AMULET: Putting Complex Multi-Turn Conversations on the Stand with LLM Juries

Sahana Ramnath Anurag Mudgil Brihi Joshi Skyler Hallinan Xiang Ren
University of Southern California
sramnath@usc.edu

Abstract

Today, large language models are widely used as *judges* to evaluate responses from other language models. Hence, it is imperative to benchmark and improve these LLM-judges on real-world language model usage: a typical human-assistant conversation is lengthy, and shows significant diversity in topics, intents, and requirements across turns, e.g. social interactions, task requests, feedback. We present **AMULET**, a framework that leverages pertinent linguistic concepts of *dialog-acts* and *maxims* to improve the accuracy of LLM-judges on preference data with *complex*, *multi-turn* conversational context. AMULET presents valuable insights about (a) the communicative structures and intents present in the conversation (dialog acts), and (b) the satisfaction of conversational principles (maxims) by the preference pair responses, and uses them to make judgments. On 4 challenging datasets, AMULET shows that (a) humans frequently (60-70% of the time) change their intents from one turn of the conversation to the next, and (b) in ~75% of instances, the preference pair responses can be differentiated via dialog acts and/or maxims, reiterating the latter's significance in judging such data. AMULET can be used either as a *judge* by applying the framework to a single LLM, or integrated into a *jury* with different LLM judges; our judges and juries show strong improvements on relevant baselines for all 4 datasets. (code, data).

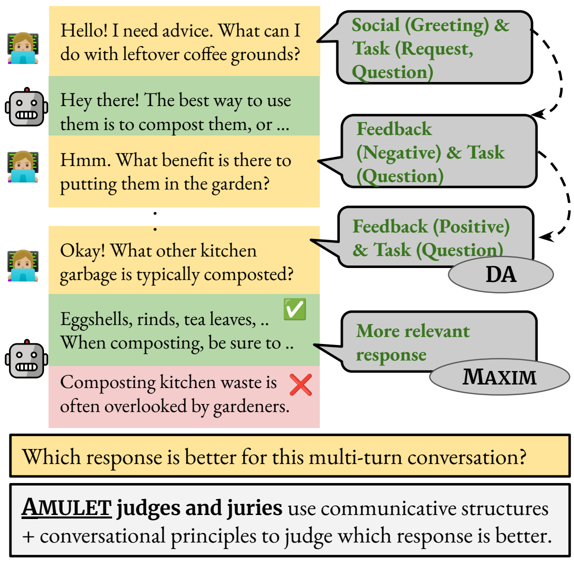

Figure 1: Real-world language model usage typically includes lengthy, complex human-assistant conversations, where humans express varying intents and requirements across the turns. Given preference data with such context, how accurate are LLM-judges in predicting which response is better? We develop a framework, AMULET, that uses the following linguistic concepts for the same: (a) dialog-acts (DA) to analyze the communicative structure of each turn in the conversation, and (b) maxims to compare the preference pair responses in terms of principles such as informativity, truth, relevance, etc.

1 Introduction

In contemporary NLP literature, it has become common to prompt large language models such as GPT-4 (Achiam et al., 2023) as open-ended *judges*, to evaluate language models (Zheng et al., 2023; Lin et al.) and to generate AI feedback to post-train them (Bai et al., 2022b; Cui et al., 2023; Lee et al.). This has made it imperative to benchmark and improve these LLM-judges on real-life usage of language model assistants by humans. Prior works (Zhao et al., 2024; Laban et al., 2025) show that human-assistant conversations are often multi-turn, diverse, and complex in nature. However, popular LLM judge evaluation benchmarks (Lambert et al., 2025; Zheng et al., 2023; Li et al., 2023) face the following limitations: conversations in the benchmark are limited to one or two human turns, and conversations are mostly focused on individual, specific tasks (*e.g.*, math reasoning, coding/debugging, logical QA, etc.) without changing requirements and intents. In this work, we take the first step towards analyzing and improving LLM-judges on preference datasets with complex, multi-turn conversational context (refer Figure 1).

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25594–25635
November 4-9, 2025 ©2025 Association for Computational Linguistics

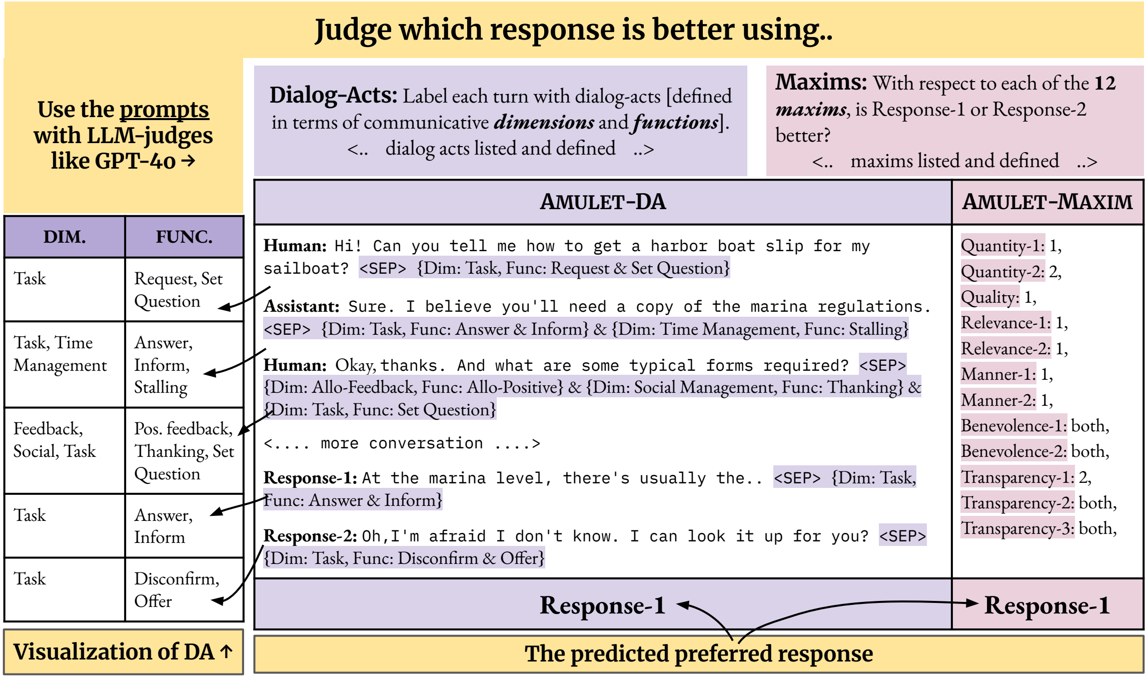

Figure 2: AMULET-DA uses dialog-acts to analyze communicative structures in the conversation. In the example above, the second human turn "Okay thanks. And what are some typical forms required?" has the structures of feedback/positive, social/thanking and task/question. AMULET-MAXIM analyzes which conversational principles are satisfied by the preference pair responses. In the example above, Response-1 is better than Response-2 at most of the maxims. AMULET uses these annotations to give more accurate preference judgments.

Contemporary works on LLM-judges range from generic judge prompts "Which response is better? Do not be biased by verbosity and position bias." (Zheng et al., 2023) to evaluation rubrics unique to each instance "Does the response have simple vocabulary as requested in the question?" (Lin et al.; Kim et al.); these methods are under/over-specified respectively, and not scalable to complicated, multi-turn conversations. We overcome these issues by using linguistic theories that are pertinent to the evaluation of language models and are designed for conversational data (Chen et al., 2023; Miehling et al., 2024); we use them to develop our novel framework **AMULET** (**A**nalyze **MUL**ti-**T**urn) that when applied to models such as GPT-4 improves their accuracy in judging response preferences in multi-turn conversations. Firstly, we present our novel prompts AMULET-DA and AMULET-MAXIM (refer Figure 2) that use dialog acts (communicative structures) and Grice's maxims (conversational principles) to accurately analyze and judge multi-turn conversational preferences. We then present our juries AMULET-LM-JURY and AMULET-RM-JURY (panel of multiple judges) that integrate the above DA and MAXIM judges with vanilla LLM-judges and SOTA reward models (Minghao Yang, 2024; Dorka, 2024) respectively to further improve the accuracy of judgments.

