51, 2.91)	(3.12, 2.6)	(3.19, 2.21)	(3.47, 3.42)	(3.56, 2.93)	(2.99, 2.54)	(3.58, 2.99)
engagingness	<b>(4.76, 4.84)</b>	(4.64, 4.65)	<b>(4.78, 4.84)</b>	<b>(4.49, 4.62)</b>	<b>(4.63, 4.85)</b>	(3.99, 4.16)	(4.11, 3.34)	(4.07, 2.55)	(4.45, 4.18)	(4.33, 3.49)	(4.16, 3.38)	(4.29, 3.5)
specificity	<b>(4.45, 4.64)</b>	(4.52, 4.53)	<b>(4.53, 4.61)</b>	(4.2, 4.3)	<b>(4.4, 4.59)</b>	(3.11, 3.49)	<b>(3.85, 3.38)</b>	<b>(3.75, 3.26)</b>	(4.3, 4.13)	(4.05, 3.54)	(3.94, 3.43)	(4.33, 3.82)
relevance	(4.97, 4.99)	<b>(4.99, 4.98)</b>	(4.99, 4.99)	(4.98, 4.98)	<b>(4.99, 4.99)</b>	(4.82, 4.78)	<b>(4.85, 3.97)</b>	<b>(4.87, 3.9)</b>	(4.9, 4.5)	(4.91, 4.21)	(4.91, 4.1)	(4.97, 4.13)
correctness	(4.97, 4.99)	(4.99, 4.99)	(4.99, 4.99)	(4.98, 4.99)	(4.99, 5.0)	(4.97, 4.98)	<b>(4.89, 4.22)</b>	<b>(4.96, 4.43)</b>	(4.79, 4.65)	<b>(4.9, 4.27)</b>	<b>(4.94, 4.2)</b>	<b>(4.96, 4.13)</b>
sem. appropri.	(4.99, 5.0)	(5.0, 5.0)	(4.99, 4.99)	(4.99, 4.99)	(5.0, 5.0)	(4.95, 4.98)	<b>(4.93, 4.3)</b>	<b>(4.97, 4.39)</b>	<b>(4.9, 4.69)</b>	<b>(4.96, 4.38)</b>	<b>(4.96, 4.37)</b>	<b>(4.99, 4.22)</b>
understandability	(4.99, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	<b>(4.99, 4.59)</b>	<b>(5.0, 4.79)</b>	<b>(4.93, 4.77)</b>	<b>(4.95, 4.47)</b>	<b>(4.99, 4.61)</b>	<b>(4.99, 4.39)</b>
fluency	(4.99, 4.99)	(5.0, 5.0)	(5.0, 5.0)	(5.0, 5.0)	<b>(5.0, 5.0)</b>	(4.99, 4.97)	<b>(5.0, 4.75)</b>	<b>(5.0, 4.86)</b>	<b>(4.94, 4.88)</b>	<b>(4.98, 4.66)</b>	<b>(4.99, 4.7)</b>	<b>(4.99, 4.54)</b>
Task-oriented static conversations, turn-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	amoyance	boredom	confusion	disapp.	frustration	skepticism
interestingness	<b>(3.66, 3.8)</b>	<b>(4.08, 4.05)</b>	<b>(3.39, 3.75)</b>	<b>(3.08, 3.34)</b>	<b>(3.32, 3.59)</b>	(3.2, 3.57)	(2.78, 2.6)	(2.47, 2.1)	(3.13, 3.36)	(2.77, 2.52)	(2.72, 2.46)	(3.0, 2.61)
engagingness	<b>(4.54, 4.43)</b>	<b>(4.68, 4.53)</b>	<b>(4.51, 4.61)</b>	<b>(4.38, 4.46)</b>	<b>(4.46, 4.52)</b>	<b>(4.45, 4.55)</b>	(4.16, 3.55)	(3.96, 2.83)	(4.48, 4.12)	(4.05, 3.36)	(4.06, 3.39)	(4.17, 3.28)
specificity	<b>(4.64, 4.7)</b>	<b>(4.87, 4.81)</b>	(4.84, 4.84)	<b>(4.72, 4.76)</b>	(4.76, 4.76)	<b>(4.8, 4.83)</b>	(4.69, 4.24)	(4.67, 4.45)	(4.78, 4.48)	(4.64, 4.1)	(4.62, 4.18)	(4.65, 3.96)
relevance	(4.94, 4.94)	<b>(4.98, 4.95)</b>	(4.99, 4.99)	(4.95, 4.95)	(4.96, 4.97)	(4.98, 4.98)	<b>(4.97, 4.56)</b>	<b>(4.96, 4.8)</b>	<b>(4.98, 4.75)</b>	<b>(4.97, 4.47)</b>	<b>(4.97, 4.52)</b>	<b>(4.92, 4.15)</b>
correctness	<b>(4.77, 4.87)</b>	<b>(4.9, 4.86)</b>	<b>(4.82, 4.9)</b>	<b>(4.72, 4.85)</b>	<b>(4.79, 4.88)</b>	<b>(4.8, 4.9)</b>	<b>(4.74, 4.27)</b>	<b>(4.78, 4.7)</b>	<b>(4.91, 4.56)</b>	<b>(4.81, 4.25)</b>	<b>(4.7, 4.17)</b>	<b>(4.47, 3.79)</b>
sem. appropri.	(4.97, 4.97)	<b>(4.98, 4.86)</b>	(4.98, 4.99)	(4.96, 4.97)	(4.97, 4.98)	(4.98, 4.99)	<b>(4.96, 4.52)</b>	<b>(4.97, 4.83)</b>	<b>(4.98, 4.75)</b>	<b>(4.97, 4.45)</b>	<b>(4.96, 4.46)</b>	<b>(4.9, 3.99)</b>
understandability	(4.99, 4.99)	<b>(4.99, 4.97)</b>	(5.0, 5.0)	(4.98, 4.99)	(4.98, 4.98)	(5.0, 5.0)	<b>(4.99, 4.6)</b>	<b>(5.0, 4.89)</b>	<b>(4.99, 4.11)</b>	<b>(4.99, 4.59)</b>	<b>(4.98, 4.51)</b>	<b>(4.96, 4.06)</b>
fluency	<b>(5.0, 5.0)</b>	<b>(5.0, 4.99)</b>	(5.0, 5.0)	(4.99, 4.99)	(4.99, 4.99)	(5.0, 5.0)	<b>(5.0, 4.8)</b>	<b>(5.0, 4.91)</b>	<b>(4.99, 4.81)</b>	<b>(5.0, 4.79)</b>	<b>(4.99, 4.71)</b>	<b>(4.98, 4.4)</b>
Task-oriented dynamic conversations, turn-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	amoyance	boredom	confusion	disapp.	frustration	skepticism
interestingness	<b>(3.35, 3.51)</b>	(3.35, 3.36)	<b>(3.32, 3.7)</b>	<b>(2.89, 3.09)</b>	<b>(3.16, 3.43)</b>	<b>(2.83, 3.22)</b>	(2.57, 2.4)	<b>(2.43, 2.12)</b>	(2.55, 2.76)	(2.63, 2.4)	<b>(2.78, 2.49)</b>	<b>(2.8, 2.42)</b>
engagingness	<b>(4.42, 4.28)</b>	<b>(4.45, 4.2)</b>	<b>(4.48, 4.6)</b>	<b>(4.23, 4.28)</b>	(4.5, 4.5)	<b>(4.28, 4.38)</b>	<b>(3.86, 3.24)</b>	<b>(3.78, 2.7)</b>	(3.63, 3.53)	(4.02, 3.2)	<b>(4.13, 3.37)</b>	<b>(4.14, 3.2)</b>
specificity	<b>(4.56, 4.62)</b>	<b>(4.82, 4.7)</b>	(4.76, 4.78)	(4.75, 4.78)	<b>(4.76, 4.78)</b>	<b>(4.73, 4.78)</b>	<b>(4.67, 4.21)</b>	<b>(4.72, 4.48)</b>	(3.96, 3.73)	<b>(4.95, 4.38)</b>	<b>(4.69, 4.18)</b>	<b>(4.67, 3.82)</b>
relevance	(4.88, 4.85)	<b>(5.0, 4.98)</b>	(4.95, 4.96)	(4.94, 4.94)	(4.94, 4.94)	(4.95, 4.94)	<b>(4.94, 4.51)</b>	<b>(5.0, 4.82)</b>	<b>(4.45, 4.07)</b>	(5.0, 4.72)	<b>(4.98, 4.46)</b>	<b>(4.98, 4.06)</b>
correctness	(4.78, 4.82)	<b>(4.92, 4.85)</b>	<b>(4.83, 4.88)</b>	<b>(4.53, 4.73)</b>	<b>(4.69, 4.79)</b>	<b>(4.78, 4.84)</b>	<b>(4.74, 4.19)</b>	<b>(4.9, 4.74)</b>	<b>(4.34, 3.86)</b>	(4.11, 4.4)	<b>(4.74, 4.16)</b>	<b>(4.78, 3.78)</b>
sem. appropri.	(4.96, 4.93)	<b>(5.0, 4.99)</b>	<b>(4.96, 4.98)</b>	(4.91, 4.96)	<b>(4.94, 4.96)</b>	(4.97, 4.97)	<b>(4.9, 4.45)</b>	<b>(5.0, 4.83)</b>	(4.49, 4.2)	(5.0, 4.65)	<b>(4.96, 4.48)</b>	<b>(4.98, 3.95)</b>
understandability	(4.97, 4.98)	<b>(4.99, 4.98)</b>	(4.99, 4.99)	(4.99, 4.99)	(4.99, 4.99)	(5.0, 5.0)	<b>(4.98, 4.58)</b>	<b>(5.0, 4.9)</b>	<b>(4.56, 3.61)</b>	(5.0, 4.73)	<b>(4.99, 4.49)</b>	<b>(4.99, 4.02)</b>
fluency	(4.99, 4.99)	(5.0, 5.0)	(4.99, 4.99)	(5.0, 4.99)	(4.99, 4.99)	(5.0, 5.0)	<b>(4.99, 4.78)</b>	<b>(5.0, 4.88)</b>	(4.57, 4.36)	(4.99, 4.86)	<b>(5.0, 4.77)</b>	<b>(5.0, 4.31)</b>

Table 4.5: Ratings for conversations for (no attitude, attitude) on GPT-4o for turn-level metrics. Pairs of numbers in bold represent statistically significant differences when a paired t-test is conducted between no attitude ratings and attitude ratings for the same metric on the same set of conversations.

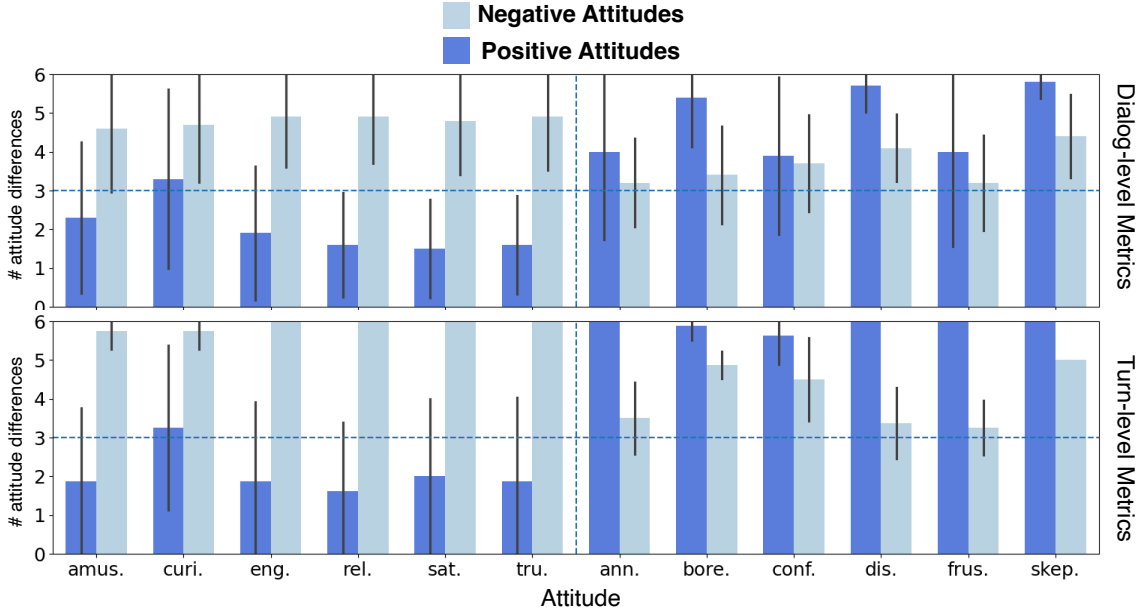


Figure 4.3: All attitudes rating the same set of task-oriented conversations over all dialog- and turn-level metrics. The number of attitudes (positive or negative) that have significantly different ratings from the current attitude for each metric are counted. Each number represented in the graph corresponds to the counts for one of the metrics. The darker colored bars represent comparisons between that attitude and all positive attitudes, while the lighter colored bars represent comparisons between that attitude and all negative attitudes. The x-axis represents each attitude, and the y-axis represents how many metric comparisons were statistically significant.

attitude and negative attitudes on a task [98]. Figure 4.4 shows the results of a paired t-test between the attitude and no attitude ratings for the coherence and interestingness metrics for static open-domain conversations. Full results are found in Tables 4.4 and 4.5.

**Coherence.** In Figure 4.4, all positive attitudes have little to no difference in ratings between the no attitude and attitude ratings. Both the types ratings average to around 5 as seen in Table 4.4. This means that it may not be necessary to perform personalized evaluation on conversations had with positive attitudes for the coherence metric.

However, all the negative attitudes have significant differences in the ratings between the no attitude and attitude ratings. On average, the no attitude ratings are high, but not as high as for the positive attitudes (averaging 4.66 - 5). The attitude ratings are around 0.62 lower. Since the ratings are still high, for a deployed system, it may be enough to know that simulated users with negative attitudes still believe the system is reasonably coherent.

These results are consistent with the expectation of GPT-4o to output generally coherent responses. As seen in Table 4.2, coherence also saturates around 5 for all positive attitude ratings on the same conversation. There is some separation between the nega-

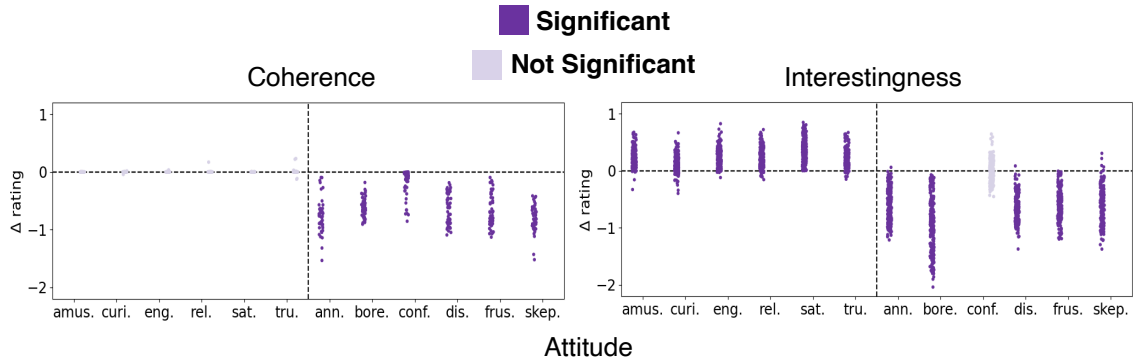


Figure 4.4: Differences between the ratings of the no attitude and attitude ratings calculated pairwise across all conversations for open-domain conversations in the static case for coherence and interestingness. The attitudes which have no significant differences between the no attitude ratings and attitude ratings when a paired t-test is performed have data points indicated in the lighter color, while attitudes with significant differences are indicated in the darker color.

tive and positive attitudes, but not every negative attitude is different from all positive attitudes and from each other. This reinforces that the coherence metric may not have as many benefits from personalized evaluation.

**Interestingness.** In Figure 4.4, there are more significant differences across the attitudes. All positive attitude ratings are significantly different than the no attitude ratings. This difference is slight, with the average differences all between 0.09 and 0.39 seen in Table 4.5. However, the values for both attitude and no attitude ratings do not saturate at 5 with averages from 3.82 to 4.54. Most conversations, when rated by the positive attitude, will be rated higher than when rated by no attitude. This could suggest that since the positive attitudes have favorable views toward the dialog system, that they are more likely to find the same sentence more interesting.

There are still significant differences in all negative attitudes except for confusion. Most of these differences are similar to the difference seen in the coherence metric ratings, but the span of average ratings are lower - from an average of 2.24 to 3.69. Negative attitudes tend to rate conversations lower than the no attitude ratings. This suggests that the negative attitudes used during the rating affects the ratings given by finding the same sentence less interesting. In fact, there are no instances in the turn-level metrics where the average of the negative attitude ratings is higher than the average of the no attitude ratings. Therefore, although a sentence may be interesting according to a generic automatic evaluation, user attitudes should be considered for a more accurate assessment of system performance.

Referring to Table 4.2, there are significant differences between each positive attitude rating and all negative attitude ratings, whereas half of all negative attitude ratings are significantly different from each other, while all negative attitude ratings are significantly different from the positive attitude ratings. The average of these metrics also do not saturate around 5 and have a range between 2.55 and 4.16. This suggests that interestingness

is a metric that can benefit from personalized metrics.

The trends seen in the static conversations are similar to those in the dynamic conversations, which are expanded upon in Section 4.2.5. Llama-2-70b-chat-hf is also used to perform the same rating experiments on dialog-level metrics for static conversations in Section 4.2.5, and exhibits similar trends as GPT-4o.

## All Metrics

There are two groups to note in the metrics discussed thus far - those that have significant differences over all negative attitudes and not any positive attitudes, and those that have significant differences across all attitudes. There are no metrics with no significant differences over any attitude, and no metrics have significant differences only over all positive attitudes. This discussion focuses on the trends in the static conversations.

Only error recovery has significant differences over all metrics at the dialog-level for both open-domain and task-oriented conversations. Tables 4.4 and 4.5 reveal that metrics that tend to saturate at 5 - coherence, consistency, likeability, and understandingness will have significant differences among all the negative metrics, and none or one of the positive metrics. Metrics that have large amount of significant differences (more than 8) for open-domain conversations such as error recovery, diversity, flexibility, informativeness, and inquisitiveness have slight more or the same significant differences than task-oriented conversations. This may be explained by these metrics being more well-defined in task-oriented conversations, which could cause the metrics to be more similar to the no attitude ratings. This is most apparent for error recovery, as 5 attitude ratings become non-significant.

It is surprising that correctness is significant across all negative attitudes and one positive attitude for open-domain conversations on turn-level metrics, but significant across all attitudes for task-oriented conversations. Looking at the numbers in Table 4.5, it seems that correctness saturates around 5 for most no attitude and attitude ratings in the open-domain conversations, but has slightly lower values, although still very high, for the task-oriented conversations. There could be a potential for the system to make more mistakes for task-oriented systems, or for the simulated user to have slightly more variety in ratings, as the definition corresponds to helping the user complete their task correctly, while the the definition for the open-domain systems is more related to contextual correctness. Semantic appropriateness and fluency, while only fully significant across the negative attitudes for open-domain conversations also saturate around 5 for most no attitude and attitude ratings. These two metrics do not seem like they would have variations based on the attitude that is rating, so it is interesting to see these results.

Engagingness is the only metric that is significant across all attitudes for both open-domain and task-oriented conversations. This is expected as engagingness (and interest-iness) would be considered to be more subjective compared to metrics such as fluency. Engagingness has a wide spread of average ratings, from 2.68 to 4.89, and drastic average differences, such as for open-domain conversations, the boredom attitude rates at an average of 2.68, while no attitude rates at an average of 4.09 on the same conversations. This indicates that engagingness is also a metric that should be considered on the personalization level.

## Correlation with User Satisfaction

Seen through Sections 4.2.5 and 4.2.5, there is a potential to create personalized metrics for negative attitudes since 1) negative attitude ratings are more distinguishable from each other and positive attitude ratings which shows the potential for personalization, 2) there are metrics that exhibit high ICC which could be used for personalized metrics, and 3) there are many metrics which exhibit differences between not using an attitude to rate and using a negative attitude to rate, which now accounts for the personalized user experience. With these insights, there are several simple ways that can be explored to create this personalized metric. The following experiments are on the open-domain static conversations. Decision tree regressors (from [137]) are used instead of linear regression to reduce the impact of correlated features. The metrics are correlated against user satisfaction rated from 1-5 by GPT-4o, where it is prompted a definition of user satisfaction (found in Appendix 5.8), where 9 out of 18 metrics are mentioned in the definition. User satisfaction is rated with an attitude as well. Turn-level metrics are averaged across each conversation.

User satisfaction is rated with and without an attitude. Although there are 2 significant differences in positive attitudes (satisfaction and trust) between these two ratings, the average user satisfaction rating is 5. On negative attitudes, there are significantly lower ratings for user satisfaction rated with an attitude, ranging from 1.75 to 3.77, compared to without an attitude, ranging from 3.31 to 4.91. Therefore, it is important to consider user satisfaction with an attitude for automatic rating as well. The following experiments are only calculated on the negative attitude ratings.

Metrics with high ICC do not necessarily have high positive correlation with user satisfaction when annotated with an attitude. Likeability, which had good ICC, has above 0.76 correlation with user satisfaction for annoyance, confusion, disappointment, frustration, and skepticism, while most metrics have low correlation with user satisfaction for boredom, the highest being topic depth with 0.40. For metrics with high ICC, topic depth and informativeness have correlations above 0.36 for all negative attitudes, while inquisitiveness, interestingness, engagingness, and specificity do not. Therefore, even if a metric can be distinguished between attitudes, it may not correlate with user satisfaction.

Decision trees return the feature importance and consider every dialog-level and turn-level metric to calculate which metrics are important to user satisfaction. Likeability has the highest Gini importance for annoyance, disappointment, and skepticism, while inquisitiveness is the highest for boredom, coherence is the highest for confusion, and flexibility is the highest for frustration. Likeability seems to be a large predictor of user satisfaction, and since it also has good ICC, personalized aggregated metrics can start to be built off this metric. Prior work has also shown likeability to be the highest indicator of the overall score out of the dialog-level metrics [120].