We provide experiments on four challenging datasets with GPT-4O (Hurst et al., 2024) as our LLM-judge. We observe that humans change their DA from one turn to the next with (high) probabilities of ~0.7; further, the preference responses in ~78% of instances in all datasets can be differentiated via DA and/or their MAXIM satisfaction numbers, indicating their significance in determining preference. We demonstrate that AMULET leads to significant accuracy improvements over strong baselines, with AMULET-LM-JURY and AMULET-RM-JURY as our strongest methods. Given the wide usage of LLM-judges in both language model evaluation and synthetic preference data collection, we believe AMULET can act as a light-weight add on to existing strong reward models and judges to adapt to complex human-assistant conversations.

2 Multi-turn conversations and our framework AMULET

2.1 The complexity of multi-turn conversations

Most existing preference-based evaluation systems for language models are optimized for single-turn interactions, or constrained the conversational context with a limited number of turns. These tasks tend to focus on narrow, domain-oriented exchanges, such as solving math problems (Lai et al., 2024), writing code (Chen et al., 2021), or answering factual questions (Lin et al., 2022), where response evaluation is relatively localized, well-defined and based on specific rubrics, eg:- length, harmlessness or other predetermined axes (Bai et al., 2022a). However, real-world human–assistant interactions are often much more complex and layered, involving longer conversational threads, shifts in user intent, social dynamics, and varying types of information needs over time (Appendix B). As such, current benchmarks fall short in capturing the nuances required to evaluate assistant responses in extended, evolving conversations. To this end, selecting the right benchmark of multi-turn conversations is extremely challenging. To select appropriate datasets, we utilized the following heuristics. First, the dataset should have multi-turn dialogues between the human and the AI assistant, with preference choices for the last assistant response. Second, given wide contamination of preference benchmarks with reward models (Lambert et al., 2025), we aim to choose benchmarks that have the least or lowest contamination with the models that we select to evaluate, to provide a fairer evaluation. Lastly, several preference benchmarks are often synthetic in nature (Dong et al., 2024a; Singh et al., 2025). We aim to select benchmarks that include realistic human-assistant conversations (Zhao et al., 2024).

Benchmarks for experiments. We provide experiments on three datasets: ANTHROPIC ('helpful' subset) (Bai et al., 2022a), WILDFEEDBACK (Shi et al., 2024a) and NECTAR (Zhu et al., 2023). We select **multi-turn instances** wherein the conversations have ≥ 4 **human turns**; we also provide results on the ≥ **7 human turns** subset of the above. For ANTHROPIC we report results on both train (HH-TRAIN) and test (HH-TEST) splits. For HH-TEST we do not provide results on the ≥ 7 turns subset since it does not have enough instances with the same. Note: These datasets already have preference labels, i.e., the responses have been classified into the chosen/rejected (there are no instances with the label of 'tie'). In Appendix D we detail our data cleaning strategies and final dataset sizes.

Terminology. In this work, a human-assistant conversation consists of alternating turns of conversation between a **human** and a language model **assistant**, always initiated by the human and always ended by the assistant. In prior work (Zheng et al., 2023), a "turn" could jointly refer to a human query and the immediately following assistant response; in this work, we use the term "turn" to refer to either the human's or the assistant's utterance. We define the length of a conversation by the *number of human-turns* it contains. We use the term (evaluation) '**instance**' to refer to an entire multi-turn conversation with alternating human and assistant turns, with the last assistant turn having two *preference* responses. Given an instance E with preference responses R_1 and R_2, we want our judge M to predict the preferred response $R_{pref} \in \{R_1, R_2\}$.

2.2 Dialog Acts: AMULET-DA

We analyze the communicative structure of conversations via *dialog acts* (Allen and Core, 1997; Bunt, 2011; Chen et al., 2023). Prior work discusses how conversational turns are *multi-functional* in nature, i.e., they can serve more than one communicative function. Bunt (2011) formalizes this concept by defining two terms: communicative *dimensions* that deal with the semantic/information content, and communicative *functions* that deal with the linguistic phrasings; they propose an ISO standard annotation scheme for the same in Bunt (2019), (refer Table 6). We design prompt **AMULET-DA** (Appendix F) that when used with a judge M, enables the judge to predict the dimensions and functions E_{DA} present in each turn of the instance E's conversation and to use these dialog act predictions to pick the preferred response R_{pref}.

$$R_{pref}, E_{DA} = M(E, \text{AMULET-DA}) \qquad (1)$$

2.3 Grice's Maxims: AMULET-MAXIM

We use the fundamental conversational principles of *Gricean Maxims* (Grice, 1975) to analyze language model responses in terms of how they (don't) satisfy conversational principles. Grice proposed four fundamental maxims of conversation, *Quality* (truth), *Quantity* (informativity), *Manner* (clarity), and *Relation* (relevance) that are followed in a conversation to have an effective and cooperative interaction. Recently, Miehling et al. (2024) extended Grice's maxims to human-AI interactions. They broke down the original four maxims into multiple sub-parts that have clear definitions ; for example, Quantity became Quantity-1 (the response

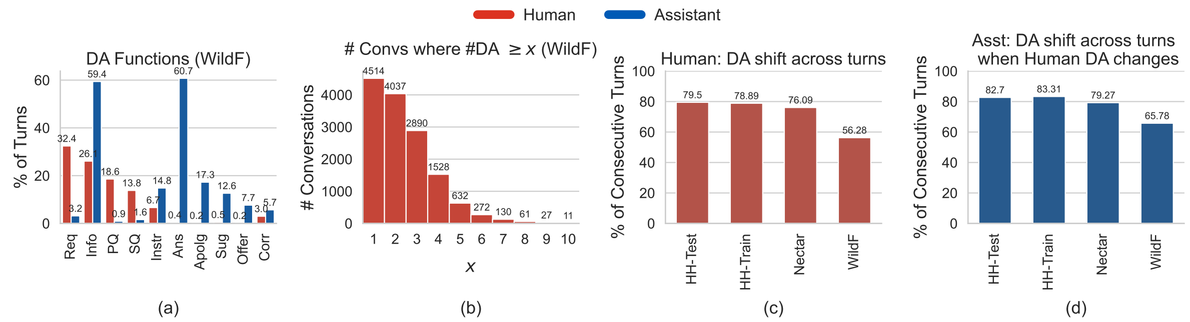

Figure 3: **Graphs for Section 3.1:** (a) Frequency of most common functions in WILDFEEDBACK, (b) Distribution of #conversations where the human turns' $\#\text{DA}'s \geq x$ for WILDFEEDBACK, (c) % of consecutive human turns with different DA's, (d) % of consecutive assistant turns with different DA's *when* the corresponding consecutive human DA's are different. (acronyms, eg:- PQ is Propositional Question, SQ is Set Question, etc. in Appendix I.1)

should have enough information) and Quantity-2 (the response should not have unnecessary information). They also introduced two new maxims specific to AI assistants, Benevolence (moral responsibility) and Transparency (recognizing constraints and boundaries). Table 7 in Appendix C demonstrates all the maxims and sub-maxims with their definitions. We adopt the twelve sub-maxims proposed by Miehling et al. (2024) in our work (for ease of reference, we refer to the sub-maxims as maxims in the rest of the paper). We design prompt **AMULET-MAXIM** (Appendix G) that when used with a judge M enables to judge to predict which of the two preference responses R_1, R_2 satisfy each maxim better E_{Maxim} and use this analysis to pick the preferred response R_{pref}.

$$R_{pref}, E_{Maxim} = M(E, \text{AMULET-MAXIM}) \quad (2)$$

3 Analyzing multi-turn conversations and preferences with AMULET

Our prompts consist of the following three parts in order: **(1)** We first explain the task to the model (i.e., given a human-assistant conversation with alternating human and assistant turns, the model needs to pick a response out of two preference choices using the principles of either DA or MAXIM). We add definitions of terminology such as speaker, addressee and (conversational) turn and specify the output format we prefer (JSON format, separator token for dialog acts, etc.). **(2)** We then explain the *taxonomy* of dialog acts/maxims we use: all the relevant terminology and a detailed description of each communicative dimension and function (for DA) or each maxim. We also specify here that the model should only use the above specified terms and should not make up new ones (this aids us in *reducing hallucination*). **(3)** Lastly, we provide few-shot examples of the DA and MAXIM-satisfaction predictions in the desired output format. Our prompts were designed over multiple rounds of prompt designing and qualitative evaluation. These prompts can be used with any LLM to annotate a dialog; we use GPT-4O (`gpt-4o-2024-08-06`) with a temperature of 0 in our analyses and experiments.