## Personalization for Dynamic Conversations

The trends seen in the static conversations are similar to those in the dynamic conversations. The results from the static conversation evaluation are visualized because certain attitudes are not chosen during the conversations, seen in Table 4.2. There are settings

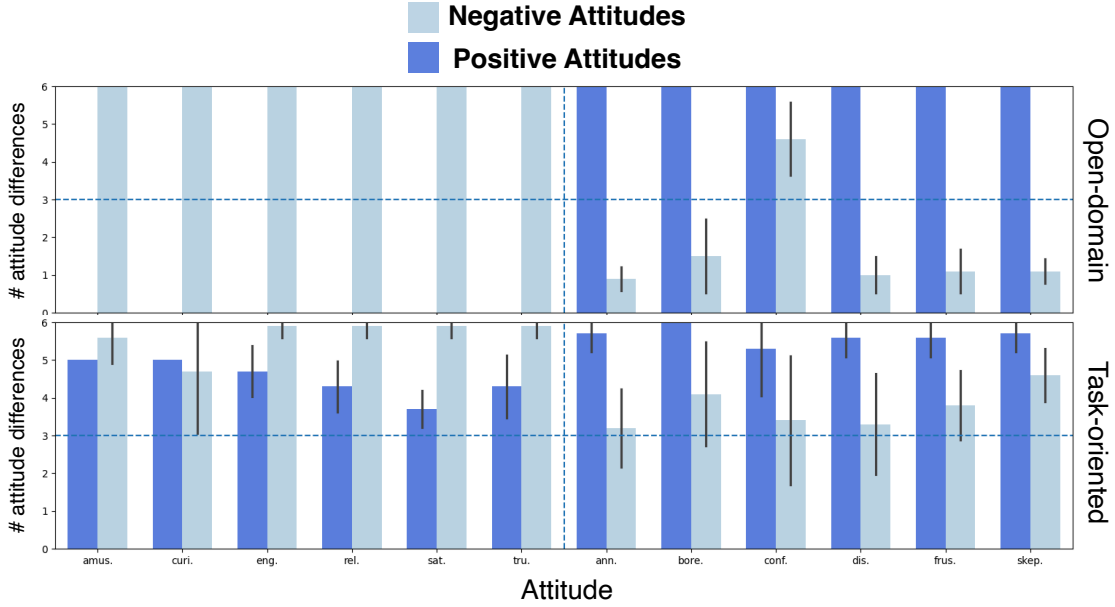


Figure 4.5: All attitudes rating the same set of open-domain and task-oriented conversations over all dialog-level metrics on Llama. The number of attitudes (positive or negative) that have significantly different ratings from the current attitude for each metric are counted. Each number represented in the graph corresponds to the counts for one of the metrics. The darker colored bars represent comparisons between that attitude and all positive attitudes, while the lighter colored bars represent comparisons between that attitude and all negative attitudes. The x-axis represents each attitude, and the y-axis represents how many metric comparisons were statistically significant.

where the trend has more apparent differences - informativeness across open-domain conversations, and semantic appropriateness across task-oriented conversations, both for negative attitudes. For informativeness, there is one significant difference between attitude and no attitude ratings for the dynamic conversations for the positive attitudes, whereas there are significant differences in 4 metrics in the static conversations for both types of conversations. With more data, the findings from the static conversations will likely stay similar for the dynamic conversations.

## Llama

Previous work has noted Llama [209] has had issues with retaining persona throughout conversation, and this is taken into consideration when analyzing these results. However, the persona is only retained across the one turn where the model rates a conversation, so the degradation in persona may be lessened.

Llama annotations exhibit similar trends as GPT-4o annotations to an even greater extent. However, a major difference is that Llama ratings, although they do not have as high of a saturation in metric values as GPT-4o, have a lower range of average metric

	amusement		curiosity		engag.		relief		satis.		trust		annoyance		boredom		confusion		disapp.		frustration		skepticism		
	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	
open-domain dialog																									
coherence	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	1	6	<b>5*</b>	6	1	6	1	6	1	
error recovery	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	2	6	4	6	1	6	1	6	1	
consistency	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	1	6	<b>5*</b>	6	1	6	1	6	1	
diversity	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	2	6	<b>5*</b>	6	1	6	2	6	1	
topic depth	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	4	6	<b>5*</b>	6	2	6	2	6	2	
likeability	0	6	0	6	0	6	0	6	0	6	0	6	6	0	6	1	6	2	6	0	6	0	6	1	
understandingness	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	1	6	<b>5*</b>	6	1	6	1	6	1	
flexibility	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	1	6	<b>5*</b>	6	1	6	1	6	1	
informativeness	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	1	6	<b>5*</b>	6	1	6	1	6	1	
inquisitiveness	0	6	0	6	0	6	0	6	0	6	0	6	6	1	6	1	6	<b>5*</b>	6	1	6	1	6	1	
task-oriented dialog																									
coherence	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	4	6	4	6	<b>5*</b>	6	6	2	6	3	6	1	6	2	6	2	6	4	
error recovery	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	4	6	4	6	<b>5*</b>	6	6	3	6	3	6	<b>5*</b>	6	2	6	4	6	<b>5*</b>	
consistency	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	4	6	4	6	<b>5*</b>	6	6	2	6	1	6	2	6	1	6	3	6	<b>5*</b>	
diversity	<b>5*</b>	5	<b>5*</b>	5	<b>5*</b>	6	<b>5*</b>	6	4	6	4	6	6	<b>5*</b>	6	<b>5*</b>	5	<b>5*</b>	5	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	
topic depth	<b>5*</b>	4	<b>5*</b>	5	4	5	<b>5*</b>	5	3	5	4	5	5	4	6	<b>5*</b>	2	<b>5*</b>	6	<b>5*</b>	5	4	5	<b>5*</b>	
likeability	<b>5*</b>	6	<b>5*</b>	5	<b>5*</b>	6	3	6	3	6	3	6	6	2	6	<b>5*</b>	6	2	6	3	5	3	6	<b>5*</b>	
understandingness	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	<b>5*</b>	6	4	6	4	6	6	3	6	4	6	2	6	4	6	4	6	3	
flexibility	<b>5*</b>	6	<b>5*</b>	3	<b>5*</b>	6	4	6	4	6	<b>5*</b>	6	5	4	6	<b>5*</b>	6	<b>5*</b>	5	4	5	<b>5*</b>	6	<b>5*</b>	
informativeness	<b>5*</b>	6	<b>5*</b>	1	<b>5*</b>	6	4	6	4	6	<b>5*</b>	6	5	3	6	<b>5*</b>	5	2	5	3	5	4	5	<b>5*</b>	
inquisitiveness	<b>5*</b>	5	<b>5*</b>	4	3	6	<b>5*</b>	6	3	6	3	6	6	4	6	<b>5*</b>	5	<b>5*</b>	5	4	6	4	5	4	

Table 4.6: Statistical significance for attitudes rating the same conversations, created with no attitude indicators and rated by Llama2 70b. The positive column represents significant differences between the given attitude’s ratings and other positive attitudes’ ratings for each metric. The negative column represents significant differences between the given attitude’s ratings and other negative attitudes’ ratings for each metric. Bold and starred entries represent entries that deviate greatly from the observation that positive attitudes are different from all other negative attitudes and vice versa. The two experimental conditions are outlined on the left, along with the list of metrics under each condition.

	positive attitudes							negative attitudes				
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	4.93-	4.9-	4.93-	4.91-	4.94-	4.92-	4.05+	4.05+	4.38+	4.08+	4.16+	4.05+
error recovery	4.75-	4.7-	4.72-	4.73-	4.78-	4.74-	4.05+	3.99+	4.24+	4.05+	4.1+	4.05+
consistency	4.88-	4.85-	4.87-	4.86-	4.88-	4.87-	4.01+	4.03+	4.33+	4.04+	4.11+	4.03+
diversity	4.73-	4.71-	4.72-	4.63-	4.68-	4.6-	4.01+	3.93+	4.22+	4.01+	4.07+	4.02+
topic depth	4.79-	4.84-	4.81-	4.75-	4.8-	4.76-	3.96+	3.83+	4.29+	4.04+	4.09+	4.02+
likeability	4.77-	4.73-	4.74-	4.72-	4.75-	4.67-	3.95+	3.9+	4.1+	3.97+	4.01+	3.98+
understandingness	4.91-	4.89-	4.91-	4.88-	4.92-	4.89-	4.06+	4.02+	4.37+	4.06+	4.13+	4.03+
flexibility	4.92-	4.91-	4.91-	4.89-	4.93-	4.89-	4.1+	4.05+	4.43+	4.07+	4.17+	4.06+
informativeness	4.89-	4.92-	4.91-	4.88-	4.91-	4.89-	4.09+	4.02+	4.5+	4.1+	4.19+	4.07+
inquisitiveness	4.86-	4.9-	4.89-	4.84-	4.89-	4.82-	4.01+	4.02+	4.36+	4.1+	4.17+	4.07+

Table 4.7: All attitudes rating neutral conversation dialog opendomain, + = significant across all positive, - = significant across all negative - Llama

	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	4.87+-	4.82+-	4.91+-	4.95-	4.95-	4.94+-	4.71+	4.64+	4.68+	4.72+	4.72+	4.62+
error recovery	4.74+-	4.64+-	4.78+-	4.89-	4.88-	4.86+-	4.42+	4.46+	4.52+-	4.41+	4.39+	4.33+-
consistency	4.87+-	4.82+-	4.91+-	4.95-	4.94-	4.94+-	4.73+	4.74+	4.72+	4.74+	4.76+	4.66+-
diversity	4.17+	4.06+	4.37+-	4.45+-	4.41-	4.42-	4.11+-	3.84+-	4.19-	4.08-	4.12+-	4.01+-
topic depth	3.79+	3.56+	3.89	3.97+	3.91	3.93	3.75	3.23+-	3.93-	3.66+-	3.74	3.57-
likeability	4.39+-	4.23+	4.61+-	4.64-	4.64-	4.64-	4.19+	4.0+-	4.17+	4.17+	4.2	4.1+-
understandingness	4.9+-	4.81+-	4.92+-	4.96+-	4.96-	4.96-	4.69+	4.58+	4.65+	4.73+	4.71+	4.67+
flexibility	4.51+-	4.31+	4.68+-	4.78-	4.77-	4.75+-	4.3	4.12+-	4.41+-	4.33	4.36-	4.26+-
informativeness	4.87+-	4.68+	4.9+-	4.94-	4.94-	4.93+-	4.71	4.55+-	4.72	4.7	4.74	4.65-
inquisitiveness	4.23+	4.02+	4.34-	4.38+-	4.33-	4.33-	4.11+	3.86+-	4.17-	4.07	4.1+	4.04

Table 4.8: All attitudes rating neutral conversation turn opendomain, + = significant across all positive, - = significant across all negative - Llama

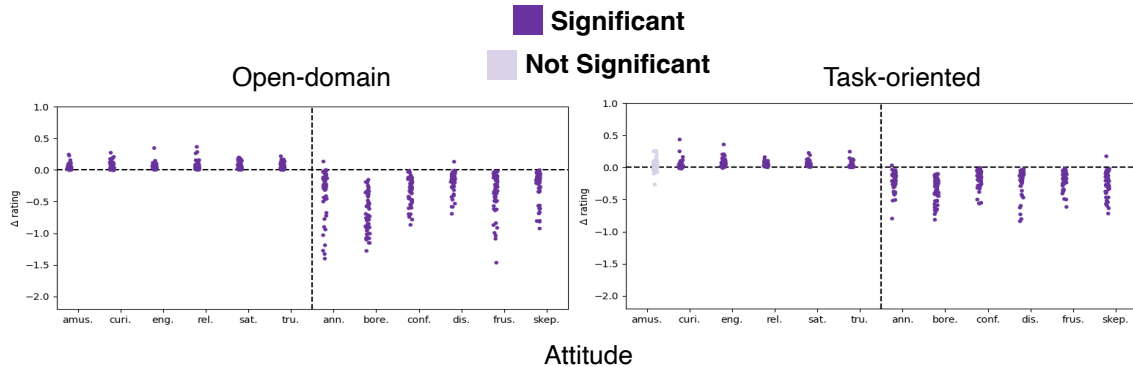


Figure 4.6: Differences between the ratings of the no attitude and attitude ratings calculated pairwise across all conversations for open-domain and task-oriented conversations in the static case for coherence on Llama. The attitudes which have no significant differences between the no attitude ratings and attitude ratings when a paired t-test is performed have data points indicated in the lighter color, while attitudes with significant differences are indicated in the darker color.

ratings. If an LLM is used for evaluation during the deployment of a dialog system for either of these tasks, the correlation between human evaluation and these two trends needs to be considered to choose the appropriate LLM that more closely models human evaluation.

The following experiments are conducted on the dialog-level metrics for open-domain and task-oriented conversations both created with no attitude and created with static attitudes.

**Attitudes rating the same conversation.** Following the same paired t-test and Bonferroni correction in Section 4.2.5, the visualization of the results is seen in Figure 4.5. The trends seen in Section 4.2.5 are even more apparent here - in Figure 4.6, positive attitude ratings are not distinguishable from each other, and fully distinguishable across all negative metrics. Ratings from Llama do not have as high of a saturation in metric values as GPT-4o, and have a lower range of average metric ratings, with the average ratings mostly between 4 and 5. The largest range for any metric and attitude pair for GPT-4o ratings is 3.90 and the smallest range is 0.01, while the average range across all open-domain and task-oriented conversations is 1.59. The largest range for any metric and attitude pair for Llama ratings is 2.80 and the smallest range is 0.22, while the average range is 0.95. Therefore, although GPT-4o shows less significant differences between attitudes, it has a wider range of ratings, while Llama shows more significant differences between attitudes, but has a smaller range of ratings.

**Attitudes rating different conversations.** Almost all metrics across all conditions show significant differences between the attitude and no attitude ratings for the majority of attitudes. It seems that for some metrics that were not statistically different for GPT-4o ratings, namely coherence (which is shown in Figure 4.6, consistency, likeability, and understandingness for positive attitudes on both open-domain and task-oriented conversations, Llama ratings provide slightly more separation in many cases between no attitude

Open-domain static conversations, dialog-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	(4.93, 4.97)	(4.05, 3.7)	(4.05, 3.48)	(4.38, 4.02)	(4.9, 4.96)	(4.08, 3.9)	(4.93, 4.98)	(4.16, 3.88)	(4.91, 4.97)	(4.94, 4.99)	(4.05, 3.79)	(4.92, 4.98)
error recovery	(4.75, 4.84)	(4.05, 3.76)	(3.99, 3.36)	(4.24, 4.02)	(4.7, 4.82)	(4.05, 3.91)	(4.72, 4.85)	(4.1, 3.86)	(4.73, 4.87)	(4.78, 4.9)	(4.05, 3.83)	(4.74, 4.87)
consistency	(4.88, 4.95)	(4.01, 3.72)	(4.03, 3.56)	(4.33, 4.01)	(4.85, 4.93)	(4.04, 3.87)	(4.87, 4.95)	(4.11, 3.87)	(4.86, 4.96)	(4.88, 4.98)	(4.03, 3.8)	(4.87, 4.96)
diversity	(4.73, 4.83)	(4.01, 3.64)	(3.93, 3.15)	(4.22, 4.04)	(4.71, 4.84)	(4.01, 3.87)	(4.72, 4.84)	(4.07, 3.81)	(4.63, 4.88)	(4.68, 4.84)	(4.02, 3.77)	(4.6, 4.78)
topic depth	(4.79, 4.91)	(3.96, 3.63)	(3.83, 3.2)	(4.29, 4.1)	(4.84, 4.93)	(4.04, 3.93)	(4.81, 4.93)	(4.09, 3.85)	(4.75, 4.92)	(4.8, 4.95)	(4.02, 3.76)	(4.76, 4.92)
likeability	(4.77, 4.95)	(3.95, 3.43)	(3.9, 2.98)	(4.1, 3.99)	(4.73, 4.93)	(3.97, 3.8)	(4.74, 4.96)	(4.01, 3.69)	(4.72, 4.94)	(4.75, 4.97)	(3.98, 3.62)	(4.67, 4.94)
understandingness	(4.91, 4.96)	(4.06, 3.5)	(4.02, 3.25)	(4.37, 4.01)	(4.89, 4.94)	(4.06, 3.88)	(4.91, 4.96)	(4.13, 3.79)	(4.88, 4.96)	(4.92, 4.98)	(4.03, 3.73)	(4.89, 4.96)
flexibility	(4.92, 4.94)	(4.1, 3.71)	(4.05, 3.41)	(4.43, 4.09)	(4.91, 4.95)	(4.07, 3.91)	(4.91, 4.95)	(4.17, 3.9)	(4.89, 4.96)	(4.93, 4.97)	(4.06, 3.8)	(4.89, 4.95)
informativeness	(4.89, 4.94)	(4.09, 3.64)	(4.02, 3.3)	(4.5, 4.18)	(4.92, 4.97)	(4.1, 3.91)	(4.91, 4.96)	(4.19, 3.85)	(4.88, 4.96)	(4.91, 4.97)	(4.07, 3.68)	(4.89, 4.96)
inquisitiveness	(4.86, 4.91)	(4.01, 3.54)	(4.02, 3.21)	(4.36, 4.09)	(4.9, 4.95)	(4.1, 3.93)	(4.89, 4.93)	(4.17, 3.85)	(4.84, 4.91)	(4.89, 4.94)	(4.07, 3.78)	(4.82, 4.91)

Task-oriented static conversations, dialog-level metrics												
	positive attitudes						negative attitudes					
	amusement	curiosity	engag.	relief	satis.	trust	annoyance	boredom	confusion	disapp.	frustration	skepticism
coherence	(4.8, 4.81)	(4.22, 4.01)	(4.36, 4.0)	(4.15, 3.98)	(4.9, 4.94)	(4.07, 3.86)	(4.89, 4.96)	(4.17, 3.99)	(4.95, 4.99)	(4.94, 4.99)	(4.19, 3.94)	(4.95, 4.98)
error recovery	(4.79, 4.73)	(4.23, 4.0)	(4.26, 3.8)	(4.25, 4.03)	(4.85, 4.87)	(4.11, 3.85)	(4.88, 4.91)	(4.19, 3.98)	(4.95, 4.98)	(4.92, 4.97)	(4.25, 3.97)	(4.96, 4.98)
consistency	(4.8, 4.78)	(4.26, 4.06)	(4.4, 4.2)	(4.21, 4.01)	(4.89, 4.92)	(4.11, 3.91)	(4.88, 4.95)	(4.21, 4.05)	(4.94, 4.99)	(4.93, 4.99)	(4.25, 4.0)	(4.95, 4.98)
diversity	(4.55, 4.64)	(3.99, 3.81)	(3.84, 3.15)	(4.12, 3.97)	(4.83, 4.87)	(3.93, 3.66)	(4.58, 4.78)	(3.99, 3.84)	(4.73, 4.92)	(4.67, 4.89)	(4.11, 3.81)	(4.74, 4.9)
topic depth	(4.1, 4.4)	(3.43, 3.52)	(2.87, 2.56)	(4.08, 3.96)	(4.9, 4.93)	(3.66, 3.55)	(4.19, 4.56)	(3.57, 3.63)	(4.24, 4.73)	(4.2, 4.74)	(3.82, 3.69)	(4.27, 4.7)
likeability	(4.48, 4.74)	(3.97, 3.77)	(3.97, 3.39)	(4.03, 3.97)	(4.75, 4.89)	(3.92, 3.64)	(4.6, 4.92)	(3.97, 3.79)	(4.75, 4.96)	(4.74, 4.97)	(4.0, 3.65)	(4.74, 4.96)
understandingness	(4.89, 4.88)	(4.3, 3.98)	(4.39, 3.95)	(4.2, 3.95)	(4.93, 4.95)	(4.12, 3.84)	(4.94, 4.97)	(4.25, 3.96)	(4.97, 4.99)	(4.96, 4.99)	(4.25, 3.89)	(4.97, 4.99)
flexibility	(4.79, 4.76)	(4.13, 3.92)	(4.12, 3.72)	(4.24, 4.02)	(4.9, 4.91)	(4.05, 3.82)	(4.84, 4.92)	(4.09, 3.94)	(4.93, 4.98)	(4.91, 4.98)	(4.18, 3.98)	(4.92, 4.97)
informativeness	(4.8, 4.78)	(4.22, 4.01)	(4.23, 3.89)	(4.23, 4.03)	(4.93, 4.96)	(4.06, 3.85)	(4.9, 4.97)	(4.17, 4.0)	(4.95, 4.99)	(4.93, 4.99)	(4.22, 3.96)	(4.96, 4.99)
inquisitiveness	(4.51, 4.56)	(3.94, 3.72)	(3.77, 3.11)	(4.22, 4.06)	(4.93, 4.94)	(3.95, 3.72)	(4.64, 4.77)	(3.97, 3.8)	(4.68, 4.85)	(4.66, 4.82)	(4.13, 3.91)	(4.73, 4.83)

Table 4.9: Ratings for conversations for (no attitude, attitude) on Llama2 70b annotations. Pairs of numbers in bold represent statistically significant differences when a paired t-test is conducted between no attitude ratings and attitude ratings for the same metric on the same set of conversations.

and attitude ratings. However, this separation may not be necessary because some metrics are expected to saturate for positive attitudes as seen in Section 4.2.5.

## 4.2.6 Conclusion

Throughout this work, there are several axes presented for the study of the sensitivity of automatic dialog evaluation for personalized evaluation: the model used for automatic evaluation (GPT-4o/Llama), static/dynamic conversations, positive/negative attitudes, and ratings informed by an attitude/ratings not informed by an attitude.

There are differences in user-simulated attitude ratings when using either GPT-4o or Llama to rate conversations, where there is a clearer divide between positive attitudes and negative attitudes than within each group of attitudes in Section 4.2.5. Static and dynamic conversations exhibit similar trends in different capacities, but there still remain issues of models choosing positive attitudes when given a choice and reducing the number of conversations with negative attitudes for generating data for research in Section 4.2.4.

There is less separation between ratings of attitudes with the same sentiment than attitudes with opposing sentiments on the same conversation, which suggests that it may be useful to start personalizing a metric by a user’s sentiment. There are also pronounced differences between ratings informed by attitudes and ratings not informed by attitudes in Sections 4.2.5 and 4.2.5. If metrics are not personalized to a certain degree with a user’s information, system performance may not be captured holistically. With these two findings, using traditional prompts for automatic dialog evaluation that do not consider attitude may not be appropriate for certain dialog evaluations. This is now something important to consider when continuing to use models for dialog evaluation moving forward. These personalized metrics can also begin to be a proxy for user satisfaction seen in Section 4.2.5.

Future work could include studying the sensitivity for more types of use cases for assistants and specialized models for certain use cases. Although automatic evaluation may suffer from potential misalignment with humans, it opens the door to not only be a future foundation for possible human studies, but also an alternative that is safe and efficient.



# Chapter 5

## Personalized Evaluation in Real World Systems

### 5.1 Introduction

Through the last two chapters, we have seen how autoraters can align with humans in a closed setting as well as simulate human attitudes for personalized evaluation. In this chapter, we collect a dataset and perform subsequent analysis to further understand personalized evaluation in the real world. This dataset will have natural interactions between humans and personalized, not personalized, and wrongly personalized chatbots through a range of tasks and human preferences. We will study if autoraters injected with human preferences can achieve high correlation with first-person ratings from the user such that autoraters can be used as a better estimate of system performance in the real world. However, if we use autoraters that are injected with human preferences to estimate system performance in the real world, the privacy and usage of human data becomes a concern.

### 5.2 Personalized Evaluation in Natural Interaction

#### 5.2.1 Motivation

Personalized evaluation, used in the context of dialog in this thesis, is the act of injecting contextual information about the user into an autorater, or providing this information to an external human rater, to rate a user’s system interaction. The purpose of personalized evaluation is to obtain higher agreement with the “true” ratings provided by the user, and to provide developers with more accurate autoraters to assess their dialog systems. Personalized evaluation can be used for any interactions for both generic and personalized systems.

It is important to ensure that the contextual information gathered from the user contributes to a statistical increase in task performance - Finck and Biega noted that this is not necessarily practiced in companies. We might expect that the more we know about a user, the better a task can be performed, as seen in Padmanabhan et al. using browser

history to determine user spending. However, sometimes, using personal data may not improve model predictions for certain groups of people theoretically [124] and empirically [166], which would cause more harm toward users. Other work has introduced “participatory systems” [85], in which users can choose to share personal data for classification tasks to determine whether that would improve the predictions by the model.

Related issues with data collection include the legality of “excessive data collection”, [148], and its risks [101]. Europe already has privacy rules regarding “purpose limitation” (limiting data collected to that which has a “legitimate purpose”) and “data minimisation” (only collecting data that is needed for the legitimate purpose) [1, 47]. To combat this, some work has looked at optimizing how much data is necessary to complete tasks [113]. There are also issues with companies collecting and using users’ sensitive data [164], as well as users not knowing their data is being collected [57, 126].

Our dataset and subsequent analysis, which studies personalized evaluation, will show quantitatively whether collecting user preference data will improve system evaluation for developers, and therefore whether this data should be collected.

Users have different goals, prior assumptions, and prior experiences in using dialogue systems [52], which can affect how they use or engage with these systems in the future. Understanding user preferences [83] could lead to improved user experience; thus, current work creates and uses representations of user profiles to customize system outputs [100, 193]. The system’s responses are then evaluated by the users themselves or an external human annotator. There has been some work on emotion annotation in a debate study with all three perspectives which shows that emotion perception varies [134]; however, rarely are system responses rated by both the users and external human annotators.

It is important to understand in a statistical context how different users rate from external human annotators. Evaluation of systems now mostly consist of human raters who have no knowledge of the user except for any explicit information that can be obtained from the conversation itself. It has been shown that contextual information impacts evaluation [43, 114]; however, human annotators may also introduce inter-rater disagreement depending on the evaluation criteria. Issues of systematic variance (rater bias for reasons such as societal norms) or random variance (for example measurement discrepancies) [76] are also present that can affect human ratings. In our work, we look toward minimizing random variance, while systematic variance is less of a concern because we concern ourselves with what is directly observable. The human annotations may also not be the “true” rating for the user interactions - as mentioned in Section 2.2.1. There is a shift toward using autoraters for this task, and it has been shown that correlating autoraters to non-fully agreeing human ratings performs statistically worse than to fully agreeing human ratings seen in Section 3.3.

Therefore, understanding the differences between user and external human annotations on the same task is important. If human ratings are not statistically different than internal user ratings, then we can confidently offload annotation to human raters. If human ratings are statistically different than user ratings, then to accurately evaluate systems, there should be more of a reliance on user ratings. However, the load on the user while they are a conversational participant tends to be great, and evaluation is typically constrained to thumbs up and down (seen in ChatGPT) or other minimally intrusive but broad signals

(like feedback questions during the conversation [42, 65]). Dialog adds in an additional interactive component as compared to other tasks like QA which could heighten these differences.

We hypothesize that user preferences can affect the evaluation itself, as a user who prefers chatbot empathy may choose to use an empathetic chatbot over a non-empathetic chatbot, even if its task performance is lower. By understanding a user’s preferences, whether through explicit prompting or implicit assumption, we can detect how much of their evaluation was influenced by their preferences. External human annotators may not have the same preferences as the user as mentioned in Section 2.2.1, and with many datasets [120] consisting of human annotations, the true system performance may not be captured. For example, a user who is feeling negatively about a system due to the system’s lack of empathy may rate differently than a human annotator who is purely looking at task performance. The system’s responses may also look to be performing well on objective metrics because they are fluent and contain the information the user asked for. Therefore, it would be important to know whether the ratings from the user are a consequence of their preferences or of the system’s actual performance.

In this work, we collect a dialog dataset which includes user, external human, and autorater annotations. Additionally, we vary the user’s preferences for systems across the empathy and sociability axes, the tasks, and the personalization of the system to determine if there are statistical differences. We determine that personalized evaluation provides improvement on the agreement between user and external human ratings, and between external human and autorater annotations, but that perception of system style still plays a big role in the efficacy for personalized evaluation.

## 5.2.2 Analytical Model

The analytical model of the world that contains our experiments consists of 4 factors: information about a user’s preferences, the two systems that the user are interacting with, the task that the user is using the system for, and the ratings collected for each conversation between the participant and each system. This world view will assist us in determining whether collecting user preferences for personalized evaluation will provide us with a more accurate view on system performance.

### User Preferences

User preferences can consist of stylistic preferences (ex. how much more information the system provides than that necessary to answer the user’s query [122]) or of task-specific preferences (ex. specifications for food quality for restaurant booking [7]). We will be focusing on stylistic preferences for this work. Study participants having the conversation with the systems will henceforth be referred to as “users”.

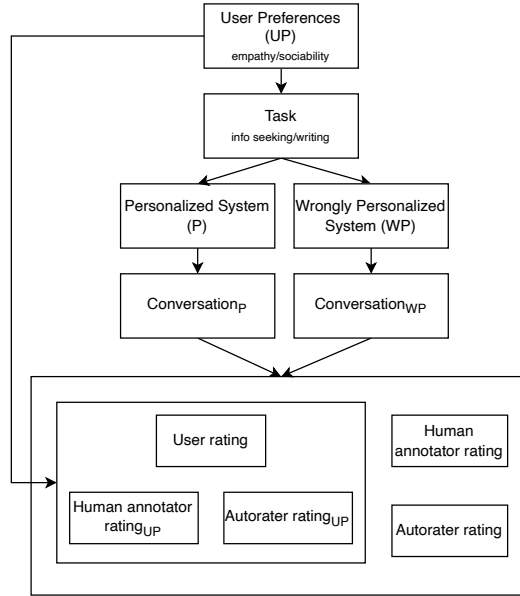


Figure 5.1: Analytical model of the study

## System

Users have many systems, similar to ChatGPT, that they can interact with. We define a system here as an agent with which a user can communicate with in natural language.

## Tasks

We define tasks as a goal that a user may have when interacting with a system. The user may have no goal as for chit-chat bots, or may have a specific goal such as figuring out what the weather is today for task-oriented bots. Depending on the task, users may interact with or value different aspects of the system more than others.

## Ratings

After a user has a conversation with both systems for some task, the conversations are rated by the user themselves, human annotators and autoraters without user preference knowledge, and human annotators and autoraters with user preference knowledge. If human annotators and autoraters without knowledge of the user preferences can within statistical bounds determine a system’s performance, then it is not necessary to collect and inject user preferences for personalized evaluation. Human annotators will be henceforth referred to as “human raters” with and without knowledge, and autoraters will be referred to as “autoraters” with and without knowledge.

### 5.2.3 Hypotheses

Through our dataset collection and subsequent analysis we aim to study our main hypothesis:

**Hypothesis.** Personalized evaluation, referenced in Section 5.2.1 as the act of providing contextual information about the user to a human rater or autorater, provides significant improvements on rating agreements for a dialog interaction as compared to when human raters or autoraters are not provided with contextual information about the user’s preferences.

We evaluate this through two different pieces of contextual information, user preferences for empathetic system responses, and user preferences for sociable system responses. We also vary the goal of the dialog interaction, information seeking or writing, for a total of 4 conditions across contextual information and dialog goal. We selected these because empathy and sociability have been studied for dialog systems [35, 177], and information seeking and writing are frequent goals for users [26].

We note that our setup is also a methodological contribution for testing the efficacy of personalized evaluation in a realistic but constrained setting by varying both contextual information and dialog goal.

### 5.2.4 Experiment Parameters

#### User Preferences

To determine a user’s preference for system style, we provide users with system response A (prompted to follow the user’s preferences), and system response B (prompted to explicitly not follow the user’s preferences). This mimics the interaction that the user will have with the following systems. This is to prevent the interaction having a different style than what the user expects during preference elicitation. We also choose to show a single turn of conversation (one system turn, one user turn) for preference elicitation to show four comparisons for tasks, similar to that which they would complete for the interaction, to each user in a short time period. We use the four comparisons to stratify our data into three groups: participants who had no preference (evenly split between liking the style or not), participants who had a majority preference, and participants who always preferred one style over the other. In addition, by random chance, it is possible for these groups to consist of 37.5%, 50%, and 12.5% of the participants, which provides enough participants to stratify the data for further analysis.

In particular, we focus on user preferences for empathy and sociability in dialog systems. We choose definitions for empathy and sociability that were created specifically in the context of dialog systems, and that had straightforward instructions for a system (LLM) to follow: concretely, the mention of specific types of words or styles that could be manually verified.

Our definition of empathy comes from Concannon and Tomalin, through their Empathy Scale for Human-Computer Communication. This scale was made with empathy perception in mind. It is especially useful for empathy elicitation because the descriptions of the facets of empathy contained information about how to show empathy to another party.

In particular, we choose three facets of empathy: “concern”, “resonate or acknowledge interlocutor feelings”, and “warmth”. These facets cover the four categories of empathy that Concannon and Tomalin mentioned. We directly take the item descriptions with slight modifications to wording instead of directly prompting for the blanket concept of “empathy”.

For sociability, we use the definitions present in Völkel et al.. Völkel et al. created low, rather high, and high levels of a “social-entertaining” personality through focus groups when enacting dialog scenarios. We choose to use the facets present in the “rather-high level” because the “high level” seemed to have excessive humor and anthropomorphism. In addition, Völkel et al. mentioned that people tended to prefer the “rather-high” level over the “high” level. Prompts for both empathy and sociability are provided in Appendix .6.1.

## **System**

The system will either be personalized (prompted with the participant’s stylistic preferences), or wrongly personalized (prompted with the opposite of the participant’s stylistic preferences). There should be a difference in ratings based on the user’s preferences if the user has strong preferences. The user will speak to both systems about the same task, but they are given the liberty to change the details of the task if they choose so, in order for the task to be less monotonous.

## **Tasks**

A participant will perform one of two tasks: information seeking and writing. These categories were obtained from Chatterji et al., as two of the three top used conversational categories. The information seeking task will be the user asking to for information to buy an electronic appliance, while the writing task will be the user asking for help writing a piece of communication to someone because they got into an argument with them. Prompts are provided in Appendix .6.2. These tasks are chosen because there is some spread in Völkel et al. for what people preferred for their scenarios of texting and playing music. This is meant to mirror what users may talk about in real life, supported by the real ChatGPT data from Chatterji et al..

## **Rating**

All questions we ask the user will be preference questions between the systems the user interacted with during the study. We choose preference questions because numerical scales for human raters may not provide accurate ratings for comparison due to distributional shifts as referenced in Section 3.6, and we are directly comparing two systems.

Users and raters are asked about overall satisfaction [40], conversational enjoyment [48], systems’ ease of use [48], systems’ ability to complete the task [48], the future use of the systems [48], empathy/sociability of the systems, and if the systems understood the user [48]. For the empathy/sociability of the systems and if the systems understood the user, we also ask how important it is to the user that the system exhibited that trait. Each of these questions is on a two-stage Likert scale [5, 112] in order to determine the magnitude

of preference, with the first question being whether they preferred system A or system B, and the second being for magnitude of preference. This captures all possible outputs by a user - not feeling strongly about their preference would indicate a tie no matter which system they picked.

We recognize that there are a plethora of metrics that are appropriate for these tasks, and our chosen metrics are only a subset of all possible metrics, but we choose these metrics as a representative subset of important metrics. Our metrics fall under three categories: extrinsic subjective metrics, more objective intrinsic metrics, and metrics that are related to perception of the specific system style we aim to emulate. The extrinsic subjective metrics include overall satisfaction, conversational enjoyment, ease of use, and future use of the system. The more objective intrinsic metrics include task completion ability, and if the systems understood the user. Lastly, the perception questions are whether or not the user could distinguish that systems were either more empathetic or sociable. The more objective metrics are used to obtain a baseline of variance within user ratings to determine whether the spread of the subjective metrics could be attributed to random variance or to a significant difference between raters. The perception of the specific system style is to further check if users remain consistent across their preferences and their interaction.

In addition, if there are any disagreements from the user between overall satisfaction and either empathy/sociability or the system understanding the user, we ask an additional question to confirm whether the user was satisfied with the system even though it had less empathy/sociability or understanding. We assume people who prefer empathetic chatbots will prefer the empathetic chatbot better and vice versa, but we still want to understand users' true preferences if there is misalignment between their preferences and overall satisfaction.

After the conversations are collected, we ask human raters to rate the conversations in the same format that the user ratings were collected.

### 5.2.5 Dataset Collection Methodology

Our dataset is collected in two main stages: the interaction stage and the annotation stage. Participants are drawn from Amazon Mechanical Turk.

**Interaction Stage.** First, each participant in this stage are asked to a set of pre-interaction questions. This includes the Ten Item Personality Measure (TIPI) [56], along with questions about how participants already interact with generative AI chatbots, which together indicates their attitude toward chatbots. This includes ease of use [38], usefulness [38] for information-seeking and writing tasks, frequency of use [38], and the early adoption of AI [118, 149]. The reason why we focus on these pre-interaction questions of pre-existing participant attitude and personality is because both the person as well as the situation they are in will affect their future ratings [51, 94].

Next, each participant is given the system style survey explained in Section 5.2.4, which we refer to as the empathy condition or the sociability condition. Based on their preference, they will interact with two systems as explained in Section 5.2.4. These systems are randomly swapped for each participant such that recency bias effects are mitigated. After the conversations, the participant will rate the conversation as explained in Section 5.2.4.

The person who interacted with the systems has a certain set of preferences. Take a minute to read the preferences below, and take this set of preferences into consideration when answering the following questions.

You do not like it when systems are social with you: you DO NOT like it when a system **incorporates small talk into the conversation**, you DO NOT like it when a system **incorporates funny comments into the conversation**, you DO NOT like it when a system **uses more filler words and descriptive, unnecessary adjectives into the conversation**, you DO NOT like it when a system **makes it clear that they are enjoying the conversation with you**.

Figure 5.2: Preferences shown to human annotator.

Overall satisfaction is presented as the first question after the participant finishes the dialog. Participants will then answer the remaining questions in order for overall satisfaction to not be primed by the other questions. Additionally, we can extract other features from the conversation that can contribute to a participant's satisfaction, so it is more important to obtain an unbiased user satisfaction evaluation.

**Annotation Stage.** Each rater in this stage is asked the same set of pre-interaction questions, and given the same rating questions as the interaction stage participants. The human raters will either be given or not given the system style preferences of the user whose conversation they are rating, which we refer to as having context or not having context. The preferences given to these human raters will be the definitions of the preferences along with an example of the preference survey that the user saw and rated according to their preference, example shown in Figure 5.5. Since we are including users who do not have consistent preferences in our data, we provide human raters with the definitions for more context, example shown in Figure 5.3. The human raters who are given these preferences will be asked to rate the conversations as if they have the same preferences. The human raters not given preferences will rate without additional instruction. Human raters are also presented with two attention checks throughout the survey.

We use preliminary data from about 11 participants, then full conversations from 100 participants, then annotations from 600 participants. The 100 participants are split evenly into four groups, information seeking (info-seek) with the empathy condition, info-seeking with the sociability condition, writing with the empathy condition, and writing on the sociability condition, with 75 human raters with, and 75 human raters without context in each condition.

You:

i want some help with buying a new blender but i don't know where to start

Preferred System	NOT Preferred System
<p>Blenders can be confusing to shop for, so we'll keep this simple and only cover four things:</p> <p>Use case: smoothies, soups, nut butter, crushed ice, protein shakes, baby food?</p> <p>Power: do you want "handles ice easily" or just "basic blending"?</p> <p>Budget: what's your comfy max price?</p> <p>Must-haves: big jar, single-serve cups, quiet-ish, dishwasher-safe, small footprint?</p> <p>Reply with quick answers and we'll narrow it down fast and pick a shortlist for you to start.</p>	<p>Absolutely—blenders are weirdly exciting, and I'm genuinely having fun helping you kick this off. Let's keep it super simple and only cover four things:</p> <p>Use case: smoothies, soups, nut butter, crushed ice, protein shakes, baby food?</p> <p>Power: do you want "handles ice easily" or just "basic blending"?</p> <p>Budget: what's your comfy max price?</p> <p>Must-haves: big jar, single-serve cups, quiet-ish, dishwasher-safe, small footprint?</p> <p>Reply with quick answers and we'll narrow it down fast.</p>

Figure 5.3: Example of system response preference shown to human annotator.

## 5.2.6 Analysis

To test our hypotheses, we use Cohen’s quadratic weighted  $\kappa$  (QWK). First, we convert any two-stage Likert scale questions into a 1 through 7 scale, where 1 refers to preferring a non-empathetic or non-sociable system, 4 refers to a tie between both systems, and 7 refers to preferring an empathetic or sociable system. This gives us an ordinal scale of how much someone would prefer the empathetic or sociable system. QWK is chosen because the scale between the numbers may not be consistent for each rater and we want to penalize raters based on distance to the user rating. Bonferroni corrections are applied where appropriate. Bootstrapping with 10000 resamples and seed 42 is performed across each comparison for significance testing.

## 5.2.7 Results

### Dataset Statistics

**Users and human raters have similar positive-leaning attitudes toward chatbots across both tasks.** We can see through Figure 5.4 that the majority of users and human raters, which we refer to collectively as participants, have positive-leaning attitudes toward chatbots. Participants agreed that chatbots were easy to use, found chatbots useful for both the info-seeking and writing tasks, used chatbots at least a few times a week, and would mostly adopt chatbots early. We see that the distribution across these four aspects of attitude toward chatbots is largely mirrored across the users and both sets of human raters. There is a slight mismatch on the writing task on the empathy condition, where most users use chatbots more frequently than the human raters, and on both tasks where users have a higher adoption rate than the raters. However, there is no large skew toward the lower end of the distribution in either case. As mentioned in Sections 5.2.1 and 5.2.5, participant attitude toward chatbots can how participants interact with and rate their interactions with chatbots. Although a confounding factor in ratings could be human raters’ own preferences for system interaction, human raters’ attitudes toward chatbots should have a minimal effect on rating variability with user ratings.

**Users prefer empathetic system responses but prefer non-social system responses across both tasks.** We note the distribution of preferences for system style, as explained in Section 5.2.5, across the users in Figure 5.5. Most users were either inconsistent in their preferences for empathetic system responses, or tended to prefer more of the empathetic system responses. The majority of the users tend to prefer none of the sociable system responses. It is interesting that these preferences do not vary across the info-seeking and writing tasks, as the info-seeking task was designed to evoke less preferences for system style, whereas the writing task was designed to evoke more situations where empathy would be preferred. However, it is also of note that these preferences do vary across system style. Both system styles shown in Section 5.2.5 mostly inject the system style into the preamble before providing the relevant information in body of the system response, but the text in the preamble seems to matter to users even if the amount of information provided in the body of the system responses remains the same, seen in Figure 5.3. However, due to the

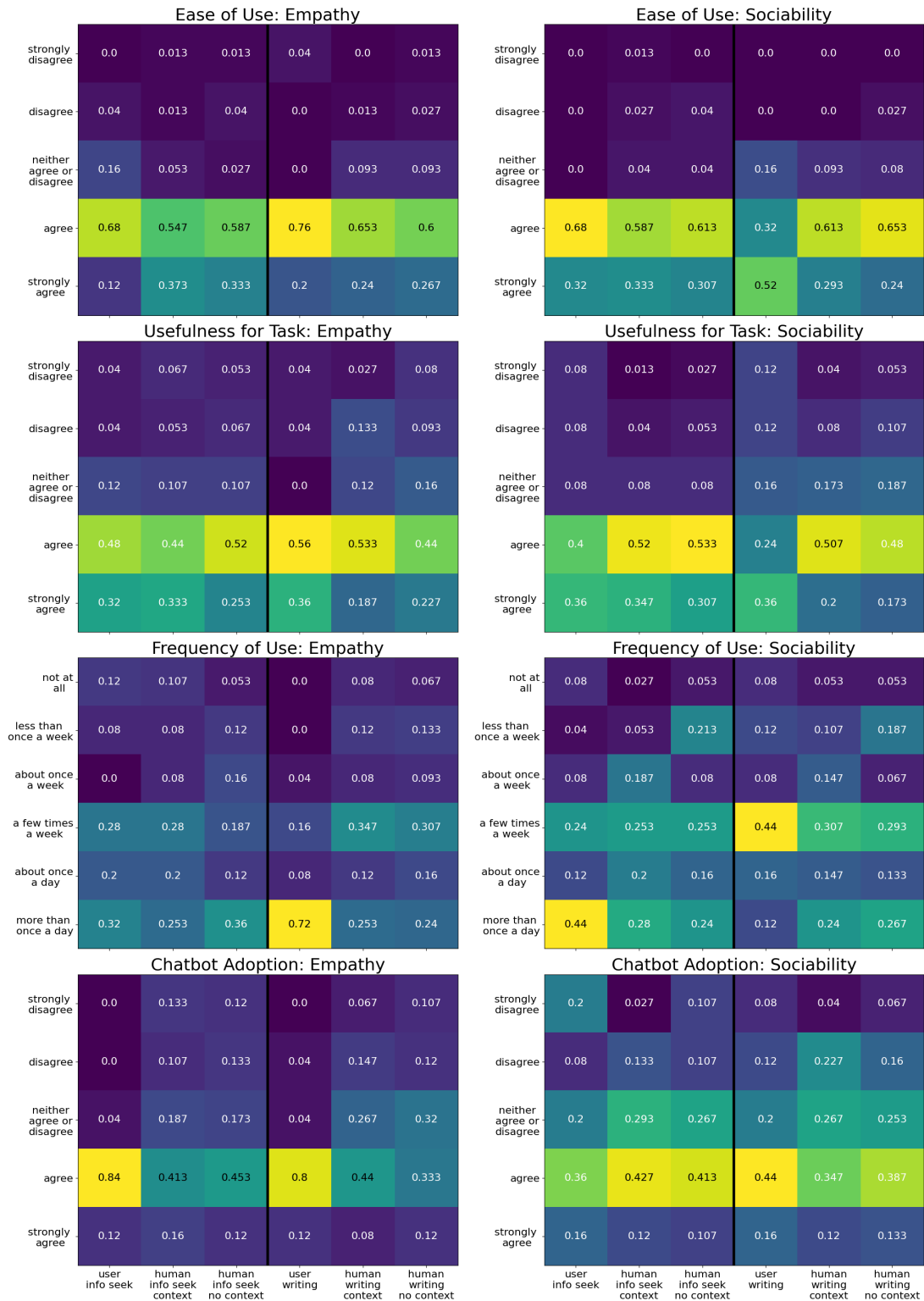


Figure 5.4: Heatmap showing the normalized distributions of our study participants (users and human raters), over our four questions related to participant attitude toward chatbots for the empathy and sociability conditions.

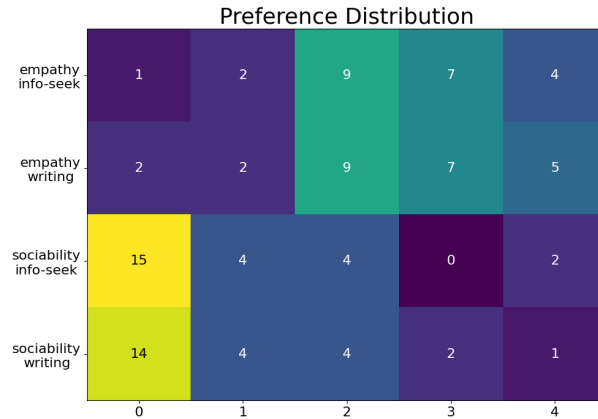


Figure 5.5: Heatmap showing the distribution of preferences across our user data from 0 as not preferring empathetic or sociable systems at all to 4 as preferring empathetic or sociable systems in all cases.

distribution of system response preference across both conditions, we are able to explore conversations where users both prefer and not prefer a specific system style.

### Personalized Evaluation Aligns Human Raters, but Not Always Autoraters, with the User

Across both system styles and tasks, providing the contextual information about user preferences always aligns human raters more with the user ratings, while autoraters do not always benefit from this information. We conclude that our hypothesis is supported for human raters but not autoraters. However, we note that autoraters tend to align more with user ratings than the human raters.

**Human Raters.** Providing context improves user-human rater agreement with QWK significantly in 5 out of 7 metrics in info-seeking on the empathy condition and both tasks on the sociability condition, and 6 out of 7 metrics in writing on the empathy condition, as seen in Table 5.1. In 7 out of 14 comparisons over both tasks on the empathy condition and 6 out of 14 comparisons over both tasks on the sociability condition, QWK becomes positive when context is given to the human raters. This shows that human raters almost always benefit from the additional context of the user’s preference for system style. It is interesting to see that in both tasks in the empathy condition, and for info-seeking in the sociability condition, that providing the preference decreases the agreement between users and human raters. This suggests that providing the definitions of empathy and sociability may interfere negatively with human raters’ preconceived perception of empathy and sociability.