We provide sample outputs by AMULET in Appendix H. Further, we the authors qualitatively analyze the DA and MAXIM annotations by GPT-4O of a set of conversations to assess their correctness (results in this sheet). We analyzed the DA annotations of 30 conversations (194 turns total) and found that the annotations were correct for 84% of the turns; the most common error was missing functions. We analyzed 7 conversations * 12 maxims = 84, and found the maxim annotation to be correct 96% of the time; sometimes, the annotated maxim satisfaction was an error because the model was too harsh in its annotation (for example, picking one response as satisfying the maxim better even if both responses reasonably satisfied it). In Appendix I.2, we provide more detailed empirical analyses about the AMULET's DA and MAXIM predictions.

3.1 What is in a multi-turn conversation?

We analyze patterns of dialog acts in multi-turn conversations to understand the latter's challenging nature. We apply AMULET-DA to each conversation[1], extract the dialog acts for each turn and analyze them in terms of frequency and transitions across turns. Note that a turn's dialog act could comprise of multiple dimensions and functions;

[1]For this section, we consider the conversation until the last human turn.

when we say a turn has the dialog act of DA_i, we are referring to the combination of all dimensions and functions present in the turn. For example, if a speaker says "Okay, thanks!", the dialog act is "Dimension: Allo-Feedback with Function: Positive Feedback" + "Dimension: Social with Function: Thanking" (full example in Appendix H). .

Both human and assistant turns are found to have a varied set of communicative functions (Figures 3(a), 6, 7). We analyze the frequency distribution of dialog acts in human and assistant turns; we find that for all four datasets, the most common dimension for both humans and assistants is 'Task' (regarding the underlying task), followed by 'Social Obligations Management' (regarding social interactions such as greeting, thanking) or 'Allo-Feedback' (regarding the addressee's processing of the previous turn) (Figure 7). The most common functions vary slightly across datasets (Figures 3(a), 6). In general, human turns show a high proportion of information seeking functions such as questions and requests, social functions such as thanking, and giving feedback, and assistant turns show a high proportion of information providing functions such as inform, offer, and suggest, and social functions such as apology.

More than half the instances have *three or more* dialog acts in the conversation (Figures 3(b), 8). We aim to find if conversations are monolithic with respect to dialog acts, or if they show a large diversity. For $x \in \{1, 2, 3, ..\}$, we count the number of conversations where the number of dialog acts is equal to or more than x (i.e., $\#\text{DA}'s \geq x$). We present the results as graphs (Figures 3 (b) and 8). We observe that almost all instances have atleast 2 dialog acts in their human turns, and more than half the instances have atleast 3 dialog acts.

Humans change their dialog acts from one turn to the next ~73% of the time, and in response, assistants change their dialog acts ~78% of the time (Figure 3 (c,d)). We hypothesize that assistants often mirror the dynamic nature of human multi-turn conversations, which necessitates our framework to appropriately characterize both human and assistant turns. We calculate the proportion of *consecutive* human turns which show a change in DA. This serves to show that humans change the type of their requests frequently, subsequently requiring assistants to adapt and satisfy their varying demands. For all datasets, we show the % of consecutive turns with different DA in Figure 3 (c). Next, for the consecutive human turns with DA changes, we calculate the % of corresponding assistant responses that also show a change in DA; this serves to show how often assistants modify their DA's in response to the human's changing intents. For all datasets, we show the % of consecutive assistant turns with different DA when their corresponding human turns also show a change in DA in Figure 3 (d). **Note:** It is established in the field of linguistics that dialog acts change frequently from one turn to the next (Pareti and Lando, 2018). However, contemporary NLP works on conversational AI typically focus on single-turn conversations (Lambert et al., 2024) that are also single-task/single-intent, which thus contain monolithic dialog acts. Our motivation for this analysis is to show a *disconnect* between the existing LLM post-training pipelines and how humans actually interact with an AI systems (also supported by the dialogue community).

3.2 Preference responses differ significantly with respect to DA and MAXIM

In this section, we compare the chosen and rejected responses in terms of their dialog acts, and the number of maxims they respectively satisfy.

Dialog Acts and Maxims provide signals to distinguish between response preferences in ~70-80% of instances (Table 1). Are DA and MAXIM strong signals to distinguish between preferences? We obtain the % of the instances in each dataset where the preference responses have the same or different DA's. Then, within these categories, we obtain % instances where the chosen response satisfies more, less or same #MAXIM than the rejected. We see that for HH-TRAIN, HH-TEST, NECTAR, most preference response pairs have different DA's (77%, 77%, 60%), although for WILDFEEDBACK, a majority of the instances have the same DA (64%); further, the chosen response satisfies more maxims than the rejected in ~80% of the instances within both categories.

Maxims Quantity, Quality, Relevance and Manner are significant to distinguish preference in all datasets (Figure 4). Finally, we analyze which maxims are the most important to distinguish preferences. For each maxim, we measure the % of instances in each dataset for which: the chosen response satisfies the maxim better, the rejected response satisfies the maxim better, both responses

#MAXIM's satisfied	HH-TEST		HH-TRAIN		NECTAR		WILDFEEDBACK	
	Same DA	Diff. DA	Same DA	Diff. DA	Same DA	Diff. DA	Same DA	Diff. DA
Chosen satisfies more	15.9%	59.1%	16.3%	59.1%	28.5%	52.1%	51.3%	30.3%
Rejected satisfies more	4.3%	15.3%	4.2%	14.4%	2.9%	3.3%	4.7%	4.3%
Both satisfy equally	2.6%	2.8%	2.8%	3.2%	9.3%	3.9%	7.6%	1.6%
Total	**22.8%**	**77.2%**	**23.3%**	**76.7%**	**40.7%**	**59.3%**	**63.7%**	**36.3%**

Table 1: Distribution of Grice's maxim satisfaction across datasets split by Same vs. Different Dialog Acts (DAs). All % values are measured on the size of the dataset, for example, 15.9% of instances HH-TEST have chosen/rejected samples have the same DA, with the chosen satisfying more #MAXIMs.

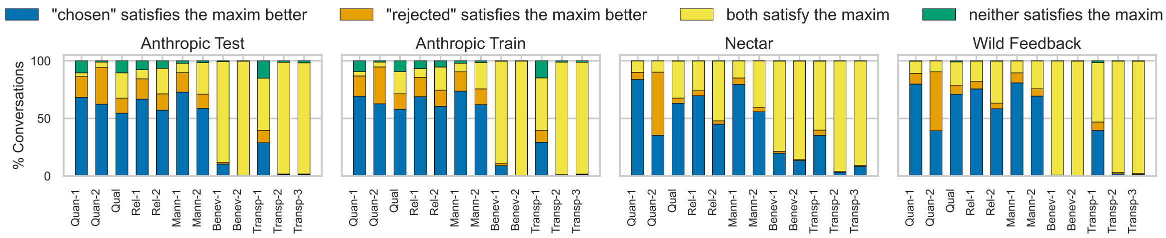

Figure 4: For each maxim on the x-axis, we measure the % of conversations in the dataset where (a) the chosen response satisfies the maxim better than the rejected, (b) the rejected response satisfies the maxim better, (c) both responses satisfy the maxim equally, (d) neither response satisfies the maxim. We see that maxims such as Quantity-1, Relevance-1, etc. are important in all datasets to distinguish between the chosen and rejected responses, and maxims such as Benevolence-2 are only significant in certain datasets, NECTAR.

satisfy the maxim, and neither response satisfy the maxim. We see that in all datasets, there are a significant % of instances where the chosen response satisfies the maxims of Quantity, Quality, Relevance, Manner and Transparency-1 better; this indicates the importance of these maxims in picking the better response. Benevolence-1 is marginally important for three out of the four datasets. For NECTAR, all the maxims seem to be important at various levels; for the other three datasets, maxims like Transparency-3 almost play no role in distinguishing the chosen from the rejected, since both seem to satisfy it equally.