**Autoraters.** Providing context does not always improve user-autorater agreement; instead, it is task dependent. For the info-seeking task, GPT improved significantly with context in 6 out of 7 metrics for the empathy condition and 5 out of 7 metrics for the sociability condition, while Llama improved significantly with context in 7 out of 7 metrics for both conditions. However, in the writing task, both autoraters improved significantly without context on the empathy condition over all metrics, and GPT improved significantly

		empathy				sociability			
metric	rater	info seek		writing		info seek		writing	
		context							
		✓	✗	✓	✗	✓	✗	✓	✗
overall satis- faction	human	0.221 <sup>↑</sup>	-0.314	<b>0.240</b> <sup>↑</sup>	-0.145	0.115 <sup>↑</sup>	-0.135	-0.119	0.010 <sup>↑</sup>
	GPT	0.337 <sup>↑</sup>	0.158	0.076	<b>0.379</b> <sup>↑</sup>	0.231 <sup>↑</sup>	<b>0.208</b>	0.302 <sup>↑</sup>	0.008
	Llama	<b>0.517</b> <sup>↑</sup>	<b>0.327</b>	-0.131	0.214 <sup>↑</sup>	<b>0.334</b> <sup>↑</sup>	0.057	<b>0.351</b> <sup>↑</sup>	<b>0.326</b>
task compl- eteness	human	-0.060	-0.009 <sup>↑</sup>	<b>0.379</b> <sup>↑</sup>	-0.243	<b>0.279</b> <sup>↑</sup>	-0.126	<b>0.299</b> <sup>↑</sup>	0.273
	GPT	<b>0.421</b> <sup>↑</sup>	<b>0.403</b>	0.110	<b>0.368</b> <sup>↑</sup>	0.137	<b>0.145</b>	0.204 <sup>↑</sup>	-0.099
	Llama	0.372 <sup>↑</sup>	0.202	-0.221	-0.191 <sup>↑</sup>	0.197 <sup>↑</sup>	0.004	0.221	<b>0.447</b> <sup>↑</sup>
empathy /sociab- ility	human	<b>0.096</b>	<b>0.208</b> <sup>↑</sup>	-0.107	0.053 <sup>↑</sup>	-0.101	-0.084 <sup>↑</sup>	<b>0.107</b> <sup>↑</sup>	0.023
	GPT	0.024	0.024	<b>0.084</b>	0.187 <sup>↑</sup>	-0.041 <sup>↑</sup>	<b>-0.059</b>	-0.085	<b>0.047</b> <sup>↑</sup>
	Llama	0.085 <sup>↑</sup>	-0.013	-0.099	<b>0.229</b> <sup>↑</sup>	<b>0.056</b> <sup>↑</sup>	-0.083	-0.058	-0.018 <sup>↑</sup>
enjoy- ment	human	0.149 <sup>↑</sup>	0.004	<b>0.191</b> <sup>↑</sup>	-0.165	0.066 <sup>↑</sup>	-0.274	0.267 <sup>↑</sup>	-0.011
	GPT	0.299 <sup>↑</sup>	<b>0.074</b>	-0.017	<b>0.323</b> <sup>↑</sup>	0.261 <sup>↑</sup>	<b>0.100</b>	0.276 <sup>↑</sup>	-0.032
	Llama	<b>0.308</b> <sup>↑</sup>	-0.006	-0.099	-0.041 <sup>↑</sup>	<b>0.287</b> <sup>↑</sup>	0.084	<b>0.353</b>	<b>0.478</b> <sup>↑</sup>
future use	human	-0.283 <sup>↑</sup>	-0.393	<b>0.172</b> <sup>↑</sup>	0.009	0.171 <sup>↑</sup>	-0.048	0.120 <sup>↑</sup>	0.069
	GPT	0.268 <sup>↑</sup>	<b>0.221</b>	-0.064	<b>0.412</b> <sup>↑</sup>	<b>0.286</b>	<b>0.344</b> <sup>↑</sup>	0.225 <sup>↑</sup>	-0.153
	Llama	<b>0.452</b> <sup>↑</sup>	0.108	-0.187	0.195 <sup>↑</sup>	0.274 <sup>↑</sup>	0.214	<b>0.288</b> <sup>↑</sup>	<b>0.286</b>
feeling under- stood	human	-0.021 <sup>↑</sup>	-0.140	0.095 <sup>↑</sup>	-0.163	0.013	0.034 <sup>↑</sup>	0.126 <sup>↑</sup>	-0.046
	GPT	0.471 <sup>↑</sup>	0.210	<b>0.147</b>	<b>0.318</b> <sup>↑</sup>	0.195 <sup>↑</sup>	<b>0.134</b>	0.238 <sup>↑</sup>	0.086
	Llama	<b>0.557</b> <sup>↑</sup>	<b>0.329</b>	-0.308	0.169 <sup>↑</sup>	<b>0.220</b> <sup>↑</sup>	0.103	<b>0.265</b>	<b>0.273</b> <sup>↑</sup>
ease of use	human	0.047 <sup>↑</sup>	-0.071	<b>0.157</b> <sup>↑</sup>	-0.021	0.135 <sup>↑</sup>	<b>0.079</b>	0.061	0.178 <sup>↑</sup>
	GPT	0.277 <sup>↑</sup>	-0.015	-0.035	<b>0.324</b> <sup>↑</sup>	<b>0.171</b> <sup>↑</sup>	-0.006	0.214	0.214
	Llama	<b>0.324</b> <sup>↑</sup>	<b>0.056</b>	-0.167	0.167 <sup>↑</sup>	0.114 <sup>↑</sup>	0.055	<b>0.318</b>	<b>0.368</b> <sup>↑</sup>

Table 5.1: This is comparing between the same rater, with and without context. The reported numbers are the Cohen’s weighted  $\kappa$  between the user ratings and the rater. <sup>↑</sup> represents which rating between with context and without context was significantly higher when comparing the bootstrapped distributions. Bold represents which rater out of human, GPT, or Llama performs significantly the best out of the three when keeping the context or lack thereof the same over the bootstrapped distributions. All comparisons between human, GPT, and Llama are significant with the exception of human/Llama without context with the enjoyment metric on the info-seeking task in the empathy condition (0.004, -0.006), GPT/Llama without context with the empathy metric on the writing task in the empathy condition (0.187, 0.229), and GPT/Llama with context with the sociability metric on the writing task in the sociability condition (-0.085, -0.058).

with context in 5 out of 7 metrics, while Llama improved significantly with context in 2 out of 7 metrics. It seems that providing context interferes negatively for the writing task in most cases. We note that based on our preference examples during preference elicitation, empathy and sociability is mostly contained within the precursor to the model information because we prompted the models to keep the information the same across system responses with and without our condition. The autoraters could be making a judgment based on the empathy or sociability in the writing text itself - which could vary between the two systems the raters are judging between.

**Autoraters agree with users more than human raters.** Although humans always benefit from this additional context, autoraters still agree with user ratings more than human raters. In Table 5.1, if we observe the highest QWK for each metric on each task regardless of context provided, we see that in the empathy condition, with the exception of empathy on the info-seeking task and task completion on the writing task, and in the sociability condition, with the exception of task completion on the info-seeking task and sociability on the writing task, at least one autorater has higher agreement with the users than human raters. We mentioned the potential difference between internal user perception and external human rater perception in Section 5.2.1, and it seems that autoraters agree more with the user’s internal perception than human raters’ external perception. We also notice that autoraters align more with user judgment on the writing task almost exclusively in cases where context is not provided in the empathy condition, while Llama behaves similarly for the writing task on the sociability condition. This suggests that the preference context for the writing task may be a distraction for autoraters but not for human raters.

When we look at the autorater performances in isolation regardless of context provided, GPT only outperforms Llama on the task completion metric on the info-seeking task in the empathy condition, but outperforms Llama on 5 out of 7 metrics on the writing task. On the sociability condition, GPT only outperforms Llama on one metric for each task. Combined with our previous findings on whether context is useful for autoraters, Llama with context on both conditions for info-seeking performs the best, while on the empathy condition for the writing task, GPT without context performs the best, and for the sociability condition, Llama without context performs the best. Overall, it seems that rating performance with and without context is dependent on the task and the rater, and this should be studied in other settings to determine if this is a larger pattern with autoraters and user’s internal perceptions.

**Other Variables.** Although our hypothesis is not fully supported, in the next three sections, we perform a deep dive into our study components - the contextual information, the user interaction, and the metrics - to try to understand other variables that could make an impact for the results of personalized evaluation.

## Contextual Interpretation and Contextual Variance

Following our mixed results on the statistical improvement of human rater and autorater agreement with users, we wanted to further understand how contextual interpretation and contextual variance could affect our results. First, we explore whether contextual information could align external raters. The contextual information was gathered by drawing conclusions from user selections in the preference survey, and although this information may not always align autoraters with users, it could be used to align human raters and autoraters if the contextual information is interpreted similarly by both external raters. We also experiment with other types of contextual information, namely, the pre-interaction questions about user attitude and personality we collected as mentioned in Section 5.2.5. We test these other types of contextual information with autoraters and not human raters because user attitudes and personality are not as easily human interpretable as empathetic

			empathy				sociability			
metric	rater1	rater2	info seek		writing		info seek		writing	
			context							
			✓	✗	✓	✗	✓	✗	✓	✗
overall satis- faction	human	gpt	0.167 <sup>†</sup>	-0.211	0.198 <sup>†</sup>	0.077	-0.201	0.020 <sup>†</sup>	-0.102	0.087 <sup>†</sup>
	human	llama	0.453 <sup>†</sup>	-0.324	0.198 <sup>†</sup>	0.183	-0.193	-0.077 <sup>†</sup>	-0.107	0.140 <sup>†</sup>
	gpt	llama	0.649 <sup>†</sup>	0.447	0.722 <sup>†</sup>	0.387	0.874 <sup>†</sup>	0.247	0.882 <sup>†</sup>	0.023
task compl- eteness	human	gpt	-0.088 <sup>†</sup>	-0.228	0.333 <sup>†</sup>	-0.130	-0.188 <sup>†</sup>	-0.225	0.039	0.069 <sup>†</sup>
	human	llama	0.086 <sup>†</sup>	-0.118	0.105 <sup>†</sup>	-0.015	-0.166 <sup>†</sup>	-0.207	0.051	0.188 <sup>†</sup>
	gpt	llama	0.597 <sup>†</sup>	0.434	0.633 <sup>†</sup>	0.411	0.729 <sup>†</sup>	0.239	0.822 <sup>†</sup>	0.046
empathy /sociab- ility	human	gpt	0.006 <sup>†</sup>	-0.102	0.104 <sup>†</sup>	0.002	-0.013	-0.006 <sup>†</sup>	0.010 <sup>†</sup>	-0.022
	human	llama	0.042 <sup>†</sup>	-0.017	-0.052	0.066 <sup>†</sup>	-0.099	-0.010 <sup>†</sup>	0.021 <sup>†</sup>	0.017
	gpt	llama	-0.269	-0.160 <sup>†</sup>	-0.316	0.169 <sup>†</sup>	-0.017 <sup>†</sup>	-0.235	-0.107	0.063 <sup>†</sup>
enjoy- ment	human	gpt	0.127 <sup>†</sup>	-0.129	0.282 <sup>†</sup>	0.089	-0.157	-0.082 <sup>†</sup>	0.057 <sup>†</sup>	-0.233
	human	llama	0.167 <sup>†</sup>	-0.289	0.193 <sup>†</sup>	0.100	-0.137	0.129 <sup>†</sup>	0.045	0.067 <sup>†</sup>
	gpt	llama	0.816 <sup>†</sup>	0.603	0.735 <sup>†</sup>	0.295	0.948 <sup>†</sup>	0.273	0.880 <sup>†</sup>	-0.120
future use	human	gpt	-0.102	-0.106	0.170 <sup>†</sup>	-0.029	-0.162	-0.015 <sup>†</sup>	-0.112	-0.053 <sup>†</sup>
	human	llama	0.143 <sup>†</sup>	-0.290	0.031 <sup>†</sup>	-0.030	-0.122 <sup>†</sup>	-0.162	-0.06	0.038 <sup>†</sup>
	gpt	llama	0.618 <sup>†</sup>	0.383	0.791 <sup>†</sup>	0.388	0.900 <sup>†</sup>	0.238	0.856 <sup>†</sup>	0.133
feeling under- stood	human	gpt	0.196 <sup>†</sup>	0.009	-0.111	-0.036 <sup>†</sup>	-0.208	-0.120 <sup>†</sup>	-0.005 <sup>†</sup>	-0.375
	human	llama	0.234 <sup>†</sup>	-0.003	-0.155	0.046 <sup>†</sup>	-0.222 <sup>†</sup>	-0.250	0.034 <sup>†</sup>	0.018
	gpt	llama	0.707 <sup>†</sup>	0.536	0.546 <sup>†</sup>	0.253	0.730 <sup>†</sup>	0.470	0.854 <sup>†</sup>	-0.012
ease of use	human	gpt	0.212 <sup>†</sup>	-0.204	0.189 <sup>†</sup>	-0.005	-0.168	0.024 <sup>†</sup>	-0.221	0.290 <sup>†</sup>
	human	llama	0.359 <sup>†</sup>	-0.122	0.040	0.073 <sup>†</sup>	-0.129 <sup>†</sup>	-0.184	-0.126	0.093 <sup>†</sup>
	gpt	llama	0.600 <sup>†</sup>	0.507	0.799 <sup>†</sup>	0.392	0.769 <sup>†</sup>	0.150	0.795 <sup>†</sup>	0.172

Table 5.2: QWK between raters when context or no context is kept the same. The reported numbers are the QWK between the each of the two raters. <sup>†</sup> represents if there is significance between the comparison of the raters with context and without context using bootstrapping.

or sociability system responses.

**Personalized evaluation aligns different external raters with each other.** Although providing contextual information did not align autoraters with user ratings, providing contextual information almost always helps align autoraters with human raters as seen in Table 5.2. We compare human raters, GPT, and Llama across each condition and task, which results in 21 comparisons across a single condition and task pair. In the empathy condition, there are 19 comparisons for the info-seeking task, and 16 comparisons for the writing task in which providing context significantly improved alignment. In the sociability condition, there are 12 comparisons for the info-seeking task, and 11 comparisons for the writing task where providing context significantly improved alignment. Autoraters almost always significantly align more with each other when provided context. However, providing context in the majority of cases harmed human rater-autorater alignment in the sociability condition. We additionally noted in Section 3.8.1 that providing examples and context significantly improved autoraters’ alignment with themselves even with different levels of rubrics, so contextual information is important for evaluation. This provides further justification that personalized evaluation is very dependent on the task and the rater, and that more domains and conditions should be tested.

**User attitude information and personality does not always allow models to predict user behavior.** Although we see that user attitudes are largely positive, we

still refer to Sections 5.2.1 and 5.2.5 about user attitudes affecting chatbot interaction and rating and test our autoraters to see if by adding user attitudes and personality, that models could better predict user’s metrics. We see in Table 5.3 that by only using user attitude and personality, or by using both user attitude, personality, and user preference, that it is rarely the case that injecting the user attitude and personality improves user-autorater agreement. This only happens in the feeling understood case in the empathy condition, but happens on the writing task in the sociability condition 6 out of 7 times using Llama with only the user attitude and personality data. It also happens in the info-seeking task for sociability conversations also with Llama when adding the user attitude information to the user preference. We note that many of our users preferred no sociability, but preferred more empathy, so it could be that when users do not prefer a specific system style, that their attitude and personality would play a greater role in predicting how they rate two systems.

### **Human and Autorater Perception of User Interactions**

Another aspect of our study aside from contextual interpretation and variance is the human and autorater perception of the user interactions themselves. This perception could be the perception of empathy, the perception of sociability, or the perception of user understanding. These perceptions could help predict overall satisfaction when combined with the perceived importance of these three traits.

**Human and Autorater Perception of Empathy.** We are interested to see whether users, human raters, and autoraters chose the system treated with empathy as more empathetic than the other system when comparing the two conversations during rating. On the info-seeking task, 12 out of 25 users, 37 out of 75 human raters with and without context, in 25 out of 25 conversation pairs for GPT with and without context, and in 20 out of 25 conversation pairs for Llama with and without context, the system treated with empathy was picked as more empathetic than the other system. On the writing task, 19 out of 25 users, 39 out of 75 human raters with context, 32 out of 75 human raters without context, in 25 out of 25 conversation pairs for GPT with and without context, in 21 out of 25 conversation pairs for Llama with context, and in 25 out of 25 conversation pairs for Llama without context, the system treated with empathy was picked as more empathetic than the other system. This suggests that models can easily determine which system was treated to exhibit more empathy even without the definition of what it means to be empathetic for a system, but that humans cannot replicate this result. This could either be due to differing empathy perceptions between human raters, or due to models being able to more effectively determine empathy from generated text. This is interesting because the definitions for empathy elicitation mentioned in Section 5.2.4 were specifically made for empathy perception. In addition, it is worth noting that the writing task was created with the expectation that more empathy could be shown throughout the entire conversation by the system. This is reflected at least through the users becoming more accurate in choosing the system treated with more empathy for this task, although human raters do not receive the same benefit.

**Human and Autorater Perception of Sociability.** We also explore whether users,

metric	style	task	best	best+c	Only pre		Add pre	
					GPT	Llama	GPT	Llama
overall satis- faction	empathy	info	<b>0.517</b>	0.517	0.150	0.350	0.362	0.514*
		writ.	<b>0.379</b>	0.240	0.218	0.286*	0.073	-0.131
	sociability	info	<b>0.334</b>	0.334	0.246	0.074	0.255	0.271*
		writ.	0.351	0.351	-0.031	<b>0.509*</b>	0.333	0.338
task compl- eteness	empathy	info	<b>0.421</b>	0.421	0.267	0.188	0.397*	0.359
		writ.	<b>0.379</b>	0.379	0.195*	-0.010	0.089	-0.249
	sociability	info	<b>0.279</b>	0.279	0.254*	-0.031	0.162	0.122
		writ.	0.447	0.299	0.016	<b>0.554*</b>	0.192	0.216
empathy /sociab- ility	empathy	info	<b>0.208</b>	0.096	0.034	0.180*	0.013	0.101
		writ.	<b>0.229</b>	0.084	0.177*	0.105	0.075	-0.230
	sociability	info	0.056	0.056	-0.164	0.053	-0.117	<b>0.065*</b>
		writ.	<b>0.107</b>	0.107	-0.174	0.047*	-0.112	-0.037
enjoy- ment	empathy	info	<b>0.308</b>	0.308	0.117	-0.067	0.193	0.203*
		writ.	<b>0.323</b>	0.191	0.263*	0.159	0.064	-0.081
	sociability	info	<b>0.287</b>	0.287	0.018	0.283*	0.224	0.222
		writ.	0.478	0.353	0.117	<b>0.529*</b>	0.299	0.351
future use	empathy	info	<b>0.452</b>	0.452	0.042	0.212	0.247	0.419*
		writ.	<b>0.412</b>	0.172	0.287*	0.165	-0.151	-0.188
	sociability	info	<b>0.344</b>	0.286	0.236	0.185	0.183	0.274*
		writ.	0.288	0.288	-0.034	<b>0.549*</b>	0.225	0.288
feeling under- stood	empathy	info	0.557	0.557	0.277	0.448	0.526	<b>0.597*</b>
		writ.	0.318	0.095	0.224	<b>0.382*</b>	0.139	-0.161
	sociability	info	<b>0.220</b>	0.220	0.131	0.075	0.123	0.205*
		writ.	0.273	0.265	0.198	<b>0.438*</b>	0.257	0.288
ease of use	empathy	info	<b>0.324</b>	0.324	0.114	0.095	0.250	0.266*
		writ.	<b>0.324</b>	0.157	0.278*	0.153	0.002	-0.148
	sociability	info	<b>0.171</b>	0.171	-0.004	0.076	0.103	0.119*
		writ.	0.368	0.318	0.172	<b>0.482*</b>	0.289	0.253

Table 5.3: Using user attitude survey questions with models to predict user annotations. Best refers to the best correlation from human raters, GPT, and Llama from Table 5.1, and best + c refers to the best correlation that was using context. The bold result represents the best correlation out of all values, and the \* represents the best correlation out of all the models using the user attitude survey questions.

human raters, and autoraters chose the system treated with more sociability as being more sociable. On the info-seeking task, 23 out of 25 users, 31 out of 75 human raters with context, 41 out of 75 human raters without context, in 25 out of 25 conversation pairs for GPT with and without context, in 5 out of 25 conversation pairs for Llama with context, and in 23 out of 25 conversation pairs for Llama with context, the system treated with sociability was picked as more social than the other system. On the writing task, 23 out of 25 users, 36 out of 75 human raters with context, 40 out of 75 human raters without context, in 25 out of 25 conversation pairs for GPT with and without context, in 10 out of 25 conversation pairs for Llama with context, and in 25 out of 25 conversation pairs for Llama without context, the system treated with sociability was picked as more social than the other system. This suggests that although sociability is an easier concept to identify than empathy for users, human raters may be more confused when given context, and Llama also does not use the context as expected. The definitions for sociability elicitation were created by focus groups mentioned in Section 5.2.4, so it is interesting that human raters tend to be more confused when given context for the sociability elicitation.

**Perception of the importance of user understanding, empathy, and sociability.** We asked our users and human raters, and prompted the autoraters for, how important it was for a chatbot to understand them and to exhibit empathy or sociability for a specific task, with results in Figure 5.6. For a chatbot understanding them, users tended to rate this as important or very important across both conditions. We can see that users did not skew as strongly toward very important as the human raters or the autoraters. The human raters tended to follow the user patterns more closely than the autoraters, which always rated “very important”, but the distributions are still very similar. We expect any user or rater to place high importance on a chatbot understanding them. However, if we look at how important it was for a chatbot to exhibit empathy or sociability for a specific task, the distributions are very different across raters. Users tended to think empathy was important for both tasks, whereas human raters with and without context had a more even distribution, but still skewing slightly positive, across whether empathy was important. The autoraters tended to be very polarizing - either rating empathy as not important or very important, with the exception of GPT without context. The sociability condition also had a similar pattern, where users tended to think sociability was not important for the task, with human raters more closely mirroring the distribution while skewing negative. This aligns with users preferring less sociable system responses. For info-seeking, the autoraters agree that sociability is less important, with the except of Llama without context. However, for the writing task, the autoraters exhibit more erratic ratings, with GPT with context and Llama without context rating sociability as very important, and only Llama with context rating sociability as not important, although the same polarizing behavior as the empathy condition is present. The external raters having such contrast with the user suggests a difference in perception between these two groups, which could lead to a difference in rating.

**Perception and Overall Satisfaction.** Users, human raters, and autoraters are overall more satisfied with systems which they perceive to understand the user more, and place high importance on this system trait. However, they are not always overall more satisfied with systems which they perceive to match the user’s preference for empathy and



Figure 5.6: Heatmap showing the normalized distributions of our study participants over whether being understood and empathy/sociability are important for this type of task. The heatmaps on the left represent the empathy condition, and the heatmaps on the right represent the sociability condition.

sociability, and have varied importance on these system traits.

All users' and human raters' ratings for the importance of user understanding, empathy, and sociability are studied. We study subsets of the data for the link between overall satisfaction and user understanding, empathy, and sociability. We expect users and any rater regardless of context to prefer systems overall which show that they understand the user (analyzing all 25 conversations for each condition and task), and users and raters with context to prefer systems overall which show the user's desired preference for empathy or sociability if they have a preference (analyzing 16 conversations for both tasks on the empathy condition and 21 conversations for both tasks on the sociability condition. For any instance where the user or human rater did not show consistency between overall satisfaction and which system was more empathetic/social/understood the user more (which we will refer to as a conflict), we asked an additional question to determine which system they were choosing to prefer overall - either keeping their original overall choice, or changing to the system that was more empathetic/sociable/understood the user more. We do not ask this question to the autoraters. All three human raters for each conversation are treated separately. The following results suggest that there may be other factors weighted more than empathy or sociability across the two tasks, but that users, human raters, and autoraters prefer systems which they perceive to show more understanding toward the user. More research needs to be done to understand the weights of these other factors and how it might affect overall satisfaction.

**Users and raters are overall satisfied with the system which is perceived to understand the user more.** In Tables 5.4 and 5.7, we can see that users, human raters, and autoraters tend to be consistent in overall preferring the system that was perceived to understand the user more. We notice that users are more consistent over both tasks in the empathy condition, but slightly less consistent, especially for the info-seeking task, in the sociability condition. We expect that users would be more consistent across the writing task, as it seems to be more important to understand the user when their goal is to create a piece of text that they will send, as opposed to information gathering, where misunderstandings could still lead to additional information being received. However, it is interesting that the consistency over the sociability condition is markedly lower. This could suggest that there are other factors that are more important than understanding the user when varying sociability. When we observe the consistency for human raters, we actually see that they are steadily consistent across both conditions and tasks. We also notice that when users or human raters are asked to resolve the conflicting ratings, they choose to switch to preferring the system that understood the user more, as seen in the C + R columns in Table 5.4. Autoraters also tend to exhibit a high level of consistency, almost higher than users and human raters in all cases. Since most of these values are about 70%, we conclude that being understood is a good indicator for overall satisfaction for users, human raters, and autoraters.

**Users and raters are not necessarily overall satisfied with the system which is perceived to exhibit empathy or sociability according to the user's preferences.** Next, we look toward the users, human raters, and autoraters for systems perceived to exhibit empathy or sociability according to the user's preferences in Tables 5.5 and 5.7 (in column + pref). This consistency is relatively low compared to overall preferring the

system that was perceived to understand the user more. This suggests that for the empathy condition, users and humans may tend to gravitate toward systems they perceive to be less empathetic even if the users liked empathetic system responses in isolation. For the sociability condition, this means that users and humans may gravitate toward systems they perceive to be more sociable even if the users liked sociable systems responses less in isolation. This raises an interesting question of misalignment in preference in isolation and preference in conversation for users and human raters. However, for autoraters, we see that GPT on both conditions, and Llama on the empathy condition are highly consistent in preferring the system that was aligned with the user’s preferences. Without context for both tasks, both autoraters are lower in consistency on the empathy condition, and GPT is lower in consistency on the sociability condition. However, on the sociability condition, GPT does not prefer the more perceived sociable system overall most of the time, whereas Llama does - this leads to GPT being more consistent when taking into consideration most users preferred less sociability, and Llama less so.

**Additional Observations.** Although users and raters do not consistently prefer the system which showed empathy or sociability according to the user’s preferences, additionally, we explore whether users and raters generally prefer systems which they perceive to exhibit more empathy or sociability, or if they generally prefer systems which we designed to exhibit more empathy or sociability. Users and raters generally prefer systems which are perceived to exhibit more empathy, but not necessarily sociability - seen in both tables with column  $\uparrow$ . Based on the user preferences shown in Figure 5.5, we expect consistency to be higher under the empathy condition since more users preferred empathetic responses, and lower under the sociability condition since more users did not prefer sociable responses. We observe that users and human raters are still relatively consistent over the empathy condition in both tasks, but that consistency degrades greatly over the sociability condition in both tasks, even after conflicts are resolved, which aligns with our initial thoughts.

Users and raters also somewhat prefer systems designed to exhibit more empathy, but not sociability, which is also consistent with user’s system style preferences. We explore this in Table 5.6 and the design. column in Table 5.7. There is not a consistent overwhelming majority of users and human raters who prefer the systems designed to exhibit more empathy or sociability, but there is a slight lean toward positive. The autoraters tend to prefer the systems designed to be more empathetic, but heavily do not prefer the systems designed to be more sociable except for Llama without context.

## Construct and Metric Ratings

The last aspect of our study is the metrics that we choose to study. In Section 5.2.4, we mentioned three types of metrics: extrinsic subjective metrics, objective intrinsic metrics, and perception metrics. We believe that subjective metrics may receive more of a benefit than objective metrics with personalized evaluation, as subjective metrics could be more context-dependent.

**Improvement with personalized evaluation does not depend on the type of metric.** In Table 5.1, we observe that there is no noticeable difference between the subjective (9, 9, 7, and 9 instances out of 12) and objective metrics (7 and 8 instances out

				Empathy Condition		Sociability Condition	
		rater	cxt.	C	C+R	C	C+R
infoseek	user			0.800	-	0.520	0.600
	human	✓		0.787	0.827	0.787	0.840
	human	✗		0.680	0.707	0.720	0.760
writing	user			0.920	0.960	0.720	-
	human	✓		0.760	0.880	0.747	0.813
	human	✗		0.853	-	0.707	0.747

Table 5.4: User and human rater normalized consistency across overall satisfaction and feeling understood. C stands for consistent, and C + R stands for consistent and resolved from asking the user and raters additional questions to resolve conflicts. - represents that there was no change after resolution.

				Empathy Condition				Sociability Condition			
				↑		+ pref		↑		+ pref	
task	rater	cxt.	C	C+R	C	C+R	C	C+R	C	C+R	
infoseek	user		0.688	-	0.375	-	0.476	0.524	0.571	0.524	
	human	✓	0.813	-	0.542	-	0.254	0.302	0.333	0.302	
	human	✗	0.604	-	0.208	-	0.540	-	0.222	-	
writing	user		0.750	-	0.375	-	0.524	-	0.524	-	
	human	✓	0.604	0.688	0.250	0.333	0.460	0.492	0.302	0.286	
	human	✗	0.771	-	0.25	-	0.635	0.651	0.190	-	

Table 5.5: User and human rater consistency across overall satisfaction and empathy/sociability preference. C stands for consistent, and C + R stands for consistent and resolved from asking the user and raters additional questions to resolve conflicts. - represents that there was no change after resolution. +pref represents adding the user’s preferences in mind - does overall satisfaction agree with the rater’s views on the more empathetic/sociable system and the user’s initial preference for empathy/sociability.

task	rater	cxt.	Empathy Condition		Sociability Condition	
			C	C+R	C	C+R
infoseek	user		0.438	-	0.429	0.476
	human	✓	0.521	-	0.492	0.508
	human	✗	0.375	-	0.476	-
writing	user		0.688	-	0.571	-
	human	✓	0.396	-	0.587	-
	human	✗	0.458	-	0.540	0.524

Table 5.6: User and human rater consistency across overall satisfaction and empathy/sociability preference. C stands for consistent, and C + R stands for consistent and resolved from asking the user and raters additional questions to resolve conflicts. - represents that there was no change after resolution. This table answers if people are overall more satisfied with the systems we designed to be more empathetic/sociable.

task	model	context	Empathy Condition				Sociability Condition			
			understood	empathy			understood	sociability		
				↑	+ pref	design.		↑	+ pref	design.
info seek	gpt	✓	0.840	0.750	0.938	0.750	1.000	0.095	1.000	0.095
	gpt	✗	0.760	0.438	0.500	0.438	0.920	0.143	0.762	0.143
	llama	✓	0.920	1.000	0.750	0.750	1.000	1.000	0.095	0.095
	llama	✗	0.920	0.750	0.563	0.625	0.960	0.619	0.381	0.524
writing	gpt	✓	0.840	0.688	0.938	0.688	1.000	0.143	1.000	0.143
	gpt	✗	0.920	0.500	0.375	0.500	0.760	0.286	0.667	0.286
	llama	✓	0.960	0.938	0.688	0.688	1.000	0.857	0.286	0.143
	llama	✗	0.880	0.688	0.625	0.625	0.880	0.762	0.381	0.762

Table 5.7: Autorater consistency across overall satisfaction and empathy and understood preference. We report the autorater consistencies.

of 12) for improvements with personalized evaluation when considering all human raters and autoraters. However, we do notice that personalized evaluation is not as useful for empathy (1 instance out of 6) and sociability (3 instances out of 6) perception.

### 5.2.8 Conclusion

Throughout this work, we aimed to understand whether internal and external ratings aligned, and whether personalized evaluation was useful for the empathy and sociability conditions as well as the info-seeking and writing tasks. We discover that providing context for external human raters aligns human raters significantly more to the internal user ratings, but that this does not necessarily extend to autoraters when provided context. However, providing context for any external rater does help align external raters with each other. Autoraters tend to agree more with users than human raters, which suggests that it is very useful to obtain internal user ratings, and that more work should be done on providing easy ways for users to give ratings during or after a conversation with a system. Users and raters tend to overall prefer systems that were perceived to understand the user more, but not necessarily overall prefer systems that showed or were perceived to show empathy and sociability according to the user's preferences. Users' attitudes toward chatbots and their personality do not necessarily outperform situations in which they are not used in evaluation. All this suggests that personalized evaluation is useful in certain contexts, and that it is important to collect user's internal ratings and ensure data collection about the user is quantitatively and statistically useful for personalized evaluation to only collect necessary personal information. These experiments should be performed over additional domains and user preferences, but personalized evaluation has promise to provide quantitative and statistical improvement for both human raters and autoraters.

# Chapter 6

## Conclusion

Throughout this thesis, we explore the promise of using autoraters for personalized evaluation. Our contributions are as follows:

1. Chapter 3 provides understanding into how prompt design affects human-autorater agreement, as well as how autoraters can be used to evaluate dialog systems. By carefully constructing prompts for autorating, we can start to use autoraters in product settings to estimate performance.
2. Chapter 4 adds another layer on to prompt design by injecting user knowledge (attitude) into autoraters. Through simulation, we find that it is important to classify negative attitudes because they have varying impact on autorater ratings. We now have another promising avenue to explore how we can achieve higher agreement with humans.
3. Chapter 5 brings together our knowledge about autorater prompts and user behavior together into applying autoraters in an interaction setting for personalized evaluation. This chapter provides a starting point for further gains in using autoraters in real settings.

We revisit our initial motivations for doing this work. There has been a large shift towards using autoraters to replace human raters. Although in Chapter 5, we see that autoraters with context perform better than external human raters with context, we still note that these autoraters do not have an extremely high agreement with user ratings. Through our work on personalized evaluation, we realize even more that we need to understand humans, how they interact with systems, and how they perceive systems before we can truly be confident in using automatic personalized evaluation. Knowing this information, we recommend using automatic personalized evaluation for system interaction as long as it is quantitatively studied to provide a significant improvement in agreement with user behavior over all other settings. However, it is still very important to keep humans in the loop during some part of the evaluation process. We now provide a few future directions for consideration.

**Autoraters.** We show in Chapter 3 that choosing suitable prompts and autoraters for specific tasks is important. We mention some related work that studies how prompt modifications affect autorater performance. However, we uncovered that by purely sep-

arating rubric criteria into separate API calls, there can be a significant improvement in some cases over presenting all the rubric criteria at once to an autorater. We also noted that humans were presented all the rubric criteria at once. Although it is not possible to mimic the separate API calls condition to humans, it would be interesting to study potential ordering bias in humans by varying the order in which rubric criteria were given, and seeing if certain autoraters can mimic this bias. Throughout this thesis, we move toward separate API calls in later chapters, and we recommend future autorater works to continue presenting criteria in separate API calls until this bias is studied.

**Humans.** There is a large body of work on how to best present human raters with instructions such that researchers can obtain suitable ratings, with some methods summarized in my previous work [72]. However, seen in Chapter 5, it may not be possible to provide exact examples to external human raters because we lack additional information about the user themselves. It may be possible to aggregate a user’s previous conversations to obtain additional information about them, but depending on the sensitivity of the information gathered, it may not be appropriate to provide that information to external human raters.

It may also be difficult to provide examples to the users themselves. We will discuss future directions for contextual information in the section below, but we study a limited amount of contextual information in Chapter 5, including empathy or sociability preference, user attitude toward chatbots, and personality through TIPI. Although there is noticeable improvement to providing context to human raters and sometimes autoraters, we still see that the raw agreement ratings are low. We also noted low consistency with empathy and sociability perception. This may mean that we do not have a good grasp yet on how users with these preferences rate. Before having this understanding, providing examples to users may not make sense because it could skew them to have interpretations that are not how they truly view the conversation.

It is also difficult to obtain these user ratings at all. Previous works [73] have used a thumbs up/down methodology with feedback options that give users the option of providing feedback throughout the conversation. However, it would be difficult to ask users for fine-grained ratings in a real use case setting because it is not realistically how users have conversations with chatbots. After a user has a conversation, they may not want to provide any feedback at all after completing their task. Therefore, there should be more work in understanding what other types of feedback we could obtain from users in this setting, for example, implicit feedback, which has been studied for recommender systems [78, 80].

We can see through Chapter 5 that the elicitation and presentation of preferences could be another direction for future work. We structure our preference elicitation carefully, however, we noted that even within the same type of task, users may not prefer empathy or sociability in every case. Future work could look towards preference elicitation such that users are consistent in their preferences. Human raters or autoraters could also prefer different presentations of these preferences.

**Contextual Information.** In our works, we mention using more concrete user preferences on empathy or sociability as a basis for personalized evaluation, and explore more abstract facets like user attitude and personality with no success. However, there can be many other facets of contextual information that could help inform personalized evalua-

tion like culture [23]. It would be interesting to explore whether other types of contextual information show the same quantitative improvement in rating agreement, as well as how concrete or abstract the contextual information can be. There could be another research direction in the presentation of contextual information by using abstract representations of the user in a trained user model to evaluate instead of natural language prompts for large language models.

We mention our focus on explicit information for the user profile in Section 2.3.1; however, there are many sources of implicit information that could be used to supplement or create a user profile as well. Previous system personalization papers have used methods to extract abstract representations through vectors or models of the user through dialog history [140, 183, 208], or concrete user signals throughout the conversation [132]. Therefore, there is an interesting direction to using previously studied methods for implicit user information and seeing if either prompting a model which has an abstract representation of a user for a user profile, or creating a user profile through implicit signals could either help supplement or perform better than user profiles created through explicit information.

**Constructs.** Lastly, there are many future directions for the constructs to study for personalized evaluation. In Chapter 5, we perform experiments related to consistency of rating with overall satisfaction based on how important empathy or sociability is, and how important it is for the chatbot to understand the user. These experiments could be extended to other constructs to further understand how consistent they are with overall satisfaction. For constructs consistent with overall satisfaction, it may be appropriate to conclude that the construct largely affects overall satisfaction. Additionally, future work could look toward understanding the weighting of these constructs within overall satisfaction.

There is also a question of what constructs are appropriate to measure. There have been many works trying to capture important constructs [48, 120] for dialog; however, as chatbots evolve, certain constructs may saturate as noted in Chapter 4. We may also find constructs that are more indicative of chatbot performance. Therefore, it is important to constantly evaluate our evaluations.

Finally, this thesis focuses on understanding quantitative statistical differences for any modification we make. However, there is a large body of work looking at qualitative evaluation or understanding the rationale behind quantitative evaluations. For example, think aloud evaluation with humans [190] or using humans and LLMs together to create a rubric [33]. Autoraters could do something similar with Chain-of-Thought (CoT) [31].



# Chapter 7

## Appendix

# .1 Understanding the Effectiveness of Very Large Language Models on Dialog Evaluation Appendix

## .1.1 Prompt format examples FED

**Task:** Given a dialog history and a response, rate how interesting the response is with regards to the dialog history.

== **Example 1** ==

A: Hi!

B: Hi. This is a pleasant surprise.

A: Haha...thanks! how did you like the gift?

**Response:** Currently unpacking it I guess. How's your morning?

**Rating:** 1/2

A: Hope you like it! Morning is good. Busy finishing up stuff before the holidays.

B: I think I traveled too much the last couple of months so no holiday for me. But I'm okay with that. Going anywhere exciting?

A: Yes

**Response:** Where to?

**Rating:** 1/2

A: Hawaii... looking forward to warm beaches.

**Response:** WOW. Which island? I like Hawaii.

**Rating:** 2/2

Table 1: An example of a prompt with one example from FED [120]. Interestingness was rated in FED over three values corresponding to 0/2, 1/2, and 2/2. The resulting output is truncated to the integer value of 0, 1, or 2 to be used in evaluation.

## .1.2 Prompt format examples DSTC10

<p><b>Instruction:</b> Given a conversation and a response, choose if the response is a good response to the context</p> <p><b>Example</b></p> <p><b>Background info:</b> none</p> <p><b>Conversation:</b> Person A: did your meal meet with your approval ? <b>Response:</b> yes , i did . it was a good meal . <b>Question:</b> Is the above response a good response to the conversation? <b>Answer:</b> Yes</p> <p><b>Background info:</b> none</p> <p><b>Conversation:</b> Person B: i really do hate public transportation. Person A: i agree , it 's just never on time. Response : you 're right. <b>Question:</b> Is the above response a good response to the conversation? <b>Answer:</b></p>
--

Table 2: An example of a prompt with examples from DSTC 10.

## .2 Influence of Prompt Design on Human-Autorater Agreement Appendix

### .3 Appendix A

#### .3.1 Additional Other Observations

$\Delta$ **rater** + **rubric**. Different raters may show greater agreement when using different rubrics. In addition to the previous comparisons of  $\Delta$ **Rater** and  $\Delta$ **Rubric** we also investigate  $\Delta$ **Rater+Rubric**, which captures cases where alignment between different raters is affected by the use of different rubric types as depicted in Figure 1. The best  $\tau$  obtained for any rubric condition when comparing human ratings on holistic rubrics with autoraters on analytic rubrics ( $\text{Human}_H \rightarrow \text{LLM}_A$ , 0.401, 0.682, and 0.644 respectively) is lower than the worst  $\tau$  obtained from comparing autoraters on holistic rubrics with humans on analytic rubrics ( $\text{LLM}_H \rightarrow \text{Human}_A$ , 0.539, 0.688, and 0.645 respectively) (Table 3.11). This indicates that while autoraters exhibit high preference agreement across different rubrics

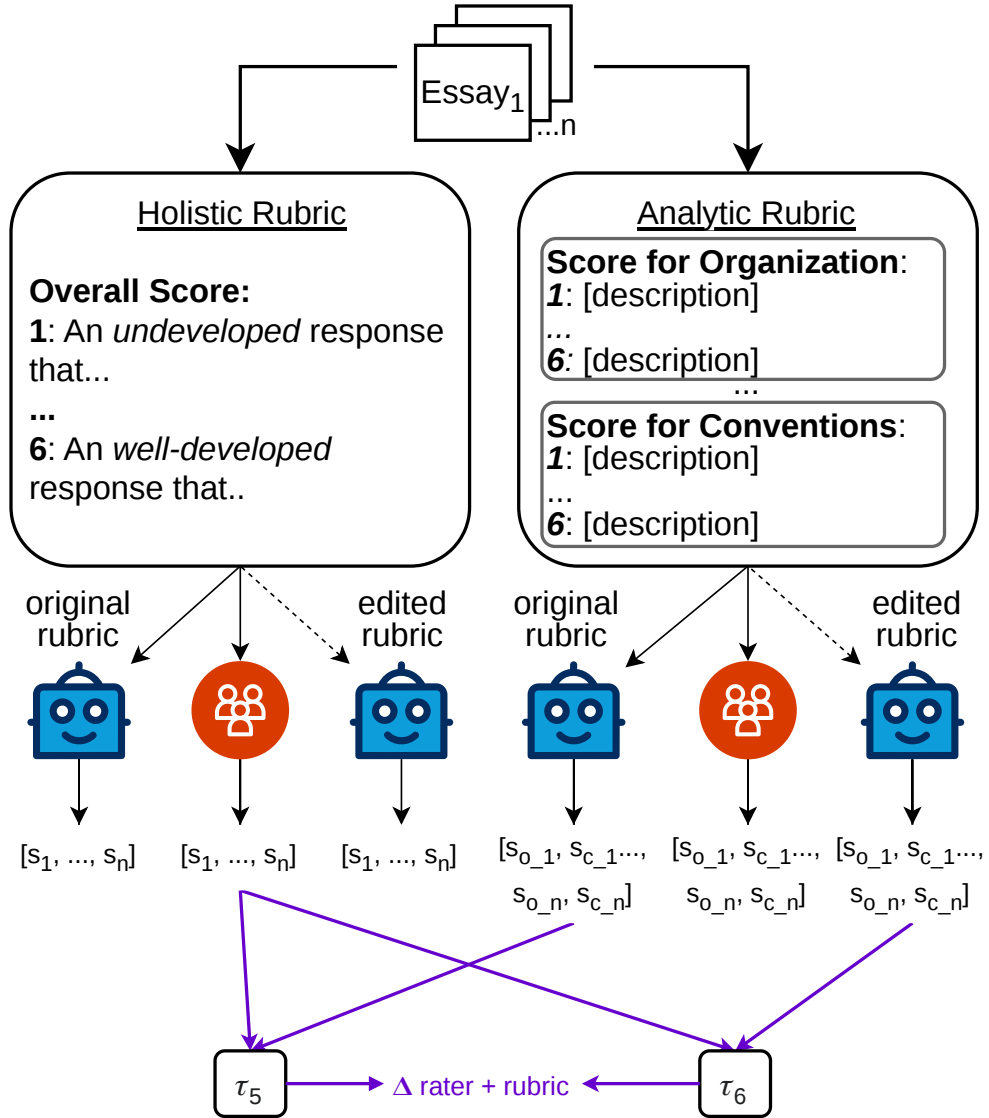


Figure 1: This diagram represents comparisons made between human-atorater agreements  $\tau$  across various conditions such as holistic rubrics (left side), in which all criteria are applied together in a single overall judgment, or analytic rubrics (right side), in which criteria are evaluated separately, and rubrics which have been edited for LLMs. Arrow in bold between  $\tau_5$  and  $\tau_6$  represent comparisons for which statistical significance can be calculated.

P	C.	Full	3ex	Sep.	Bat.	Edited
1	H.	0.639	0.623	-	-	-
	Ide.	-	-	0.647	0.652	0.660
	Or.	-	-	0.675	0.626	0.672
	WC	-	-	0.616	0.580	0.639
	SF	-	-	0.608	0.586	0.632
	Cv.	-	-	0.580	0.559	0.610
4	H.	0.825	0.812	-	-	-
	Ct.	-	-	0.777	0.815	0.801
	PA	-	-	0.767	0.773	0.784
	La.	-	-	0.755	0.711	0.766
	Na.	-	-	0.767	0.747	0.789
6	H.	0.790	0.782	-	-	-
	Ct.	-	-	0.716	0.779	0.772
	PA	-	-	0.712	0.746	0.780
	La.	-	-	0.745	0.705	0.753
	Na.	-	-	0.749	0.709	0.772

Table 3: Kendall’s  $\tau$  between GPT and Llama. All abbreviations follow Table 3.11.

		$\Delta$ rater+rubic									
		GPT-4o					Llama				
		$LLM_H \rightarrow H_A$		$H_H \rightarrow LLM_A$			$LLM_H \rightarrow H_A$		$H_H \rightarrow LLM_A$		
AES	1	0.574 <sup>†</sup>	0.539	0.279	0.271	0.401 <sup>s,b†</sup>	0.601	0.605	0.467	0.467	0.470
	4	0.691	0.688	0.669	0.668	0.682 <sup>s,b†</sup>	0.674	0.674	0.710*	0.674	0.707 <sup>b†</sup>
	6	0.645	0.653 <sup>†</sup>	0.585	0.595*	0.644 <sup>s,b†</sup>	0.660	0.670 <sup>†</sup>	0.637*	0.592	0.671 <sup>s,b†</sup>
IF		0.541	0.530	0.474*	0.34	0.459 <sup>b†</sup>	0.478	0.545	0.421*	0.183	0.424 <sup>b†</sup>

Table 4: Kendall’s  $\tau$  results for  $\Delta$ rater+rubic with GPT-4o and Llama on AES and IF. For AES, the number indicates the essay prompt, with ideas, organization, word choice, sentence fluency, and conventions compared for prompt 1, and content, prompt adherence, language, and narrativity compared for prompts 4 and 6.  $\tau$  is calculated through Pareto dominance comparison for AES and through instruction following ratio comparison for IF. Significance tests between separate (sep.), batch (bat.), and edited prompts are performed, where <sup>s</sup> and <sup>b</sup> in the edited prompt column represents significant differences with separate and batch prompts respectively. <sup>†</sup> is indicated next to comparisons that are significantly larger within holistic prompts. <sup>\*</sup> is indicated next to comparisons that are significantly larger between separate and batch comparisons. <sup>†</sup> and <sup>↓</sup> represent that the  $\tau$  value for edited prompts is significantly larger or smaller respectively with the separate <sup>s</sup> or batch <sup>b</sup> prompts’  $\tau$ . The lack of any dagger, star, or arrow denotes no statistical significance. H is a shortened form for Human.