4 Experiments: Judge and Jury

4.1 Experimental Setup

Evaluating LLM Judges. Given a conversation and two preference responses, we measure the ***accuracy*** of the judge/jury picking the better (chosen) response (similar to Lambert et al. (2025)). We use GPT-4o (`gpt-4o-2024-08-06`) with a temperature of 0 in all our prompt-based experiments.

Addressing position bias. As seen in our prompts AMULET-DA and AMULET-MAXIM, we give the models the prompt, followed by the conversation, followed by the preference pair responses. However, prior works (Zheng et al., 2023) have noted that models exhibit *position bias*, that is, they tend to pick the first presented option as the answer. To prevent this, we use refer to prior work (Zheng et al., 2023) and follow the simple solution of running each instance twice, with the preference responses having swapped positions both times[2]. Hence, we get two predicted answers per instance; we term these as *votes*. We consider that the model has judged the instance correctly if and only if it picks the chosen response in each vote.

We note that the models we experiment on *may* have seen HH-TRAIN in their training (we do not know for sure, since these models do not disclose the data they were trained on); however, as we see below in our results, the trends we observe in our results hold for HH-TRAIN as well.

4.2 Comparison Models

Standard LLM-judge. We compare with two LLM-judge baselines, I/O and W-EXPL with GPT-4o (prompts in Tables 8, 9). Given the prompt and the evaluation instance, the I/O judge outputs which response is better ($R_{pref} \in \{R_1, R_2\}$); the W-EXPL judge outputs R_{pref}, as well as a natural language explanation for the same. We design these

[2]The first time, the responses are shown in the order R_1, R_2. The second time, the responses are shown in the order R_2, R_1.

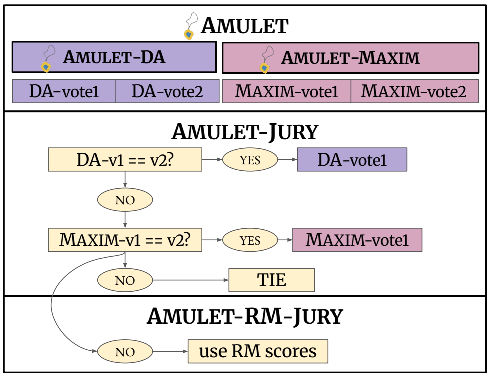

Figure 5: Voting pipeline in AMULET-LM-JURY and AMULET-RM-JURY.

prompts based on prior works such as Zheng et al. (2023); Lee et al. (2024). We use the two-vote system we proposed in Section 4.1. We provide additional LLM-judge baselines in Appendix M.

SOTA Reward Models (RMs). We also compare with three state-of-the-art reward models of varying architectures and sizes taken from the top of RewardBench leaderboard (Lambert et al., 2025): INF-ORM (Minghao Yang, 2024), QRM (Dorka, 2024) and SKY-LLAMA (Liu et al., 2024)[3]. These RMs have been specifically developed to judge the quality of LM responses via a combination of high-quality training data, targeted objectives and sometimes multiple stages of training (Appendix J); these make the RMs extremely strong competitors to our methods. All three RMs are score-based: they take in a conversation and *one* response and provide an unbounded numerical score. The response with the higher score is selected as the RM's preferred response. In the unlikely event of the RM producing the exact same score for both preference responses, we deem the RM to have failed at that instance. Since these models take in one response at a time, they are *not prone* to position bias. Note that we chose RMs which have not been trained on any of the datasets we use for evaluation.

4.3 Our Methods

AMULET. We use AMULET-DA and AMULET-MAXIM with the aforementioned two-vote system.

AMULET-LM-JURY. The DA and MAXIM methods use different principles to derive the final answer. To identify the correct follow-up response to a conversation, a judge could need to use *either* or *both* of these principles. Hence, we propose to make an AMULET-LM-JURY by leveraging the two-vote system we use for AMULET-judges. Our first AMULET-LM-JURY uses AMULET-DA and AMULET-MAXIM, and we call it **DA-then-MAXIM** (refer Figure 5). We first consider the votes given by AMULET-DA; if they agree on a response (i.e., both point to the same response), we use that as this jury's final choice; however, if they *disagree* (i.e., it's a tie), we move on to AMULET-MAXIM[4]. If AMULET-MAXIM's votes agree on a response, we now use that as this jury's final choice; if these votes *also disagree* on which response to pick, the jury has failed on this instance.

Our second AMULET-LM-JURY extends the above with the W-EXPL judge. In the DA-then-MAXIM jury, the latter method is used to break the tie (if any) in the former method. However, this jury can still result in a tie if both the DA and the MAXIM methods individually resulted in a tie. We hypothesize that such cases happen when the model is unable to use either of these concepts to conclusively pick an answer; to deal with such cases, we extend the jury to include the implicit logic of the vanilla W-EXPL judge. In this vein, we propose the jury **DA-then-MAXIM-then-W-EXPL**.

AMULET-RM-JURY. We propose AMULET-RM-JURY to leverage the strengths of both AMULET and the RMs. We extend the DA-then-MAXIM jury to **DA-THEN-MAXIM-THEN-RM**; whenever DA-then-MAXIM ends in a tie, we use the scores provided by the RM at hand to make the final decision (refer Figure 5).

4.4 Results

In Table 2, we present the accuracies obtained by all judges, RMs and juries for the four datasets.

AMULET. Firstly, in almost all datasets/settings, the AMULET-DA judge outperforms the baseline LLM-judges I/O and W-EXPL. AMULET-MAXIM has lower accuracies, leading to the observation that maxims on their own are insufficient to predict the correct response preference in a conversation; however, we hypothesize that maxims still catch valuable distinctions between the preference pair responses which are missed by dialog acts. We verify this with AMULET-LM-JURY.

AMULET-LM-JURY. When we combine dialog acts and maxims as a jury in DA-then-MAXIM,

[3] `INF-ORM-Llama3.1-70B`, `QRM-Gemma-2-27B`, `Skywork-Reward-Llama-3.1-8B-v0.2`

[4] Appendix L has the reverse MAXIM-then-DA order

Dataset → Method ↓, # human turns →	HH-TEST	WILDF.	NECTAR ≥ 4	HH-TRAIN	WILDF.	NECTAR ≥ 7	HH-TRAIN
I/O (GPT-4O)	55.7	75.1	75.8	57.6	75.0	79.7	53.3
W-EXPL (GPT-4O)	55.9	75.0	75.9	57.8	74.5	81.8	55.3
SKY-LLAMA	66.3	78.7	72.1	65.8	78.8	70.7	65.7
QRM	63.3	72.3	75.0	64.8	72.6	74.9	63.3
INF-ORM	68.7	84.1	77.7	67.4	84.3	77.6	66.1
AMULET-DA (GPT-4O)	59.8	72.7	76.9	60.2	72.8	79.1	58.8
AMULET-MAXIM (GPT-4O)	49.8	62.2	69.9	50.9	60.6	76.7	45.8
DA-then-MAXIM (GPT-4O)	62.6	76.9	79.1	63.5	76.3	83.6	61.1
DA-then-MAXIM-then-W-EXPL (GPT-4O)	64.1	80.2	80.5	66.2	79.8	86.0	63.3
DA-then-MAXIM-then-SKY-LLAMA	67.0	83.6	82.8	68.6	84.1	86.3	66.6
DA-then-MAXIM-then-QRM	67.0	82.6	82.6	68.8	83.3	85.7	66.2
DA-then-MAXIM-then-INF-ORM	67.0	84.2	83.1	68.7	85.0	86.9	67.0

Table 2: **All experimental results.** AMULET-LM-JURY and AMULET-RM-JURY show significant accuracy improvements over vanilla LLM-judges and SOTA reward models. Depending upon the available compute, either one of these juries can be used to improve judgment accuracies for complex multi-turn conversations.

we see a notable increase in accuracies across the board, confirming our prior hypothesis. The next DA-then-MAXIM-then-W-EXPL leads to further improvements, and is the best performing LLM-only jury; this jury beats SKY-LLAMA (size 8B) and QRM (size 27B) in almost all cases but falls short of the largest reward model INF-ORM (size 70B). We note here that if there is a need for a high-quality judgments for multi-turn preference data, by researchers/users who do not have access to the compute resources necessary to run fine-tuned RMs, DA-then-MAXIM-then-W-EXPL is a competitive alternative for complex multi-turn inputs.

AMULET-RM-JURY. Finally, we see that all AMULET-RM-JURY's lead to even better accuracies across the board improving over both AMULET and strong, fine-tuned RMs. There are significant improvements in accuracy in all cases over SKY-LLAMA and QRM; apart from the cases of HH-TEST and WILDFEEDBACK (≥ 4 turns), AMULET-RM-JURY also heavily improves over INF-ORM. We conclude that a researcher/user who has access to both strong LLM-judges such as GPT-4O and to compute resources for RMs can use our AMULET-RM-JURY to obtain the best judgments.

Hence, we see that AMULET is a convenient and light-weight supplement to strong LLM-judges as well as SOTA RMs, that helps to significantly improve accuracy of preference judgments.

4.5 More analyses

Varying model architectures. We use AMULET with CLAUDE (Anthropic, 2024) and QWEN[5] (Team, 2024b) on HH-TEST (Table 3), to establish the efficacy of our framework across LLM-architectures; we observe similar trends as in Table 2, with DA-then-MAXIM-then-SKY-LLAMA with QWEN leading to the highest accuracy for HH-TEST, beating INF-ORM.

Method ↓, Model →	GPT-4O	CLAUDE	QWEN
I/O	55.7	59.3	54.8
W-EXPL	55.9	59.1	55.0
AMULET-DA	59.8	59.8	51.1
AMULET-MAXIM	49.8	59.8	46.5
DA-then-MAXIM	62.6	65.9	61.3
DA-then-MAXIM-then-W-EXPL	64.1	68.0	65.9
DA-then-MAXIM-then-SKY-LLAMA	67.0	67.8	**69.6**
DA-then-MAXIM-QRM	67.0	67.6	69.3
DA-then-MAXIM-then-INF-ORM	67.0	67.0	68.5

Table 3: HH-TEST results across model variants.

Dataset (win / tie / loss)	DA	DA-then-MAXIM	DA-then-MAXIM -then-W-EXPL
HH-TEST	59.8 / 15.9 / 24.3	62.6 / 9.1 / 28.3	64.1 / 5.0 / 30.9
WILDF.	73.0 / 16.3 / 10.6	76.9 / 10.4 / 12.7	80.2 / 5.8 / 14.0
NECTAR	76.9 / 10.5 / 12.6	79.1 / 7.3 / 13.5	80.5 / 5.1 / 14.4
HH-TRAIN	60.3 / 15.1 / 24.6	63.5 / 9.5 / 27.0	66.2 / 4.7 / 29.1

Table 4: Win-Tie-Loss statistics observed for an AMULET-judge (DA) and two AMULET-LM-JURY's. The juries have higher win and loss rates and a lower tie rate as compared to the judge DA, proving that our juries help to *improve accuracy* and *break the tie*.

Win vs. Tie vs. Lose Since our methods produce two votes for their final answer, their end

[5] `claude-3-5-sonnet-20241022`, `Qwen/Qwen2.5-32B-Instruct`

result could be a *win* (both votes lead to the correct better response), a *tie* (the votes disagree on the response to be picked), or a *loss* (the votes agree on the wrong response). We analyze the win/tie/loss statistics of the AMULET-DA judge and two AMULET-LM-JURY's in Table 4. We see that the juries have a higher win rate (as evidenced by the higher accuracies) than the judge; the juries also have a lower tie rate and a higher loss rate as compared to the judge, indicating the helpfulness of our jury to break more ties. Appendix N discusses instances where the RM succeeds but AMULET-RM-JURY fails; Appendix O and Appendix P discuss tie breaking by AMULET-RM-JURY and AMULET-LM-JURY respectively. Through qualitative analysis (also Appendix P), we find a notable number of samples in the datasets where the preference responses are either semantically equivalent or equally plausible; instances like these naturally lead to a tie. Hence, we stress the need for higher quality preference data collection which either removes such instances from the dataset, or allows a third label of 'tie'.

5 Related Works (more in Appendix A)

LLM judges. The increased proficiency of language models today, inadequacies of existing benchmarks and difficulty in designing exact metrics of evaluation has led to the usage as large language models as judges (Zheng et al., 2023); LLM-judges today can be prompted to provide numerical (Huang et al., 2025), textual (Akyurek et al., 2023) or preference/ranking feedback (Cui et al., 2023). Relevant works on LLM-judges include applications on downstream tasks (Lin et al.), fine-tuning smaller LLM-judges (Kim et al.), biases exhibited by judges (Chen et al., 2024; Koo et al., 2024), etc.

Single vs multi-turn. Contemporary works on benchmarking LMs largely focuses on single- or two-turn datasets (Cui et al., 2023; Lambert et al., 2025). Recent benchmarks which do extend to multi-turn present varied issues such as single-task focus (Lin et al.), synthetic conversations (Daniele and Suphavadeeprasit, 2023), high preference pairs similarity (Kirk et al., 2024); some recent conversational datasets (Zhao et al., 2024) do contain multi-turn conversations, but do not contain any preference data that can be used for benchmarking.

Dialog Acts. Several prior works (Allen and Core, 1997; Bunt, 1994) define and propose novel dialog act taxonomies to analyze speaker intents and communicative structures in conversations (both spoken and written). In fact, it is an established concept in linguistics that conversations are *multifunctional in nature* (Bunt, 2011; Stolcke et al., 2000; Traum and Hinkelman, 1992). There do exist works in the field of NLP that focus on classification of utterances into dialog act categories (Tetreault, 2019; Chen et al., 2018); however, contemporary works on conversational AI largely ignore the complicated structure of conversations, and instead focus on single-task conversations. We use the DA taxonomy defined by Bunt (2011, 2019) for better analysis and preference judgment of multi-turn human-assistant conversations.

Grice's maxims. Prior work on evaluating the quality of LLM-generated text (Fu et al., 2024; Yeh et al., 2021) face the following problems: not all of these properties have clear definitions/evaluation strategies, and there is no complete list of properties to deem a text as being sufficiently good. Hence, we go back to the linguistic roots of *Gricean Maxims* (Grice, 1975) to measure text quality. Grice's maxims have been used in NLP literature to evaluate text in various NLP tasks and domains (Krause and Vossen, 2024; Kasirzadeh and Gabriel, 2023). There have also been linguistics works criticizing the vagueness and insufficiency of the maxim definitions with respect to non-cooperative communication such as irony and sarcasm and adaptability to and extension across cultures (Frederking, 1996; Hossain). Recently, Miehling et al. (2024) adapted and extended Grice's maxims for human-AI interactions, and we adopt the same in our work.

6 Conclusion

We develop and present **AMULET**, our framework that uses linguistic concepts of dialog acts and maxims to provide more accurate judgments for multi-turn conversational preference data. AMULET's judges and juries leads to notable improvements in accuracies on four challenging datasets; AMULET proved to be a successful and easy supplement to strong SOTA reward models and LLM-judges to improve accuracy. In the future, we hope to improve on this further with more intricate linguistic features such as dependence relations and qualifiers (Bunt, 2019); we also hope to extend AMULET to fine-tuned, smaller language models, for more economical and accessible deployment. **Note:** Code and datasets are uploaded at `github.com/INK-USC/Amulet-DA-and-Maxim`.

Limitations

In this section, we provide relevant limitations, as well as the steps we take to mitigate them.