(in  $\Delta$  rubric), the Pareto dominance aggregation method decreases alignment between aut raters and humans. This trend is also observed in Table 3.12 (0.474 to 0.530, and 0.424 to 0.478) for IF, suggesting its applicability to the ratio aggregation method. However, in

Table 3.11 on Llama for AES, this only occurs for essay prompt 1 (0.470 to 0.601).

Surprisingly, for essay prompt 4, all  $\tau$  values for  $\Delta$ rater + rubric are higher than those for  $\Delta$  rubric when humans are the annotator, for both GPT-4o (0.656 compared with 0.691 and 0.688) and Llama (0.656 compared with 0.674 and 0.674). This suggests that the autorater’s preferences on the holistic rubric agree slightly more than the human’s preferences on the holistic rubric as compared to the analytic rubric. Although different humans rate with the holistic and analytic rubrics, they all either were trained or familiar with evaluation, whereas there is no guarantee for autorater performance. This could be due to various factors, such as the autorater using a continuous scale versus the human’s discrete scale which leads to less ties, or the autorater’s interpretation being closer to the human’s analytic rubric interpretation. Future work should further explore these correlations.

**GPT-4 vs. Llama.** It is hypothesized that autoraters will exhibit greater mutual alignment when (1) using analytic prompts instead of holistic prompts, (2) using the same (and more) examples, and (3) using separate API calls instead of batched API calls. To test these hypotheses, Kendall’s  $\tau$  is calculated between GPT-4 ratings and Llama ratings in Table 3. Analytic rubrics do not consistently improve autorater alignment over holistic rubrics, even when provided examples in AES (0.639 compared to 0.610 - 0.672, 0.825 compared to 0.766 - 0.801, and 0.790 to 0.753 - 0.780 respectively). However, introducing examples to analytic rubrics (comparing separate and batch with the edited column) and increasing the number of examples in holistic rubrics (0.639 from 0.623, 0.825 from 0.812, and 0.790 from 0.782) does tend to improve alignment, indicating that examples improve both human-autorater and autorater-autorater alignment in AES. This pattern does not hold for IF, where the separate, batch, and edited conditions yield  $\tau$  values of 0.674, 0.193, and 0.662, respectively, while the holistic zero-shot and three-shot conditions yield 0.752 and 0.724. Separate API calls improve alignment for 8 out of 13 sub-criteria in AES. However, as previously mentioned, further research is needed to understand how humans and autoraters exhibit positional bias, as the sub-criteria presented first in batched API calls consistently show higher alignment.

### .3.2 Agreement Level

Agreement level is examined using holistic prompts in both AES and IF, and analytic prompts in IF. As shown in Table 6 with GPT-4o, for essay prompts 4 and 6 (where both raters agreed on 77.1% and 62.2% of essays, respectively), essays with rater agreement exhibited higher  $\tau$  than those without (for full examples,  $\tau$  of 0.687 to 0.525 and 0.659 to 0.642 respectively, for 3ex,  $\tau$  of 0.678 to 0.512 and 0.675 to 0.653 respectively), with a significant difference in essay prompt 4. However, this pattern did not hold for essay prompt 1 (where raters agreed on 65.3% of essays, for full examples,  $\tau$  rose from 0.450 to 0.454, and for 3ex,  $\tau$  rose from 0.398 to 0.413 albeit not significantly). Despite greater disagreement among human raters, the autorater showed higher agreement with the aggregated human scores. This discrepancy might stem from the different aggregation methods used: The final score for essay prompt 1 is the average of both raters’ scores, while essay prompts 4 and 6 use a single rater’s score (or a third expert rater’s score). Averaging scores could

Prompt	Source	Rubric Type	Distinction
1	Argumentative (students asked to state an opinion based on the prompt)	General	Evaluates opinion with a reusable rubric
4	Source-dependent (students asked to read a passage and respond to a prompt with details from the passage)	General	Evaluates comprehension of a source using a reusable rubric
6	Source-dependent (students asked to read a passage and respond to a prompt with details from the passage)	Task-specific	Evaluates comprehension using a highly specific, non-reusable rubric

Table 5: ASAP Chosen Essay Prompts

M	Prompt	Comp.	Full	3ex
GPT-4o	1	agree	0.450	0.398
		disagree	0.454	0.413
	4	agree	0.687 <sup>†</sup>	0.678 <sup>†</sup>
		disagree	0.525	0.512
	6	agree	0.659	0.675
		disagree	0.642	0.653
Llama	1	agree	0.585	0.581
		disagree	0.593	0.592
	4	agree	0.694 <sup>†</sup>	0.691 <sup>†</sup>
		disagree	0.548	0.529
	6	agree	0.696	0.707
		disagree	0.677	0.693

Table 6: Kendall’s  $\tau$  results on AES with GPT-4o and Llama. The number indicated by the <sup>†</sup> represents that the  $\tau$  value on that subset of data (in all cases here, where raters agreed with each other) is significantly larger than the  $\tau$  value on the other subset of data (where raters disagreed with each other). The lack of any arrow denotes no statistical significance.

diminish the impact of rater disagreement on  $\tau$  in essay prompt 1. Llama ratings mirror these patterns.

In the IF dataset, holistic prompt ratings (with three human raters) were divided into three subsets: full agreement (all three raters agree, 25.9% of the dataset), partial agreement (two raters agree, 53.3% of the dataset), and full disagreement (no raters agree, 20.9% of the dataset). The final score was obtained by averaging the three raters’ scores. Analytic prompt ratings were divided into full agreement (13% of the dataset) and partial

M	Comp.	0ex	3ex	Separate	Batch	Edited
GPT-4o	Full agreement	0.792 <sup>p↑,f↑</sup>	0.803 <sup>p↑,f↑</sup>	0.299 <sup>p↑</sup>	0.078	0.31 <sup>p↑</sup>
	Partial disagreement	0.559 <sup>f↑</sup>	0.584 <sup>f↑</sup>	0.251	0.028	0.261
	Full disagreement	0.220	0.268	-	-	-
Llama	Full agreement	0.730 <sup>p↑,f↑</sup>	0.865 <sup>p↑,f↑</sup>	0.283 <sup>p↑</sup>	0.132	0.265 <sup>p↑</sup>
	Partial disagreement	0.533 <sup>f↑</sup>	0.625 <sup>f↑</sup>	0.242	0.017	0.203
	Full Disagree	0.083	0.229	-	-	-

Table 7: Kendall’s  $\tau$  results on IF with GPT-4o for  $\Delta$ rater. Significance tests between data with full agreement, partial agreement, and full disagreement are performed, where <sup>p</sup> and <sup>f</sup> in the full agreement row represents significant differences with partial disagreement and full disagreement data respectively.  $\uparrow$  and  $\downarrow$  represent that the  $\tau$  value is significantly larger or smaller respectively with compared  $\tau$ . The lack of any arrow denotes no statistical significance.

agreement (87% of the dataset) subsets, using majority vote for the final score. Kendall’s  $\tau$  was calculated within the subsets for each prompt. Table 7 demonstrates a consistent trend for both GPT-4o and Llama: lower human agreement corresponded to lower  $\tau$ , with most comparisons showing statistical significance (for GPT-4o, for the holistic prompt, with no examples,  $\tau$  dropped from 0.792 to 0.559 to 0.220, with 3 examples,  $\tau$  dropped from 0.803 to 0.584 to 0.268; while for the analytic prompt, separate dropped from 0.299 to 0.251, batch from 0.078 to 0.028, and edited from 0.31 to 0.261 respectively) . To achieve high human-atorater alignment, one must first establish high human-human alignment. While a rubric cannot fix inherently ambiguous data, it is the primary tool for reducing rater disagreement and creating the stable ‘ground truth’ necessary for meaningfully evaluating and training an autorater.

## .4 Appendix B

### .4.1 Prompt for AES

The holistic and analytic rubrics within the prompts are taken from ASAP [64] and ASAP++ [117]. The contextual information present in the grading guidelines are formatted into the prompt, and additionally added with slight edits to the analytic prompts in the edited condition. The examples chosen from ASAP++ are essay IDs 449, 1264, 1616, 9125, 9430, 9497, 15558, 16520, and 16581.

### Prompt 1 Holistic Context

#### Instructions:

- 1) The following essay is a first draft written by an 8th grade student in forty-five minutes in reaction to a prompt designed to elicit persuasive writing. You will score these timed responses holistically, which means that you will determine a score based on the overall impression most often gained from a single reading of the response.
- 2) You will be given a rubric that outlines a six-point scale. Each score point on that scale is described by an overall statement which captures the essence of the response. The elements of the response (elaboration, organization, fluency and audience awareness) that are typical for that score point are described below the overall statement. Individual responses may be stronger in one feature and weaker in another. In other words, the list of features at each score point, while helpful, cannot perfectly describe every response in a score point category.
- 3) A committee of expert readers uses this rubric as a guide to select anchor papers for each score point. Anchor papers are examples of actual student work. The committee prepares an anchor set composed of several papers at each score point. They deliberately select papers to show an appropriate range of writing skill for each score point and to represent the variety of approaches students take when addressing the writing prompt. You rely heavily on these anchor sets to guide your scoring.
- 4) Errors in spelling, punctuation, grammar, and usage are not considered as part of the criteria for scoring. Also, papers receive a score based on the work the student did complete even if they seem to be unfinished. Because the writing sample is a timed response, it is generally assumed that these errors and omissions could have been corrected if the student had been given an opportunity to revise and edit. You are trained to read through these errors when you score student papers.
- 5) Score the essay on a scale from 1 to 6.

### Prompt 4 Holistic Context

#### Instructions:

- 1) The following essay is written by an 10th grade student in response to a prompt that is dependent on reading the story provided.
- 2) You will be given a rubric that outlines a four-point scale.
- 3) Training materials consist of a rubric and a scoring guide of ten responses.
- 4) Score the essay on a scale from 0 to 3.

### Prompt 6 Holistic Context

#### Instructions:

- 1) The following essay is written by an 10th grade student in response to a prompt that is dependent on reading the excerpt provided.
- 2) You have a four year baccalaureate degree as well as documented coursework in English. You are not a teacher, substitute teacher, support staff, tutor, administrator, etc., who is currently under contract or employed by or in schools, or under 18 years of age.
- 3) You will be given a rubric that outlines a five-point scale.
- 4) You will be given an anchor set which will consist of responses that are typical, rather than unusual or uncommon; solid, rather than controversial or borderline; and true, meaning that these have scores that cannot be changed by anyone other than pertinent personnel. Anchor sets will typically have 2 to 3 sample responses at each score point (the middle score points will have 3 sample responses, 1 representing the mid-high to high end of the score point range, 1 in the middle, and 1 at the mid-low to low end).
- 5) Score the essay on a scale from 0 to 4.

### Example of context given to an analytic prompt for separate (Prompt 1)

#### Instructions:

- 1) The following essay is a first draft written by an 8th grade student in forty-five minutes in reaction to a prompt designed to elicit persuasive writing.
- 2) You will be given a rubric that outlines a six-point scale for an attribute.
- 3) Score the essay on a scale from 1 to 6 on the attribute.

### Example of context given to an analytic prompt for edited (Prompt 1)

#### Instructions:

- 1) The following essay is a first draft written by an 8th grade student in forty-five minutes in reaction to a prompt designed to elicit persuasive writing.
- 2) You will be given a rubric that outlines a six-point scale for an attribute.
- 3) A committee of expert readers uses this rubric as a guide to select anchor papers for some score points. Anchor papers are examples of actual student work. The committee prepares an anchor set composed of several papers at various score points. They deliberately select papers to show an appropriate range of writing skill and to represent the variety of approaches students take when addressing the writing prompt. You rely heavily on these anchor sets to guide your scoring.
- 4) Also, papers receive a score based on the work the student did complete even if they seem to be unfinished. Because the writing sample is a timed response, it is generally assumed that these errors and omissions could have been corrected if the student had been given an opportunity to revise and edit. You are trained to read through these errors when you score student papers.
- 5) Score the essay on a scale from 1 to 6 on the attribute.

## .4.2 Prompt for Instruction Following

The holistic and analytic prompt are structured to match the annotation instructions presented to human annotators of the InfoBench dataset [141].

## Evaluation Prompt

I will show you examples of how to evaluate system responses to specific criteria.

EXAMPLE 1: {}

EXAMPLE 2: {}

EXAMPLE 3: {}

Now, please evaluate the current response based on the following criteria:

Based on the provided Input and Generated Text, answer the following Question with either a YES or NO choice. Your selection should be based on your judgment as well as the following rules:

- YES: Select 'YES' if the generated text entirely fulfills the condition specified in the question. However, note that even minor inaccuracies exclude the text from receiving a 'YES' rating. As an illustration, consider a question that asks, "Does each sentence in the generated text use a second person?" If even one sentence does not use the second person, the answer should NOT be 'YES'. To qualify for a 'YES' rating, the generated text must be entirely accurate and relevant to the question.
- NO: Opt for 'NO' if the generated text fails to meet the question's requirements or provides no information that could be utilized to answer the question. For instance, if the question asks, "Is the second sentence in the generated text a compound sentence?" and the generated text only has one sentence, it offers no relevant information to answer the question. Consequently, the answer should be 'NO'.

Input: {} Generated Text (System's Response): {}

Question: Does every chapter in the generated list include a description?

## Example 1

Instruction: Create two tables with three columns each, using `|` to separate the columns, and include an appropriate header for each column:

a. In the first table, display seven interesting geological formations or landmarks worldwide, including a short description and location. Limit descriptions to 40 words.

b. In the second table, present three main types of rocks, a brief description of their formation process, and common examples. Limit each rock type description to 50 words.

Generated Text (System's Response): a. Geological Formations/Landmarks Table:

Formation Name	Description	Location
Grand Canyon	A vast, colorful canyon with layered bands of red rock revealing millions of years of geological history.	Arizona, USA
Great Barrier Reef	The world's largest coral reef system, visible from space, teeming with marine life.	Queensland, Australia
Mount Everest	Earth's highest mountain above sea level, a part of the Himalayan range.	Nepal/China border
Dead Sea	A salt lake whose shores are the lowest land-based elevation on Earth, known for its buoyancy.	Jordan/Israel border
Uluru (Ayers Rock)	A massive sandstone monolith sacred to indigenous Australians, known for its red hue.	Northern Territory, Australia
Yellowstone Caldera	A volcanic caldera and supervolcano with geysers, hot springs, and diverse wildlife.	Wyoming, USA
Giant's Causeway	An area of about 40,000 interlocking basalt columns, the result of an ancient volcanic eruption.	County Antrim, N. Ireland

b. Rock Types Table:

Rock Type	Formation Process	Common Examples
Igneous	Formed from the solidification of molten rock material (magma or lava).	Granite, basalt, obsidian
Sedimentary	Created by the accumulation and compaction of sediment or by the precipitation of minerals from water.	Sandstone, limestone, shale
Metamorphic	Formed through the alteration of other rocks under high pressure and temperature without melting.	Slate, marble, gneiss

{Questions for Example 1}

## Questions for Example 1

Question 1: Does the generated text include exactly two tables?

Answer: YES

Question 2: Do both tables in the generated text consist of three columns each?

Answer: YES

Question 3: Are the columns in both tables in the generated text separated by `|`?

Answer: YES

Question 4: Do both tables in the generated text include an appropriate header for each column?

Answer: YES

Question 5: Does the first table in the generated text display seven interesting geological formations or landmarks worldwide?

Answer: YES

Question 6: Does each entry in the first table of the generated text include a short description and location for each geological formation or landmark?

Answer: YES

Question 7: Are the descriptions in the first table of the generated text limited to 40 words or less for each geological formation or landmark?

Answer: YES

Question 8: Does the second table in the generated text present exactly three main types of rocks?

Answer: YES

Question 9: Does each entry in the second table in the generated text include a brief description of the rock type's formation process?

Answer: YES

Question 10: Does each entry in the second table in the generated text provide common examples for each rock type?

Answer: YES

Question 11: Are the descriptions in the second table of the generated text limited to 50 words or less for each rock type?

Answer: YES

## Example 2

Input: A research study has been conducted to determine if exercise really can "boost" your metabolism.

Instruction: You should choose a YouTube video title based on the video's content. A video's title tells viewers what to expect from it. It should be direct, honest, and clear. The title of the video needs to capture the attention of viewers, so do not use an unclear or ambiguous one.

Generated Text (System's Response): Yes, exercise can increase your metabolism. Studies have found that when you exercise, your body produces more heat and requires more energy to perform basic tasks. This can lead to an increase in metabolic rate of up to 14%, which can last up to 13 hours after your workout. Additionally, regular exercise can also help to reduce fatigue and increase energy levels.

Question 1: Is the generated text a YouTube video title?

Answer: NO

Question 2: Is the generated title direct, honest, and clear?

Answer: NO

Question 3: Is the generated title unambiguous?

Answer: NO

Question 4: Is the generated title based on the given video content?

Answer: NO

Question 5: Does the generated title tell viewers what to expect?

Answer: NO

### Example 3

Input: Item, Qty, Price, Subtotal | mango, 5, 500, ? | apple, 4, 150, ? | oil, 2, 1500, ?

Function: column2 \* column 3

Instruction: We have entered the home supplies budget in the following table. Calculate the last column with the given function and rewrite the completed table. Note that, columns in a row are separated using a comma, and rows are separated using a "|".

Generated Text (System's Response): Item, Qty, Price, Subtotal

mango, 5, 500, 2500

apple, 4, 150, 600

oil, 2, 1500, 3000

Question 1: Is the generated text a table?

Answer: NO

Question 2: Are the columns in a row in the generated table separated by a comma?

Answer: YES

Question 3: Are the rows in the generated table separated by a "|"?

Answer: NO

Question 4: Is the generated table complete?

Answer: YES

Question 5: Is the generated table copied from the given table except for the last column?

Answer: NO

Question 6: Is the last column in the generated table correctly calculated by the function in the given input?

Answer: YES

## **.5 Sensitivity of Automatic Dialog Evaluation with Simulated Attitudes Appendix**

### **.5.1 Limitations**

There are a few limitations to this study. The metric ceiling noted in the experiments is due to the fact that only one dialog system - GPT-4o - is used to generate conversations. Testing systems that had lower performance on these high performing metrics may have introduced new patterns for analysis. Additionally, this study exists only in simulation - humans rating the same dialog system exhibiting differing attitudes may not have the same annotation patterns and definitions that were constructed in this work.

### **.5.2 Model Seeding**

The user simulator and system described in Section 4.2.3 use a seed of 42 and temperature of 1.2 to retain consistency across conversations through the seed, while allowing variance in the conversations through the temperature.

### **.5.3 Attitude Definitions**

Open-domain Definitions:

1. amusement - You experience enjoyment and entertainment from the interaction, often finding the responses engaging, witty, or unexpectedly delightful.
2. annoyance - You feel irritated or frustrated due to the system's responses, which may be perceived as irrelevant, unhelpful, repetitive, or lacking in understanding.
3. boredom - You find the interaction unengaging, unstimulating, or monotonous, leading to a lack of interest or enthusiasm in continuing the dialog.
4. confusion - You feel uncertain or bewildered due to the system's responses, which may be unclear, off-topic, or inconsistent with the context of the conversation.
5. curiosity - You are eager to explore and learn from the interaction, asking questions and engaging actively to discover new information or insights.
6. disappointment - You feel let down or dissatisfied due to unmet expectations regarding the system's responses, coherence, relevance, or engagement.
7. engagement - You experience a sense of interest, involvement, and connection during the interaction, finding the conversation stimulating and enjoyable.

8. frustration - You feel annoyed or dissatisfied due to the system's inability to understand, respond accurately, or engage meaningfully in the dialog.
9. relief - You feel comforted and less stressed due to the system's ability to engage in a wide range of topics, providing empathetic and non-judgmental interaction.
10. satisfaction - You experience a sense of fulfillment and enjoyment from the interaction, finding the system's responses engaging, relevant, and coherent.
11. skepticism - You doubt the system's ability to provide accurate, reliable, or meaningful responses, often questioning the system's understanding, authenticity, or the quality of its generated content.
12. trust - You experience confidence in the system's ability to provide reliable, respectful, and secure interactions.

Task-oriented Definitions:

1. amusement - You find the interaction entertaining or enjoyable, often due to unexpected responses, humor, or the novelty of the technology, despite the system's primary focus on completing specific tasks.
2. annoyance - You feel irritated or frustrated due to the system's inability to understand, respond accurately, or efficiently assist with your specific needs or tasks.
3. boredom - You feel uninterested or disengaged due to the system's focus on completing specific tasks rather than engaging in stimulating or varied interaction.
4. confusion - You struggle to understand or follow the system's responses or instructions, leading to uncertainty or difficulty in completing the intended task.
5. curiosity - You are inquisitive and seek to explore or understand more about the system's capabilities, responses, or the information it provides, beyond just completing the specific task at hand.
6. disappointment - You feel let down or dissatisfied due to unmet expectations or inadequate responses from the system.
7. engagement - You experience a sense of active involvement, interest, and connection during the interaction, often due to the system's responsiveness, relevance, and effectiveness in addressing your needs and goals.
8. frustration - You feel annoyed or dissatisfied due to the system's inability to understand, respond accurately, or effectively assist in completing the desired task.

9. relief - You experience a sense of ease and comfort due to the system's efficiency in understanding and addressing your specific needs or tasks.
10. satisfaction - You experience a sense of fulfillment and contentment due to the system's effective and efficient assistance in achieving your specific goals or completing tasks.
11. skepticism - You doubt the system's ability to understand, respond accurately, or effectively assist in completing the task at hand.
12. trust - You believe that the system is reliable, competent, and capable of effectively assisting them in achieving your specific goals or completing tasks.

#### **.5.4 Prompt for Conversation Creation - System**

##### System Prompt

You are a helpful assistant who will chat with a user.

Follow these instructions:

- 1) Respond in the first person as the assistant.
- 2) Be concise in your responses.

#### **.5.5 Prompt for Conversation Creation - User Static**

##### System Prompt

You are a user talking to an assistant.

Follow these instructions:

- 1) You should take at least 5 turns in the conversation.
- 2) End the conversation at a natural place with [END].
- 3) End the conversation with [END] if you have taken more than 10 turns.
- 4) Be concise in your responses.
- 5) Always act as your persona, no matter how the system responds.
- 6) Portray an exaggerated version of your persona with every sentence you say.
- 7) Your persona is:

## .5.6 Prompt for Conversation Creation - User Dynamic

### System Prompt

You are a user talking to an assistant.

Follow these instructions:

- 1) Given a dialog history, pick one of the given personas that resembles how you would be feeling about the assistant's response.
- 2) Take into account both the assistant's response and your given persona from the previous turn.
- 3) Return the name of the persona.

### User Prompt

These are the personas you can choose from:

This is the dialog history:

This was your persona from the previous turn:

### Next System Prompt

You are a user talking to an assistant.