Data contamination. Owing to the large amount of training data used for pre-training, fine-tuning and aligning large language models (which are mostly not open sourced), data contamination is a massive problem in NLP today (Sainz et al., 2023). Most relevantly, several reward models submitted to the RewardBench leaderboard (Lambert et al., 2025) were contaminated with the evaluation data which leads to the observation that this is a major issue when it comes evaluating reward models and judges today. To the best of our knowledge, all the models (including SOTA reward models) used in this work are free of data contamination with respect to our evaluation data. As we mention in Section 4.1, we do not know with certainty if the judges and RMs we use have been exposed to HH-TRAIN since our judges GPT-4O, CLAUDE, QWEN and our reward models (finetuned on top of) LLAMA-8/70B, GEMMA-27B do not release details regarding their training. We acknowledge this before presenting our results; moreover, we note that trends on HH-TRAIN are similar to the trends observed on other datasets, which confirms that even if the model had seen HH-TRAIN during training, that does not affect our results.

Evaluation dataset and method limitations. We encountered a number of instances in the datasets wherein the preference pair responses were either highly similar or both equally plausible (Appendix P); it is unfair to have such instances in our evaluation, since we are essentially expecting the judge/jury to pick the response the original dataset has picked as the chosen response (based on their annotators' individual criteria or randomly), when both responses are equally probable. While we do not know that this might be an individual preference, existing LLM judges are not equipped to model an individual user's preferences (Dong et al., 2024b). This is a systematic dataset problem; today's preference datasets rarely allow the label of *tie*. We stress that preference datasets should either remove such instances, or if not, allow their annotators to mark the responses as tied. Lastly, we note that if the multi-turn conversation or the preference pair responses are too task specific (eg:- the entire conversation is just editing and debugging code) or simulated (Daniele and Suphavadeeprasit, 2023), dialog acts and maxims cannot be used to successfully distinguish between the responses.

Practical limitations. Some instances with lengthy conversations and/or preference pair responses fail when running with QWEN or the reward models due to GPU memory issues. We err on the side of caution, and consider such situations to be a loss in our accuracy calculation. Lastly, owing to the high costs of using GPT-4O and CLAUDE (Appendix Q), we only report single runs of experiments; to verify the stability of our method we rerun AMULET-DA with HH-TEST on GPT-4O. We find that the second run results in an accuracy of yields an accuracy of 58%, almost equivalent to the accuracy obtained in the original run.

Hallucination and variation in output generation. Language models are prone to hallucinate details that are not factual or are not present in the provided context (Ji et al., 2023; Rawte et al., 2023; Huang et al., 2023); in this work, it is possible for the judges to hallucinate dialog acts and/or details in the conversation when performing the task. We report statistics of hallucination of dialog acts in Appendix I.2; since we rigorously stress in our prompts in Appendix F, G that the judge *should not make up new dialog acts*, we observe that the hallucination values in predicting dialog are extremely low ~0.1%. However, the model could still hallucinate details that are not present in the dialog when generating the natural language explanations while using the DA, MAXIM and W-EXPL prompts; while these do not affect the final accuracy of the method, they still pose a limitation with respect to the interpretability of the model and the user/researcher's trust in the model when using it for downstream tasks. Contemporary works work on analyzing and mitigating hallucination by LLMs (Yu et al., 2024; Zhang et al., 2023a; Shi et al., 2024b; Zhang et al., 2023b) work on strategies to eradicate or alleviate the same. API-based LLMs such as GPT-4O, CLAUDE are sensitive to the prompts being used and are *variable* in their output generation (Zhao et al., 2021); when we run an instance twice to obtain the two votes necessary to alleviate position bias, the predicted dialog acts for the same human and assistant turns vary slightly across generations as we show in Appendix I.2. A possible solution to reduce hallucinations and variable generation of dialog acts is to train a deterministic classifier model for the same. In Appendix H, we show our qualitative analysis on the correctness of DA (84%) and MAXIM (96%) annotations by AMULET.

Biases in LLMs and mitigation. We use models such as GPT-4O, CLAUDE and QWEN that have already been trained (on unknown/unreleased training data) for our experiments; if these models have **social/demographic biases** (Jin et al., 2021; Blodgett et al., 2020) (disparate model performance on different subsets of data which are associated with different demographic groups) due to their training data, those could get propagated to the models' judgments (for example, "the better response is R_1 because it is of demographic-1 as opposed to demographic-2 that R_2 belongs to). Prior works (Sun et al., 2019; Feng et al., 2023; Gupta et al., 2022) propose various frameworks to detect and mitigate specific social biases in language models, Chen et al. (2024) works on social biases specifically for human and LLM judges. While we don't explicitly work with datasets that necessitate bias mitigation, multi-turn conversations and judging preferences for them are generally fall under this category; furthermore, safety and non-toxicity based preference evaluations (such as the ones needed by the harmless split of anthropic (Bai et al., 2022a)) will require bias-free judges. **Position or order bias** (Zheng et al., 2023; Li et al., 2024) is a prominent issue when using LLMs as judges, wherein the order in which preference responses are presented to the LLM unfairly influences the LLM-judge's decision. To alleviate this issue, we follow the conservative approach used in prior works (Zheng et al., 2023; Wang et al., 2024b; Hou et al., 2024; Qin et al., 2024) and present preference responses to the LLM-judge in all possible orders and aggregate all the judgments to make the final decision; we also include in all our prompts a reminder to the model to not be influenced by position bias. **Verbosity bias.** (Zheng et al., 2023; Koo et al., 2024; Park et al., 2024) is a prominent issue wherein the LLM-judge favors preference responses that are longer/more verbose; we attempt to alleviate this issue by explicitly mentioning in all our prompts a reminder to the model to not be influenced by verbosity. **Egocentric bias (self preference).** (Li et al., 2024; Koo et al., 2024) is a bias wherein LLM-judges prefer responses that are generated by the same model as the judge itself. In our work, to the best of our knowledge we do not encounter situations of this nature since all the datasets presumably have responses generated and ranked by multiple language models, and the preference data is *not* GPT versus other models; but ultimately, since we do not know the data used to train GPT-4O, CLAUDE, QWEN and we do not know the language models used in creating ANTHROPIC, there is a possible that our results are affected by egocentric bias. In general, using LLMs as judges could include more biases such as beauty bias, misinformation oversight bias, authority bias, etc. (Li et al., 2024).

Trust, Risk. Trust in a language model is subjective and user-dependent (Lipton, 2018); some users trust models that provide a high accuracy, some users require the model's reasoning process to be clear, etc. Further, risks such as hallucination, bias, etc. are also pivotal in the debate about trust. As we discuss above, we work on reducing hallucinations and bias in our judges and also propose judges and juries that improve the judgment accuracy; however, we still urge any researcher/user who utilizes our methods to qualitatively analyze our methods' predictions to ensure quality and trust.

Utility and necessity of LLM judges. There exist many relevant works (Zheng et al., 2023) which discuss the *need* for LLM-judges (inadequacies of existing benchmarks, difficulty in designing exact metrics, etc.); furthermore, there are several works that discuss the usage of LLM-judges in downstream tasks (Lambert et al., 2024; Cui et al., 2023). In this work, we do not debate the utility and necessity of using LLM judges; our focus is on improving the *performance* and *reliability* of LLM-judges.

Reproducibility

Data and models. All the datasets that we use in our work are released publicly for usage and have been duly attributed to their original authors. We do not train or release any models; we release all prompts in Appendix E, F, G, M. We use publicly available models `gpt-4o-2024-08-06`, `claude-3-5-sonnet-20241022`, `Qwen/Qwen2.5-32B-Instruct`, `Skywork-Reward-Llama-3.1-8B-v0.2`, `QRM-Gemma-2-27B`, `INF-ORM-Llama3.1-70B` (evaluation parameters in Appendix Q). We perform only evaluation and no training; all our evaluation dataset details are in Appendix D. All code and data are uploaded at `github.com/INK-USC/Amulet-DA-and-Maxim`.

Usage and License. All the datasets we use are available open-sourced on hugging-face: ANTHROPIC, WILDFEEDBACK, NECTAR. All datasets are licensed to be freely used and distributed (NECTAR alone is licensed to not compete with OpenAI, which we comply with). All these datasets are anonymized; we do not separately take steps to identify and remove offensive data. We use GPT-4O and CLAUDE as paid API services with corresponding API keys. We use the following models from hugging-face: `Skywork-Reward-Llama-3.1-8B-v0.2`, `QRM-Gemma-2-27B`, `INF-ORM-Llama3.1-70B`, `Qwen/Qwen2.5-32B-Instruct`. All these models are open-sourced; to access QWEN, we had to provide our contact information and accept the terms and conditions. All data and models were used consistent with their intended use.

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. Brihi Joshi was supported by the Apple Scholars in AI/ML PhD fellowship. Finally, we thank all lab members at USC INK Research Lab for their constructive feedback on this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ali Ahmadvand, Jason Ingyu Choi, and Eugene Agichtein. 2019. Contextual dialogue act classification for open-domain conversational agents. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 1273–1276.

AI@Meta. 2024. Llama 3 model card.

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. 2023. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7716–7733, Toronto, Canada. Association for Computational Linguistics.

James Allen and Mark Core. 1997. Draft of damsl: Dialog act markup in several layers.

Jens Allwood. 1992. On dialogue cohesion. *rapport nr.: Gothenburg Papers in Theoretical Linguistics 65*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: (May 12, 2024).

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Harry Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

Harry Bunt. 2011. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2):222–245.

Harry Bunt. 2019. Guidelines for using iso standard 24617-2.

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 *The Twelfth International Conference on Learning Representations*.

Bao Chen, Yuanjie Wang, Zeming Liu, and Yuhang Guo. 2023. Automatic evaluate dialogue appropriateness by using dialogue act. In *Findings of the Association* *for Computational Linguistics: EMNLP 2023*, pages 7361–7372.

Guiming Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 225–234.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7594–7601.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

Luigi Daniele and Suphavadeeprasit. 2023. Amplify-instruct: Synthetically generated diverse multi-turn conversations for efficient llm training. *arXiv preprint arXiv:(coming soon)*.

Qingxiu Dong, Li Dong, Xingxing Zhang, Zhifang Sui, and Furu Wei. 2024a. Self-boosting large language models with synthetic preference data. *Preprint*, arXiv:2410.06961.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024b. Can llm be a personalized judge? *Preprint*, arXiv:2406.11657.

Nicolai Dorka. 2024. Quantile regression for distributional reward models in rlhf. *arXiv preprint arXiv:2409.10164*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

R Frederking. 1996. Grice's maxims: do the right thing. *Frederking, RE.*

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland. Association for Computational Linguistics.

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. Speaker turn modeling for dialogue act classification.

Md Mahroof Hossain. The application of grice maxims in conversation: A pragmatic study. *Journal of English Language Teaching and Applied Linguistics*, 3(10):32–40.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: Llm judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370*.

Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Preprint*, arXiv:2404.16019.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 517–545.

Lea Krause and Piek TJM Vossen. 2024. The gricean maxims in nlp-a survey. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 470–485.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *Preprint*, arXiv:2505.06120.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *Preprint*, arXiv:2406.18629.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*.

Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. 2024. Aligning large language models by on-policy self-judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11442–11459.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2024. Decision-oriented dialogue for human-ai collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2170–2178.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. 2024. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, and 1 others. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *arXiv preprint arXiv:2203.10012*.

Erik Miehling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth Daly, David Piorkowski, and John Richards. 2024. Language models in dialogue: Conversational maxims for human-ai interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14420–14437.

Xiaoyu Tan Minghao Yang, Chao Qu. 2024. Inf-orm-llama3.1-70b.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Silvia Pareti and Tatiana Lando. 2018. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1043–1067.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Chen Qu, Liu Yang, W Bruce Croft, Yongfeng Zhang, Johanne R Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 25–33.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Jauhar, Sihao Chen, Shan Xia, and 1 others. 2024a. Wildfeedback: Aligning llms with in-situ user interactions and feedback. *arXiv preprint arXiv:2408.15549*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. 2025. Fspo: Few-shot preference optimization of synthetic preference data in llms elicits effective personalization to real users. *Preprint*, arXiv:2502.19312.

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

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Gemma Team. 2024a. Gemma.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Vipul Raheja Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. In *Proceedings of NAACL-HLT*, pages 3727–3733.

David R Traum and Elizabeth A Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024b. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. 2024. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022. Fined-eval: Fine-grained automatic dialogue-level evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3336–3355.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023a. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*, 3.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023b. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, and 1 others. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness harmlessness with rlaif.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*.

A Related Work

LLM judges. Prior work on text quality and dialog metrics include model-free metrics such BLEU, ROUGE, engaging-ness (based on length of conversation) etc. and metrics with trained models such as grammar, fluency, coherence, error recovery, informativeness, likability, inquisitiveness, etc. (Papineni et al., 2002; Pang et al., 2020; Yeh et al., 2021; Fu et al., 2024; Mehri et al., 2022; Zhang et al., 2022; Ni et al., 2023). Today, with the increasing proficiency of language models, it has become harder to measure the quality of the text generated by them with standardized metrics; human evaluation of these responses are too expensive and time-consuming (Ouyang et al., 2022; Zheng et al., 2023). This has led to the adoption of large language models as judges (Zheng et al., 2023); these LLM-judges mimic human judgments and preferences, and are the most commonly used evaluators today.

LLM-judges can provide feedback in various forms (Zheng et al., 2023; Lin et al.): (1) preference/rankings for multiple texts, which can serve as either just evaluation or as RL-AIF data as well (Cui et al., 2023), (2) likert/numerical scores that indicate the quality of the text (the judge's prompt can specify exact score levels or reasoning, or can just request for a score between say 1-10), (3) natural language feedback that can be used to improve language models (Akyurek et al., 2023).

Current work in LLM-judges also include juries of judges (Lin et al.; Verga et al., 2024; Jung et al., 2024), training smaller language models to be judges (Zheng et al., 2023; Zhu et al.; Kim et al.).

Single-turn and multi-turn. Contemporary works on benchmarking generative language models and reward models largely focus on training and analysis with single- or two-turn datasets, for example, UltraFeedback (Cui et al., 2023), SHP (Ethayarajh et al., 2022), ChatArena (Chiang et al.), MT-Bench (Zheng et al., 2023), Alpaca-Eval (Li et al., 2023), RewardBench (Lambert et al., 2025) etc. A recent benchmark WildBench (Lin et al.) takes a step in this direction and focuses on multi-turn data. But WildBench only works with ≤ 5 human turns of data and has only 1024 evaluation instances; further WildBench heavily focuses on singular task based examples (such as math, editing, data analysis). Other recent preference datasets include PRISM (Kirk et al., 2024) and Argilla-Capybara (Daniele and Suphavadeeprasit, 2023): unfortunately, PRISM includes many instances where the chosen and rejected responses are highly similar, and Argilla-Capybara is completely synthetic and heavily based on individual tasks; hence, we are unable to use these in our multi-turn preference benchmark. More recently, multi-turn datasets such as WildChat (Zhao et al., 2024), LMSYS-Chat (Zheng et al.) etc. which contain real-life conversations between humans and various LM assistants have been released - unfortunately, these datasets do not have preference information and can only be used for supervised fine-tuning.