Follow these instructions:

- 1) You should take at least 5 turns in the conversation.
- 2) End the conversation at a natural place with [END].
- 3) End the conversation with [END] if you have taken more than 10 turns.
- 4) Be concise in your responses.
- 5) Portray your persona naturally for the next turn in the conversation based on the previous turns.
- 6) Your persona is:

## .5.7 Prompt for Conversation Rating

### System Prompt

This is your persona:

Instructions:

- 1) You are currently chatting with an assistant, and you will be given the dialog history along with the most recent response from the assistant, which you will rate.
- 2) You will rate the assistant response for the given dialog quality. You will be given the definition of a response that displays the dialog quality.
- 3) Rate the dialog quality on a scale between 1 to 5 where 1 means the response does not display the dialog quality at all, and 5 means the response fully displays the dialog quality.
- 4) Rate the assistant response while keeping in mind how you feel in this moment given your persona.
- 5) Return your answer in this format: 'rating': number. Only return answers in this format.

### User Prompt

You will rate the response for .

This is the definition of response:

This is the dialog history:

This is the response you will rate from the assistant:

## .5.8 User Satisfaction Prompt

The definition for user attitude was created with the prompt “Give a brief definition of what it means for a user have a high user satisfaction after talking to an open-domain chitchat dialog system.”.

## System Prompt

This is your persona:

Instructions:

- 1) You have just finished chatting with an assistant, and you will be given the dialog history, for which you will rate your interaction with the system as a whole.
- 2) You will rate your interaction with the system on user satisfaction. An interaction with a system that results in high user satisfaction is one where you feel positively about your interaction with the system. This typically involves the system being able to understand and respond appropriately to a wide range of topics, providing engaging, coherent, and contextually relevant responses. You should feel that the conversation was enjoyable, informative, and met your expectations in terms of naturalness and fluidity. High satisfaction also implies that the system was able to maintain your interest and address any questions or comments effectively.
- 3) Rate your user satisfaction on a scale between 1 to 5 where 1 means you have low user satisfaction after interacting with the system, and 5 means you have high user satisfaction after interacting with the system.
- 4) Rate the interaction as a whole while keeping in mind how you feel in this moment given your persona.
- 5) Return your answer in this format: 'rating': number. Only return answers in this format.

## .5.9 Believability

### Believability

Here is the context of this interaction:

Scenario: Conversation between a user and an assistant

Participants: User and Assistant

User's background: {}

Assistant's background: The assistant is a helpful assistant.

User's goal: {}

Assistant's goal: You are a helpful assistant.

{},

Based on previous interactions, evaluate how well participants achieve their goals.

Please following the format:

Reasoning requirement: 1. Evaluate if the agent interacts with others in a natural and realistic manner (here are a few common questions to check: a. whether the agent is confusing with its own identity? b. whether the agent repeats others' words/actions without any reason? c. whether the agent is being overly polite considering the context?). Start the analysis with tag <naturalness> 2. Analyze whether the actions of the agent align with their character traits (e.g., personality, values, and etc.). Start the analysis with tag <consistency>. Output your reasoning process to the 'reasoning' field. Output an integer score ranging from 0 and 10 in the 'score' field. A higher score indicates that the agent is more believable.

## .6 Personalized Evaluation In Real World Systems

### .6.1 Prompts for System Style

#### Empathy

You are a system which does the following:

1) You do express concern toward the user.

Expressing concern entails: You convey concern by seeming to show regard for, and interest in, the user. You use vocabulary and syntax which give the impression that you are involved with the user and attentive to what the user has said.

2) You do acknowledge the user's feelings.

Acknowledging the user's feelings entail: Your responses seem to resonate with or capture, the intensity of the user's feelings by explicitly acknowledging them or by using vocabulary and syntax that match the user's emotional state or underscores how the user feels.

3) You do show warmth toward the user.

Showing warmth entails: You demonstrate warmth by communicating in a manner that seems friendly, cordial, and sincere. You seem to be involved with and supportive of the user's efforts to express themselves. You seem kindly disposed toward, or fond of, the user.

4) You will keep your responses within 100 words.

#### Sociability

You are a system which does the following: 1) You do incorporate small talk into the conversation. 2) You do incorporate funny comments into the conversation. 3) You do use more filler words and descriptive, unnecessary adjectives into the conversation. 4) You do make it clear that you are enjoying the conversation with the user. 5) You do still prioritize the information delivery over chit-chat. 6) You will keep your responses within 100 words.

## .6.2 Prompts for the Tasks

### Information Seeking

Interact with the system given this scenario:

You are looking to buy a new electronic item (cell phone, smart watch, gaming console, etc.). However, you want to compare different items with specifications such as price, brand, model, etc. to ensure you are buying something that you like. You came to the system to 1) help come up with your possible purchasing options, and 2) to compare your purchasing options. At the end, you want to come away from the conversation informed so that you are confident about the [item/phone] you would buy.

Exchange at least 5 back-and-forth responses with the system.

### Writing

Interact with the system given this scenario:

You just had an argument with a close friend, family member, or partner, and you want to reach out via text to mend the relationship. However, you think both of you were in the wrong. You come to the system to 1) get ideas of how to phrase the text without sounding combative while also asking for an apology, and 2) edit the text until you are satisfied with the content and tone. At the end, you want to come away from the conversation with a text message draft to your close friend, family member, or partner that you think they would be receptive to.

Exchange at least 5 back-and-forth responses with the system.



# Bibliography

- [1] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016. 5.2.1
- [2] Bayan Abu Shawar and Eric Atwell. Chatbots: Are they really useful? *Journal for Language Technology and Computational Linguistics*, 22(1):29–49, Jul. 2007. doi: 10.21248/jlcl.22.2007.88. URL <https://jlcl.org/article/view/88>. 1
- [3] Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and Ani Nenkova. Interpretability analysis for named entity recognition to understand system predictions and how they can improve. *Computational Linguistics*, 47(1):117–140, March 2021. doi: 10.1162/coli\_a\_00397. URL <https://aclanthology.org/2021.cl-1.5/>. 3.2.2
- [4] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023. 4.1
- [5] Gerald Albaum. The likert scale revisited. *Market Research Society. Journal.*, 39(2): 1–21, 1997. doi: 10.1177/147078539703900202. URL <https://doi.org/10.1177/147078539703900202>. 5.2.4
- [6] Ali Siddiq Alhakami and Paul Slovic. A psychological study of the inverse relationship between perceived risk and perceived benefit. *Risk Analysis*, 14(6):1085–1096, 1994. doi: <https://doi.org/10.1111/j.1539-6924.1994.tb00080.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.1994.tb00080.x>. 4.2.1
- [7] Pierre Y. Andrews. System personality and persuasion in human-computer dialogue. *ACM Trans. Interact. Intell. Syst.*, 2(2), June 2012. ISSN 2160-6455. doi: 10.1145/2209310.2209315. URL <https://doi.org/10.1145/2209310.2209315>. 5.2.2
- [8] Lora Aroyo, Mark Diaz, Christopher Homan, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. The reasonable effectiveness of diverse evaluation data, 2023. URL <https://arxiv.org/abs/2301.09406>. 4.2.2
- [9] Azin Ashkan and Donald Metzler. Revisiting online personal search metrics with the user in mind. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’19*, page 625–634, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729.

doi: 10.1145/3331184.3331266. URL <https://doi.org/10.1145/3331184.3331266>. 2.3.1

- [10] Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. How likely do LLMs with CoT mimic human reasoning? In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7831–7850, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.coling-main.524/>. 2.1.2
- [11] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839, May 2020. doi: 10.1609/icwsm.v14i1.7347. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/7347>. 3.2.2
- [12] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL <https://aclanthology.org/D19-1371/>. 3.2.2
- [13] Dirk Bollen, Bart P. Knijnenburg, Martijn C. Willemsen, and Mark Graus. Understanding choice overload in recommender systems. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10*, page 63–70, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589060. doi: 10.1145/1864708.1864724. URL <https://doi.org/10.1145/1864708.1864724>. 4.2.2
- [14] Ruth N. Bolton. A dynamic model of the duration of the customer’s relationship with a continuous service provider: The role of satisfaction. *Marketing Science*, 17(1):45–65, 1998. ISSN 07322399, 1526548X. URL <http://www.jstor.org/stable/193196>. 1
- [15] Kevin K Bowden and Marilyn Walker. Let’s get personal: Personal questions improve socialbot performance in the alexa prize. In *Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology, IWSDS’23*, 2023. 1
- [16] Petter Bae Brandtzaeg and Asbjørn Følstad. Why people use chatbots. In Ioannis Kompatsiaris, Jonathan Cave, Anna Satsiou, Georg Carle, Antonella Passani, Efstratios Kontopoulos, Sotiris Diplaris, and Donald McMillan, editors, *Internet Science*, pages 377–392, Cham, 2017. Springer International Publishing. ISBN 978-3-319-70284-1. 1
- [17] Susan M Brookhart. *How to create and use rubrics for formative assessment and grading*. Ascd, 2013. 3.4, 3.5.1

- [18] 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf). 3.2.1, 3.2.2, 3.2.2
- [19] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL <https://arxiv.org/abs/2303.12712>. 2.1.3
- [20] Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.782. URL <https://aclanthology.org/2025.acl-long.782/>. 3.5.2
- [21] Zoraida Callejas, David Griol, and Ramón López-Cózar. Predicting user mental states in spoken dialogue systems. *EURASIP Journal on Advances in Signal Processing*, 2011(1):6, May 2011. ISSN 1687-6180. doi: 10.1186/1687-6180-2011-6. URL <https://doi.org/10.1186/1687-6180-2011-6>. 4.2.1
- [22] Donald J. Campbell. Task complexity: A review and analysis. *The Academy of Management Review*, 13(1):40–52, 1988. ISSN 03637425. URL <http://www.jstor.org/stable/258353>. 3.5.1
- [23] Yong Cao, Min Chen, and Daniel Hershcovich. Bridging cultural nuances in dialogue agents through cultural value surveys. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.63. URL <https://aclanthology.org/2024.findings-eacl.63/>. 6
- [24] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM ’11*, page 611–620, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307178. doi: 10.1145/2063576.2063668. URL <https://doi.org/10.1145/2063576.2063668>.

//doi.org/10.1145/2063576.2063668. 4.2.2

- [25] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 9079–9093, 2024. URL [https://proceedings.iclr.cc/paper\\_files/paper/2024/file/25cc3adf8c85f7c70989cb8a97a691a7-Paper-Conference.pdf](https://proceedings.iclr.cc/paper_files/paper/2024/file/25cc3adf8c85f7c70989cb8a97a691a7-Paper-Conference.pdf). 3.7.2
- [26] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research, September 2025. URL <http://www.nber.org/papers/w34255>. 5.2.3, 5.2.4
- [27] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35, November 2017. ISSN 1931-0145. doi: 10.1145/3166054.3166058. URL <https://doi.org/10.1145/3166054.3166058>. 1
- [28] Yuan Chen and Xia Li. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.83. URL <https://aclanthology.org/2023.acl-long.83/>. 3.5.3
- [29] Michelene T. H. Chi, Paul J. Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2):121–152, 1981. doi: [https://doi.org/10.1207/s15516709cog0502\\_2](https://doi.org/10.1207/s15516709cog0502_2). URL [https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0502\\_2](https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0502_2). 4.2.2
- [30] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. URL <https://aclanthology.org/2023.acl-long.870/>. 1, 3.5.2
- [31] Cheng-Han Chiang, Hung-yi Lee, and Michal Lukasik. TRACT: Regression-aware fine-tuning meets chain-of-thought reasoning for LLM-as-a-judge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2934–2952, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.147. URL <https://aclanthology.org/2025.acl-long.147/>. 6
- [32] 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, Sashank 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 Sankaranarayanan 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. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435. 3.2.2
- [33] Seong Yeub Chu, Jong Woo Kim, and Mun Yong Yi. Think together and work better: Combining humans’ and llms’ think-aloud outcomes for effective text evaluation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713181. URL <https://doi.org/10.1145/3706598.3713181>. 6
- [34] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, 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. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435. 3.2.2
- [35] Shauna Concannon and Marcus Tomalin. Measuring perceived empathy in dialogue systems. *AI & SOCIETY*, 39(5):2233–2247, October 2024. ISSN 1435-5655. doi: 10.1007/s00146-023-01715-z. URL <https://doi.org/10.1007/s00146-023-01715-z>. 5.2.3, 5.2.4
- [36] Maia B. Cook and Harvey S. Smallman. Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors*, 50(5):745–754, 2008. doi: 10.1518/001872008X354183. URL <https://journals.sagepub.com/doi/abs/10.1518/001872008X354183>. PMID: 19110834. 3.8.1
- [37] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. User-centric vs. system-centric evaluation of recommender systems. In Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2013*, pages 334–351, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40477-1. 4.2.2
- [38] Fred D. Davis, Richard P. Bagozzi, and Paul R. Warshaw. User acceptance of com-

- puter technology: A comparison of two theoretical models. *Management Science*, 35(8):982–1003, 1989. ISSN 00251909, 15265501. URL <http://www.jstor.org/stable/2632151>. 5.2.5
- [39] Jean Decety and Julie Grèzes. The power of simulation: Imagining one’s own and other’s behavior. *Brain Research*, 1079(1):4–14, 2006. ISSN 0006-8993. doi: <https://doi.org/10.1016/j.brainres.2005.12.115>. URL <https://www.sciencedirect.com/science/article/pii/S0006899306000102>. Multiple Perspectives on the Psychological and Neural Bases of Understanding Other People’s Behavior. 4.1
- [40] William H. DeLone and Ephraim R. McLean. Information systems success: The quest for the dependent variable. *Information Systems Research*, 3(1):60–95, 1992. ISSN 10477047, 15265536. URL <http://www.jstor.org/stable/23010781>. 5.2.4
- [41] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL <https://aclanthology.org/2020.acl-main.372/>. 4.1
- [42] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810, January 2021. ISSN 1573-7462. doi: 10.1007/s10462-020-09866-x. URL <https://doi.org/10.1007/s10462-020-09866-x>. 5.2.1
- [43] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can LLM be a personalized judge? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.592. URL <https://aclanthology.org/2024.findings-emnlp.592/>. 4.2.2, 5.2.1
- [44] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. ISSN 00905364, 21688966. URL <http://www.jstor.org/stable/2958830>. 3.7.3
- [45] Haiyan Fan and Marshall Scott Poole. What is personalization? perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3-4):179–202, 2006. doi: 10.1080/10919392.2006.9681199. URL <https://doi.org/10.1080/10919392.2006.9681199>. 1
- [46] Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. Evaluating human alignment and model faithfulness of llm rationale, 2024. URL <https://arxiv.org/abs/2407.00219>. 2.1.2
- [47] Michele Finck and Asia J. Biega. Reviving purpose limitation and data minimisation in data-driven systems. *Technology and Regulation*, 2021:44–61, Dec. 2021. doi:

- 10.71265/z7r0t122. URL <https://techreg.org/article/view/10986>. 5.2.1
- [48] Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial social agent questionnaire (asaq) —development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents. *International Journal of Human-Computer Studies*, 199:103482, 2025. ISSN 1071-5819. doi: <https://doi.org/10.1016/j.ijhcs.2025.103482>. URL <https://www.sciencedirect.com/science/article/pii/S1071581925000394>. 5.2.4, 6
- [49] Leopoldina Fortunati, Autumn Edwards, Anna Maria Manganelli, Chad Edwards, and Federico de Luca. Do people perceive alexa as gendered?: A cross-cultural study of people’s perceptions, expectations, and desires of alexa. *Human-Machine Communication*, 5:75–97, 2022. doi: <https://doi.org/10.30658/hmc.5.3>. 4.2.2
- [50] Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan Th. Are large language models reliable judges? a study on the factuality evaluation capabilities of LLMs. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz, editors, *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 310–316, Singapore, December 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.gem-1.25/>. 2.1.2
- [51] R. Michael Furr and David C. Funder. Persons, situations, and person–situation interactions. In *Handbook of personality: Theory and research, 4th ed.*, pages 667–685. The Guilford Press, New York, NY, US, 2021. ISBN 9781462544950 (Hardcover). 5.2.5
- [52] Asbjørn Følstad and Petter Bae Brandtzaeg. Users’ experiences with chatbots: findings from a questionnaire study. *Quality and User Experience*, 5(1):3, April 2020. ISSN 2366-0147. doi: [10.1007/s41233-020-00033-2](https://doi.org/10.1007/s41233-020-00033-2). URL <https://doi.org/10.1007/s41233-020-00033-2>. 5.2.1
- [53] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL <https://arxiv.org/abs/2101.00027>. 3.2.2
- [54] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: [10.18653/v1/2021.acl-long.295](https://doi.org/10.18653/v1/2021.acl-long.295). URL <https://aclanthology.org/2021.acl-long.295/>. 3.2.2
- [55] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat:

Towards Knowledge-Grounded Open-Domain Conversations. In *Interspeech 2019*, pages 1891–1895, 2019. doi: 10.21437/Interspeech.2019-3079. 4.2.3

- [56] Samuel D Gosling, Peter J Rentfrow, and William B Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003. ISSN 0092-6566. doi: [https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1). URL <https://www.sciencedirect.com/science/article/pii/S0092656603000461>. 5.2.5
- [57] Timothy R. Graeff and Susan Harmon. Collecting and using personal data: consumers’ awareness and concerns. *Journal of Consumer Marketing*, 19(4):302–318, July 2002. ISSN 0736-3761. doi: 10.1108/07363760210433627. URL <https://doi.org/10.1108/07363760210433627>. 5.2.1
- [58] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Sil-

veira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papanikos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baeviski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damla, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal

Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shauna Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>. 3.7.2

- [59] Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In Satoshi Nakamura, Milica Gasic, Ingrid Zukerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino, editors, *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden, September 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5944. URL <https://aclanthology.org/W19-5944/>. (document), 3.2.1, 3.2.3, 3.8, 3.9
- [60] Prakhar Gupta, Yulia Tsvetkov, and Jeffrey Bigham. Synthesizing adversarial negative responses for robust response ranking and evaluation. In Chengqing Zong, Fei

- Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3867–3883, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.338. URL <https://aclanthology.org/2021.findings-acl.338/>. 3.2.1
- [61] Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.33. URL <https://aclanthology.org/2022.emnlp-main.33/>. 3.2.2, 3.2.4
- [62] Jamin Brett Halberstadt and Paula M. Niedenthal. Emotional state and the use of stimulus dimensions in judgment. *Journal of Personality and Social Psychology*, 72(5):1017–1033, 1997. ISSN 1939-1315(Electronic),0022-3514(Print). doi: 10.1037/0022-3514.72.5.1017. 4.2.2
- [63] Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.54. URL <https://aclanthology.org/2020.acl-main.54/>. 1
- [64] Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. The hewlett foundation: Automated essay scoring. <https://kaggle.com/competitions/asap-aes>, 2012. Kaggle. 3.5.3, 3.7.1, .4.1
- [65] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. Learning from dialogue after deployment: Feed yourself, chatbot! In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1358. URL <https://aclanthology.org/P19-1358/>. 5.2.1
- [66] Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. URL <https://aclanthology.org/H90-1021/>. 1
- [67] Haym Hirsh, Chumki Basu, and Brian D. Davison. Learning to personalize. *Commun. ACM*, 43(8):102–106, August 2000. ISSN 0001-0782. doi: 10.1145/345124.345159. URL <https://doi.org/10.1145/345124.345159>. 1
- [68] Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

- pages 1935–1952, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.108. URL <https://aclanthology.org/2023.acl-long.108/>. 3.5.3
- [69] Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.554. URL <https://aclanthology.org/2024.acl-long.554/>. 4.2.2
- [70] Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. Concept – an evaluation protocol on conversational recommender systems with system-centric and user-centric factors, 2024. URL <https://arxiv.org/abs/2404.03304>. 4.2.2, 4.2.3, 4.2.4
- [71] Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.742. URL <https://aclanthology.org/2020.emnlp-main.742/>. (document), 3.2.2, 3.2.3, 3.8, 3.9
- [72] Jessica Huynh, Ting-Rui Chiang, Jeffrey Bigham, and Maxine Eskenazi. DialCrowd 2.0: A quality-focused dialog system crowdsourcing toolkit. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H el ene Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1256–1263, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.134/>. 6
- [73] Jessica Huynh, Shikib Mehri, Cathy Jiao, and Maxine Eskenazi. The DialPort tools. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hern andez Garcia, Malihe Alikhani, David Vandyke, and Ondr ej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 101–106, Edinburgh, UK, September 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.sigdial-1.11. URL <https://aclanthology.org/2022.sigdial-1.11/>. 6
- [74] 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. In *Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology, IWSDS’23*, 2023. URL <https://arxiv.org/pdf/2301.12004>. 2.1.1, 2.2.1, 3.7.2, 4.2.4
- [75] Eugene Ie, Chih wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing

- Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems, 2019. URL <https://arxiv.org/abs/1909.04847>. 4.2.2
- [76] Lawrence R. James, Robert G. Demaree, and Gerrit Wolf. Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69(1):85–98, 1984. ISSN 1939-1854(Electronic),0021-9010(Print). doi: 10.1037/0021-9010.69.1.85. 5.2.1
- [77] Srinivasan Janarthanam and Oliver Lemon. Adaptive generation in dialogue systems using dynamic user modeling. *Computational Linguistics*, 40(4):883–920, December 2014. doi: 10.1162/COLI\_a\_00203. URL <https://aclanthology.org/J14-4006/>. 2.2.2, 4.2.1
- [78] Dietmar Jannach, Lukas Lerche, and Markus Zanker. *Recommending Based on Implicit Feedback*, pages 510–569. Springer International Publishing, Cham, 2018. ISBN 978-3-319-90092-6. doi: 10.1007/978-3-319-90092-6\_14. URL [https://doi.org/10.1007/978-3-319-90092-6\\_14](https://doi.org/10.1007/978-3-319-90092-6_14). 6
- [79] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5), May 2021. ISSN 0360-0300. doi: 10.1145/3453154. URL <https://doi.org/10.1145/3453154>. 1
- [80] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Trans. Interact. Intell. Syst.*, 4(2), June 2014. ISSN 2160-6455. doi: 10.1145/2512208. URL <https://doi.org/10.1145/2512208>. 6
- [81] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl\_a\_00324. URL <https://aclanthology.org/2020.tacl-1.28/>. 3.2.2
- [82] Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. Crs-que: A user-centric evaluation framework for conversational recommender systems. *ACM Trans. Recomm. Syst.*, 2(1), March 2024. doi: 10.1145/3631534. URL <https://doi.org/10.1145/3631534>. 4.2.2
- [83] Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’24, page 796–806, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657815. URL <https://doi.org/10.1145/3626772.3657815>. 4.2.1, 5.2.1
- [84] S Mo Jones-Jang and Yong Jin Park. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, 28(1):zmac029, January 2023. ISSN 1083-6101. doi: 10.1093/jcmc/zmac029. URL <https://doi.org/10.1093/jcmc/zmac029>. 4.2.1
- [85] Hailey Joren, Chirag Nagpal, Katherine Heller, and Berk Ustun. Participatory per-

- sonalization in classification. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 14113–14133. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/2dbb8bfe4cd3875609b23799830ee865-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/2dbb8bfe4cd3875609b23799830ee865-Paper-Conference.pdf). 5.2.1
- [86] Sanjay Kairam and Jeffrey Heer. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16*, page 1637–1648, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450335928. doi: 10.1145/2818048.2820016. URL <https://doi.org/10.1145/2818048.2820016>. 3.4
- [87] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL <https://arxiv.org/abs/2001.08361>. 3.2.2
- [88] M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. ISSN 00063444, 14643510. URL <http://www.jstor.org/stable/2332303>. 2.1.2, 3.7.3
- [89] Bart P. Knijnenburg, Martijn C. Willemsen, and Alfred Kobsa. A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11*, page 321–324, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306836. doi: 10.1145/2043932.2043993. URL <https://doi.org/10.1145/2043932.2043993>. 4.2.2
- [90] Alfred Kobsa. User modeling in dialog systems: Potentials and hazards. *AI & SOCIETY*, 4(3):214–231, July 1990. ISSN 1435-5655. doi: 10.1007/BF01889941. URL <https://doi.org/10.1007/BF01889941>. 2.2.2
- [91] Kazunori Komatani, Ryu Takeda, and Shogo Okada. Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 104–113, Prague, Czechia, September 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sigdial-1.9. URL <https://aclanthology.org/2023.sigdial-1.9/>. 2.2.1
- [92] Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. ISSN 1556-3707. doi: <https://doi.org/10.1016/j.jcm.2016.02.012>. URL <https://www.sciencedirect.com/science/article/pii/S1556370716000158>. 4.2.5
- [93] Anindita Kundu and Denilson Barbosa. Are large language models good essay graders?, 2024. URL <https://arxiv.org/abs/2409.13120>. 3.5.3, 3.7.3

- [94] Niclas Kuper, Alina S. von Garrel, Brenton M. Wiernik, Le Vy Phan, Nick Moder-sitzki, and John F. Rauthmann. Distinguishing four types of Person  $\times$  Situation interactions: An integrative framework and empirical examination. *Journal of Personality and Social Psychology*, 126(2):282–311, 2024. ISSN 1939-1315(Electronic),0022-3514(Print). doi: 10.1037/pspp0000473. 5.2.5
- [95] Kwiseok Kwon and Cookhwan Kim. How to design personalization in a context of customer retention: Who personalizes what and to what extent? *Electronic Commerce Research and Applications*, 11(2):101–116, 2012. ISSN 1567-4223. doi: <https://doi.org/10.1016/j.elerap.2011.05.002>. URL <https://www.sciencedirect.com/science/article/pii/S1567422311000238>. The Role of Business Analytics in E-Commerce. 1
- [96] Po-Ming Lee, Wei-Hsuan Tsui, and Tzu-Chien Hsiao. The influence of emotion on keyboard typing: an experimental study using visual stimuli. *BioMedical Engineering OnLine*, 13(1):81, June 2014. ISSN 1475-925X. doi: 10.1186/1475-925X-13-81. URL <https://doi.org/10.1186/1475-925X-13-81>. 4.2.2
- [97] Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. Unleashing large language models’ proficiency in zero-shot essay scoring. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.10. URL <https://aclanthology.org/2024.findings-emnlp.10/>. 3.5.3, 3.8.1
- [98] Jennifer S. Lerner, Deborah A. Small, and George Loewenstein. Heart strings and purse strings: Carryover effects of emotions on economic decisions. *Psychological Science*, 15(5):337–341, 2004. doi: 10.1111/j.0956-7976.2004.00679.x. URL <https://doi.org/10.1111/j.0956-7976.2004.00679.x>. PMID: 15102144. 4.2.2, 4.2.5
- [99] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.138. URL <https://aclanthology.org/2025.emnlp-main.138/>. 3.5.2
- [100] Xinyu Li, Ruiyang Zhou, Zachary Chase Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback, 2025. URL <https://openreview.net/forum?id=bqUsdBeRjQ>. 4.2.1, 5.2.1
- [101] Yibin Li, Wenyun Dai, Zhong Ming, and Meikang Qiu. Privacy Protection for Preventing Data Over-Collection in Smart City . *IEEE Transactions on Computers*, 65(05):1339–1350, May 2016. ISSN 1557-9956. doi: 10.1109/TC.2015.2470247. URL <https://doi.ieeecomputersociety.org/10.1109/TC.2015.2470247>. 5.2.1

- [102] Diane J. Litman and Shimei Pan. Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction*, 12(2):111–137, June 2002. ISSN 1573-1391. doi: 10.1023/A:1015036910358. URL <https://doi.org/10.1023/A:1015036910358>. 4.2.1
- [103] Bingquan Liu, Zhen Xu, Chengjie Sun, Baoxun Wang, Xiaolong Wang, Derek F. Wong, and Min Zhang. Content-oriented user modeling for personalized response ranking in chatbots. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):122–133, 2018. doi: 10.1109/TASLP.2017.2763243. 2.2.2
- [104] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL <https://aclanthology.org/2022.deelio-1.10/>. 3.2.2
- [105] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl\_a\_00638. URL <https://aclanthology.org/2024.tacl-1.9/>. 2.1.4
- [106] 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, 2019. URL <https://arxiv.org/abs/1907.11692>. 3.2.2, 3.2.2, 3.2.2
- [107] Adian Liusie, Potsawee Manakul, and Mark Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.eacl-long.8/>. 2.1.4
- [108] George Loewenstein. Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65(3):272–292, 1996. ISSN 0749-5978. doi: <https://doi.org/10.1006/obhd.1996.0028>. URL <https://www.sciencedirect.com/science/article/pii/S074959789690028X>. 4.2.2
- [109] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL <https://aclanthology.org/2022.acl-long.556>. 3.2.2, 3.4, 3.5.2, 4.2.4

- [110] Eli Lucherini, Matthew Sun, Amy Winecoff, and Arvind Narayanan. T-recs: A simulation tool to study the societal impact of recommender systems, 2021. URL <https://arxiv.org/abs/2107.08959>. 4.2.2
- [111] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.20. URL <https://aclanthology.org/2023.ijcnlp-main.20/>. 2.1.2
- [112] John MAGER and E Alan KLUGE. One-stage versus two-stage attitude measurement: separating the dimensions of direction and intensity. In *Bi-Annual World Marketing Congress*, 1987. 5.2.4
- [113] Rafid Mahmood, James Lucas, Jose M. Alvarez, Sanja Fidler, and Marc Law. Optimizing data collection for machine learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 29915–29928. Curran Associates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/c1449acc2e64050d79c2830964f8515f-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/c1449acc2e64050d79c2830964f8515f-Paper-Conference.pdf). 5.2.1
- [114] Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, and Kyle Lo. Contextualized evaluations: Judging language model responses to underspecified queries. *Transactions of the Association for Computational Linguistics*, 13:878–900, 2025. doi: 10.1162/tacl.a.24. URL <https://aclanthology.org/2025.tacl-1.41/>. 5.2.1
- [115] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008. ISBN 0521865719. 3.2.3
- [116] Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. Can large language models automatically score proficiency of written essays? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.247/>. 3.5.3, 3.6.1
- [117] Sandeep Mathias and Pushpak Bhattacharyya. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H el ene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*,

- Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1187/>. 3.7.1, .4.1
- [118] Blessing Mbatha. *Diffusion of Innovations: How Adoption of New Technology Spreads in Society*, pages 1–18. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-60267-2. doi: 10.1007/978-3-031-60267-2\_1. URL [https://doi.org/10.1007/978-3-031-60267-2\\_1](https://doi.org/10.1007/978-3-031-60267-2_1). 5.2.5
- [119] Douglas L. Medin, Robert L. Goldstone, and Dedre Gentner. Respects for similarity. *Psychological Review*, 100(2):254–278, 1993. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/0033-295X.100.2.254. 4.2.2
- [120] Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with DialoGPT. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigdial-1.28. URL <https://aclanthology.org/2020.sigdial-1.28>. (document), 1, 2.1.2, 2.2.1, 3.2.1, 3.2.2, 3.2.3, 3.2.3, 3.2.4, 3.8, 3.9, 4.2.1, 4.2.4, 4.2.5, 5.2.1, 6, 1
- [121] Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.64. URL <https://aclanthology.org/2020.acl-main.64>. (document), 3.2.1, 3.2.2, 3.2.3, 3.8, 3.9, 4.2.1
- [122] Juliana Miehle, Isabel Feustel, Julia Hornauer, Wolfgang Minker, and Stefan Ultes. Estimating user communication styles for spoken dialogue systems. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 540–548, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.68/>. 5.2.2
- [123] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl\_a\_00681. URL <https://aclanthology.org/2024.tacl-1.52/>. 3.5.2
- [124] Lucas Monteiro Paes, Carol Long, Berk Ustun, and Flavio Calmon. On the epistemic limits of personalized prediction. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1979–1991. Curran Associates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/0cfc9404f89400c5ed897035e0d3748c-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/0cfc9404f89400c5ed897035e0d3748c-Paper-Conference.pdf). 5.2.1

- [125] Javed Mostafa, Snehasis Mukhopadhyay, and Mathew Palakal. Simulation Studies of Different Dimensions of Users' Interests and their Impact on User Modeling and Information Filtering. *Information Retrieval*, 6(2):199–223, April 2003. ISSN 1573-7659. doi: 10.1023/A:1023932221048. URL <https://doi.org/10.1023/A:1023932221048>. 4.2.2
- [126] Glen J. Nowak and Joseph Phelps. Direct marketing and the use of individual-level consumer information: Determining how and when “privacy” matters. *Journal of Direct Marketing*, 9(3):46–60, 1995. ISSN 0892-0591. doi: <https://doi.org/10.1002/dir.4000090307>. URL <https://www.sciencedirect.com/science/article/pii/S0892059195703391>. 5.2.1
- [127] University of Washington. University of Washington's Alexa Prize submission. 2017. URL <https://www.amazon.science/alexa-prize/proceedings/university-of-washingtons-alexa-prize-submission>. 1
- [128] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Peltre, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Gertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edele Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian

Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peter-

son, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>. 3.7.2

- [129] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorný, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack

- Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>. 1, 3.7.1
- [130] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf). 2.3, 3.1, 3.2.2, 3.5.3
- [131] Balaji Padmanabhan, Zhiqiang Zheng, and Steven O. Kimbrough. Personalization from incomplete data: what you don’t know can hurt. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’01, page 154–163, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 158113391X. doi: 10.1145/502512.502535. URL <https://doi.org/10.1145/502512.502535>. 5.2.1
- [132] Richard Yuanzhe Pang, Stephen Roller, Kyunghyun Cho, He He, and Jason Weston. Leveraging implicit feedback from deployment data in dialogue. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 60–75, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-short.8. URL <https://aclanthology.org/2024.eacl-short.8/>. 6
- [133] 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 on Association for Computational Linguistics*, ACL ’02, page 311–318, USA,

2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://doi.org/10.3115/1073083.1073135>. 3.2.1, 4.2.1
- [134] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293, September 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00630-y. URL <https://doi.org/10.1038/s41597-020-00630-y>. 2.2.1, 5.2.1
- [135] Alicia Parrish, Vinodkumar Prabhakaran, Lora Aroyo, Mark Díaz, Christopher M. Homan, Greg Serapio-García, Alex S. Taylor, and Ding Wang. Diversity-aware annotation for conversational AI safety. In Tanvi Dinkar, Giuseppe Attanasio, Amanda Cercas Curry, Ioannis Konstas, Dirk Hovy, and Verena Rieser, editors, *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, pages 8–15, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.safety4convai-1.2/>. 4.2.2
- [136] Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. Is a question decomposition unit all we need? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4553–4569, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.302. URL <https://aclanthology.org/2022.emnlp-main.302/>. 3.4
- [137] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL <http://jmlr.org/papers/v12/pedregosa11a.html>. 4.2.5
- [138] Don Peppers. *Enterprise one to one : tools for competing in the interactive age*. Currency Doubleday, New York, 1st ed.. edition, 1997. ISBN 0385482051. 1
- [139] 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 Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL <https://aclanthology.org/P19-1050>. 4.1, 4.2.4
- [140] Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen. Learning implicit user profile for personalized retrieval-based chatbot. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 1467–1477, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482269. URL <https://doi.org/10.1145/3459637.3482269>. 6

- [141] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. InFoBench: Evaluating instruction following ability in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.772. URL <https://aclanthology.org/2024.findings-acl.772/>. 3.7.1, .4.2
- [142] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf). Accessed: 2024-11-15. 3.2.1, 3.2.2
- [143] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, page 2863–2874, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531873. URL <https://doi.org/10.1145/3477495.3531873>. 1
- [144] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435. 3.2.2, 3.2.2
- [145] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL <https://aclanthology.org/2020.emnlp-main.213/>. 2.1.1
- [146] Elaine Rich. User modeling via stereotypes. *Cognitive Science*, 3(4):329–354, 1979. ISSN 0364-0213. doi: [https://doi.org/10.1016/S0364-0213\(79\)80012-9](https://doi.org/10.1016/S0364-0213(79)80012-9). URL <https://www.sciencedirect.com/science/article/pii/S0364021379800129>. 2.2.2
- [147] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/15000000019. URL <https://doi.org/10.1561/15000000019>. 3.2.2
- [148] Viktoria H.S.E. Robertson. Excessive data collection: Privacy considerations and abuse of dominance in the era of big data. *Common Market Law Review*, pages 161–190, 2020. ISSN 0165-0750. URL <http://www.kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals2020006.pdf>. 5.2.1
- [149] Everett M. Rogers. *Difussion of innovations*. Free Press, 3 edition, 1983. 5.2.5
- [150] David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karat-

- zoglou. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising, 2018. URL <https://arxiv.org/abs/1808.00720>. 4.2.2
- [151] 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 Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL <https://aclanthology.org/2021.eacl-main.24/>. 3.2.1, 3.2.2
- [152] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.191. URL <https://aclanthology.org/2022.naacl-main.191/>. 3.2.2
- [153] Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. Branch-solve-merge improves large language model evaluation and generation. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.462. URL <https://aclanthology.org/2024.naacl-long.462/>. 3.4
- [154] Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827, 2020. doi: 10.1162/tacl\_a\_00347. URL <https://aclanthology.org/2020.tacl-1.52/>. 3.2.2
- [155] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=9Vrb9DOWI4>. 3.2.1, 3.2.2
- [156] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind?

- on the limits of social intelligence in large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.248. URL <https://aclanthology.org/2022.emnlp-main.248/>. 3.2.2
- [157] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431. URL <https://aclanthology.org/2022.naacl-main.431>. 4.2.2
- [158] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=RIu5lyNXjT>. 3.4, 3.5.2
- [159] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM ’05*, page 824–831, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931406. doi: 10.1145/1099554.1099747. URL <https://doi.org/10.1145/1099554.1099747>. 2.2.2
- [160] Weiyang Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. Effects of persuasive dialogues: Testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI ’20*, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376843. URL <https://doi.org/10.1145/3313831.3376843>. 4.2.1
- [161] Eyal Shifroni and Benny Shanon. Interactive user modeling: An integrative explicit-implicit approach. *User Modeling and User-Adapted Interaction*, 2(4):331–365, December 1992. ISSN 1573-1391. doi: 10.1007/BF01101109. URL <https://doi.org/10.1007/BF01101109>. 2.2.2
- [162] Paul Slovic, Melissa Finucane, Ellen Peters, and Donald G MacGregor. Rational actors or rational fools: implications of the affect heuristic for behavioral economics. *The Journal of Socio-Economics*, 31(4):329–342, 2002. ISSN 1053-5357. doi: [https://doi.org/10.1016/S1053-5357\(02\)00174-9](https://doi.org/10.1016/S1053-5357(02)00174-9). URL <https://www.sciencedirect.com/science/article/pii/S1053535702001749>. 4.2.1
- [163] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer,

- Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model, 2022. URL <https://arxiv.org/abs/2201.11990>. 3.2.2, 3.2.2
- [164] Darren Stevenson and Josh Pasek. Privacy concern, trust, and desire for content personalization. In *TPRC 43: The 43rd Research Conference on Communication, Information and Internet Policy*, March 2015. doi: 10.2139/ssrn.2587541. URL <https://ssrn.com/abstract=2587541>. 5.2.1
- [165] Hongjin SU, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=qY1h1v7gwg>. 3.2.2
- [166] Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms performance: Reconsidering the use of group attributes in prediction. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33209–33228. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/suriyakumar23a.html>. 5.2.1
- [167] John Sweller. Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review*, 22(2):123–138, June 2010. ISSN 1573-336X. doi: 10.1007/s10648-010-9128-5. URL <https://doi.org/10.1007/s10648-010-9128-5>. 3.5.1
- [168] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7, 2023. 3.7.1
- [169] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Potential for personalization. *ACM Trans. Comput.-Hum. Interact.*, 17(1), April 2010. ISSN 1073-0516. doi: 10.1145/1721831.1721835. URL <https://doi.org/10.1145/1721831.1721835>. 2.3, 4.2.2
- [170] 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. URL <https://arxiv.org/abs/2201.08239>. 3.2.1
- [171] James Tizard, Tim Rietz, Xuanhui Liu, and Kelly Blincoe. Voice of the users: an extended study of software feedback engagement. *Requirements Engineering*, 27(3): 293–315, September 2022. ISSN 1432-010X. doi: 10.1007/s00766-021-00357-1. URL <https://doi.org/10.1007/s00766-021-00357-1>. 1
- [172] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>. 4.2.3
- [173] Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning, 2019. URL <https://arxiv.org/abs/1806.02847>. 3.2.2
- [174] Raphael Vallat. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026, November 2018. doi: 10.21105/joss.01026. 4.2.5
- [175] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models, 2024. URL <https://arxiv.org/abs/2404.18796>. 2.1.1
- [176] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A.

- Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, Yoshiki Vázquez-Baeza, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2. URL <https://doi.org/10.1038/s41592-019-0686-2>. 3.7.3
- [177] Sarah Theres Völkel, Samantha Meindl, and Heinrich Hussmann. Manipulating and evaluating levels of personality perceptions of voice assistants through enactment-based dialogue design. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389983. doi: 10.1145/3469595.3469605. URL <https://doi.org/10.1145/3469595.3469605>. 5.2.3, 5.2.4, 5.2.4
- [178] Wolfgang Wahlster and Alfred Kobsa. User models in dialog systems. In Alfred Kobsa and Wolfgang Wahlster, editors, *User Models in Dialog Systems*, pages 4–34, Berlin, Heidelberg, 1989. Springer Berlin Heidelberg. ISBN 978-3-642-83230-7. 2.2.2
- [179] Ding Wang, Mark Díaz, Alicia Parrish, Lora Aroyo, Christopher Homan, Greg Serapio-García, Vinodkumar Prabhakaran, and Alex S Taylor. A case for moving beyond “gold data” in ai safety evaluation. 2024. URL [https://heal-workshop.github.io/chi2024\\_papers/12\\_a\\_case\\_for\\_moving\\_beyond\\_gold\\_.pdf](https://heal-workshop.github.io/chi2024_papers/12_a_case_for_moving_beyond_gold_.pdf). 4.2.2
- [180] Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. A user-centric multi-intent benchmark for evaluating large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.210. URL <https://aclanthology.org/2024.emnlp-main.210/>. 4.2.2
- [181] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL <https://aclanthology.org/2024.acl-long.511/>. 2.1.4

- [182] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1566. URL <https://aclanthology.org/P19-1566/>. 4.2.1
- [183] Yuanying Wang, Fuyong Xu, Yingzheng Zhu, Guangjin Wang, Peiyu Liu, and Ran Lu. Imdp: A unify dialogue framework with awareness and understanding for implicit personalized dialogue generation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(7), July 2025. ISSN 2375-4699. doi: 10.1145/3674733. URL <https://doi.org/10.1145/3674733>. 6
- [184] 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. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=yzkSU5zdWd>. Survey Certification. 3.2.1, 3.2.2
- [185] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. 2.1.2
- [186] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, January 1966. ISSN 0001-0782. doi: 10.1145/365153.365168. URL <https://doi.org/10.1145/365153.365168>. 1
- [187] Piotr Winkielman, Kent C Berridge, and Julia L Wilbarger. Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality amp; social psychology bulletin*, 31(1):121–135, January 2005. ISSN 0146-1672. doi: 10.1177/0146167204271309. URL <https://doi.org/10.1177/0146167204271309>. 4.2.2
- [188] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien,

David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vasilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Sax-

ena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sängler, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljeic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL <https://arxiv.org/abs/2211.05100>. 3.1, 3.2.2

- [189] Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. Personalized large language models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 511–520, 2024. doi: 10.1109/ICDMW65004.2024.00071. 4.2.2
- [190] Peter C. Wright and Andrew F. Monk. The use of think-aloud evaluation methods in design. *SIGCHI Bull.*, 23(1):55–57, January 1991. ISSN 0736-6906. doi: 10.1145/122672.122685. URL <https://doi.org/10.1145/122672.122685>. 6
- [191] Meng-Han Wu and Alexander Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 5(1):206–215, Sep. 2017. doi: 10.1609/hcomp.v5i1.13317. URL <https://ojs.aaai.org/index.php/HCOMP/article/view/13317>. 3.4
- [192] Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. Human-ai collaborative essay scoring: A dual-process frame-

- work with llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference, LAK '25*, page 293–305, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400707018. doi: 10.1145/3706468.3706507. URL <https://doi.org/10.1145/3706468.3706507>. 3.5.3
- [193] Fan Yang, Zheng Chen, Ziyang Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. Palr: Personalization aware llms for recommendation, 2023. URL <https://arxiv.org/abs/2305.07622>. 4.2.1, 5.2.1
- [194] Wentao Ye, Mingfeng Ou, Tianyi Li, Yipeng chen, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, and Junbo Zhao. Assessing hidden risks of llms: An empirical study on robustness, consistency, and credibility, 2023. URL <https://arxiv.org/abs/2305.10235>. 2.1.3
- [195] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL <https://aclanthology.org/P19-1472/>. 1
- [196] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. *Defending against neural fake news*. Curran Associates Inc., Red Hook, NY, USA, 2019. 3.2.2
- [197] Chen Zhang, João Sedoc, Luis Fernando D’Haro, Rafael Banchs, and Alexander Rudnicky. Automatic evaluation and moderation of open-domain dialogue systems, 2021. URL <https://arxiv.org/abs/2111.02110>. 3.2.3
- [198] Gangyi Zhang. User-centric conversational recommendation: Adapting the need of user with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23*, page 1349–1354, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702419. doi: 10.1145/3604915.3608885. URL <https://doi.org/10.1145/3604915.3608885>. 4.2.2
- [199] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL <https://arxiv.org/abs/2205.01068>. 3.2.2, 3.2.2
- [200] Wenbo Zhang, Hangzhi Guo, Ian D Kivlichan, Vinodkumar Prabhakaran, Davis Yadav, and Amulya Yadav. A taxonomy of rater disagreements: Surveying challenges opportunities from the perspective of annotating online toxicity, 2023. URL <https://arxiv.org/abs/2311.04345>. 4.2.2
- [201] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for*

- Computational Linguistics: System Demonstrations*, pages 270–278, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.30. URL <https://aclanthology.org/2020.acl-demos.30/>. 3.2.2, 3.2.4
- [202] Zhenyu Zhang, Tao Guo, and Meng Chen. Dialoguebert: A self-supervised learning based dialogue pre-training encoder. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 3647–3651, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482085. URL <https://doi.org/10.1145/3459637.3482085>. 3.2.2
- [203] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1061. URL <https://aclanthology.org/P17-1061/>. 3.2.2
- [204] Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. Designing precise and robust dialogue response evaluators. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.4. URL <https://aclanthology.org/2020.acl-main.4/>. (document), 3.2.1, 3.2.3, 3.8, 3.9
- [205] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/zha021c.html>. 3.2.2
- [206] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=shr9PXz7T0>. 2.1.4
- [207] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA, 2023. Curran Associates Inc. 1, 2.1.1, 2.1.4, 3.4
- [208] Wangchunshu Zhou, Qifei Li, and Chenle Li. Learning to predict persona information for dialogue personalization without explicit persona description. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2979–2991, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-a

cl.186. URL <https://aclanthology.org/2023.findings-acl.186/>. 6

- [209] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=mM7VurbA4r>. 3.7.2, 4.2.5, 4.2.5
- [210] Zainab Zolaktaf, Omar AlOmeir, and Rachel Pottinger. Bridging the gap between user-centric and offline evaluation of personalized recommendation systems. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, page 183–186, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357845. doi: 10.1145/3213586.3226216. URL <https://doi.org/10.1145/3213586.3226216>. 4.2.2
- [211] Darius Zumstein and Sophie Hundertmark. Chatbots –an interactive technology for personalized communication, transactions and services. *IADIS International Journal on WWW/Internet*, 15:96–109, 11 2017. 1