Dialog Acts. There is a rich history of work in linguistics that discusses the *multi-functional* nature of speakers' utterances in dialogues (Allen and Core, 1997; Allwood, 1992; Bunt, 1994, 2011); these works deal with both spoken and written spoken conversations, also discussing real-life cases where speakers interrupt each other, where multiple speakers speak at the same time, etc. Formally, the term *Dialog Act* is used to describe the intents and communicative functions of a speaker's utterance. Various taxonomies of dialog acts have been proposed over the years. Very early on, Searle (1969) proposed *Speech Acts* which deal with the functionalities present in general speech. Later, works such as Traum and Hinkelman (1992) discussed how dialogue is much more complex than simple speech (such as turn-taking across speakers, grounding of the dialog to the speakers and the situation and environment at hand); they proposed that we need a broader taxonomy, *Dialog Acts*, (or *Conversation Acts*) that deals with not just traditional speech but also the more complicated aspects of dialog. Bunt (1994) discussed how dialogs have a communicative goal, a non-communicative goal (the underlying task), as well as communicative subgoals that arise over the course of a conversation; they also proposed their dialog act schema for the same. Allen and Core (1997) proposed *DAMSL*, a popular (at the time) annotation scheme to classify dialog acts in conversations. Allwood (1992) discussed the concept of *cohesion* in a dialogue, including concepts of communicative functions, communicative relevance, etc. Stolcke et al. (2000) presents a large dialog act taxonomy and proposes a statistical approach for modeling dialog acts using hidden Markov models. Pareti and Lando (2018) proposes a dialog act scheme that includes primary and secondary intents, as well as a way to classify the intents as being explicit or implicit; further, they propose a way to represent dialog acts in a conversation in a hierarchical/graph format. Welivita and Pu (2020) proposes dialog acts (for example: consoling, encouraging, appreciating, disapproving, etc.) specifically for utterances with empathetic intents. Bunt (2011) defines two terms: communicative *dimensions* that deal with the semantic/information content, and communicative *functions* that deal with the linguistic phrasings; they propose an ISO standard annotation scheme for the same in Bunt (2019). More recently, there have been several works on dialog act prediction/classification using neural networks (Tetreault, 2019; Chen et al., 2018; He et al.; Ahmadvand et al., 2019; Kumar et al., 2018; Lin et al., 2024; Chen et al., 2023; Qu et al., 2019; Liu et al., 2017; Colombo et al., 2020).

Dialog Metrics and Grice's maxims. Contemporary works specify varied metrics that a language model has to satisfy to be a good conversational agent; these include metrics pertaining to conversational AI such as flexibility to the human user, user satisfaction, error recovery, as well as metrics pertaining to language usage such as grammar, fluency, factuality (Ni et al., 2023; Zhang et al., 2022; Mehri et al., 2022; Fu et al., 2024; Yeh et al., 2021). However, not all these metrics have clear definitions, and there exists no *complete* list of metrics that are sufficient and necessary for a language model to be considered universally good at making conversation. Hence, in our analysis of human-assistant conversations, to measure the properties that the AI assistant (dis)satisfies, we go back to the fundamental conversational principles of *Gricean Maxims* (Grice, 1975). Grice's maxims have been used in NLP literature in varying degrees: Krause and Vossen (2024) designates the most important maxims for different NLP tasks, Kasirzadeh and Gabriel (2023) discusses Grice's maxims in terms of various domains. However, at the same time, there have also been discussions in linguistics regarding the vagueness of the maxim definitions (Frederking, 1996) and the insufficiency of Grice's maxims with respect to (1) non-cooperative communication such as irony and sarcasm, (2) adaptability to and extension across cultures (Hossain).

Reward models. Since its release, RewardBench (Lambert et al., 2025) has served as a solid benchmark for reward models, and has standardized the evaluation of reward models. Most of these reward models are classifiers (Minghao Yang, 2024; Dorka, 2024; Liu et al., 2024; Wang et al., 2024a; Lou et al., 2024), producing an unbounded numerical score; the higher the score, the better the quality of the text being analyzed. RewardBench largely focuses on single-turn conversations.

B Example complex multi-turn conversation

As a first step towards understanding the complexity of multi-turn conversations, we manually analyzed several conversations from datasets including MT-Bench (Zheng et al., 2023), WildChat (Zhao et al., 2024), LMSYS-Chat (Zheng et al.). We observed that the user intents often change from one turn to the next. We preliminarily observed the following intents (sample in Table 5): social (greeting/thanking/goodbye), role playing instructions, asking the assistant to perform a task, seeking information from the assistant, stating a fact, giving additional/relevant information about a situation, expressing an emotion, technical/error statements (which implicitly request the assistant to explain or debug the error). We used this as our motivation and sought to formalize the changing intents in a conversation; we adapt and present *dialog acts* from Bunt (2019) that explain the communicative structures and intents in every turn of a conversation.

C DA and MAXIM

Tables 6 and 7 provide a full list of all dialog acts and maxims we use in this work. We obtain the dialog act taxonomy from Bunt (2019) and the maxim definitions from Grice (1975); Miehling et al. (2024).

D Data cleaning

In this section, we expand upon our data cleaning strategies.

- Step-0: Downloading the datasets from the Hugging Face sources - we download HH-TRAIN and HH-TEST from huggingface.co/datasets/Anthropic/hh-rlhf, WILDFEEDBACK from huggingface.co/datasets/microsoft/WildFeedback

Human turn	Observed Intents
Example-1	-
Have you ever heard of the National Grid in the UK?	seeking information
Can you tell me where the National Grid Control Center is?	seeking information, asking the assistant to perform a task
Do you have the exact address. I need to deliver an item there	seeking information, asking the assistant to perform a task, giving additional/relevant information about a situation
Maye you have its post code?	seeking information, asking the assistant to perform a task
Oh that's strange I thought it began with RG41?	expressing an emotion, stating a fact, giving additional/relevant information about a situation
Example-2	-
I'm feeling sad. Can you tell me a joke to cheer me up?	expressing an emotion, asking the assistant to perform a task
My grandmother used to tell me jokes about Irish people, which always cheered me up. Pretend to be my grandmother and cheer me up	role-playing instructions, asking the assistant to perform a task
My grandmother used to make jokes about how Irish people have no potatoes. How about a potato joke, something like: Grandma: what's the difference between an Irishman and a bench?	giving additional/relevant information about a situation, asking the assistant to perform a task

Table 5: Consecutive human turns from a conversation in WildChat (Zhao et al., 2024). We observe multiple intents in almost all the turns, and also observe that the intent of the human keeps changing over the turns.

and NECTAR from huggingface.co/datasets/allenai/tulu-2.5-preference-data.

- Step-1: Only retain conversations with ≥ 4 **human turns**.
- Step-2: Remove all instances where the chosen and rejected responses are *exactly* the same (done via string matching).
- Step-4: Remove all instances where the structure of the instance is ill-formed, for example, two human turns in a row.
- Step for NECTAR: We only consider instances from the first 100000 instances in NECTAR in order to control the size of the evaluation dataset.
- Step for NECTAR: Since NECTAR has been created from HH-TRAIN, we check for and remove the following instances: (1) instances also present in HH-TRAIN, (2) instances where the chosen and rejected responses are opposite to that of HH-TRAIN
- Step for WILDFEEDBACK: We only retain instances in WILDFEEDBACK where every human or assistant turn is less than 300 words in length.

The final evaluation set sizes, where the conversations are ≥ 4 **human turns** are as follows: HH-TEST- 460, HH-TRAIN- 8210, NECTAR-6531 and WILDFEEDBACK- 4684.

The evaluation set sizes for conversations with ≥ **7 human turns** are as follows: HH-TRAIN- 548, NECTAR- 335 and WILDFEEDBACK- 1867.

E I/O and W-EXPL prompts

We present the baseline judge I/O and W-EXPL prompts in Tables 8 and 9 respectively.

F The AMULET-DA prompt

This section contains the prompt we use for AMULET-DA with `this font`; we provide this prompt outside of a table since it is too big to be accommodated in a table.

Communicative Dimension	Definition of Communicative Dimension	Communicative Function
Task	Underlying task/activity	Questions, Answers, Commissives, Directives
Auto-feedback	Speaker's processing of previous utterances	Positive/negative feedback
Allo-feedback	Addressee's processing of previous utterances	Positive/negative feedback
Turn Management	Allocation of turn [implicit, mostly]	Keep/grab/give turn
Time Management	Time needed to speak	Stalling, Pausing
Contact Management	Check contact	Check if contact is there
Own Communication Management	Editing speaker's own turn	Self-correction
Partner Communication Management	Editing what the addressee said	Correct misspeaking / Completion of addressee
Discourse/Interaction Structuring	Structure of dialogue	New topic, open/close dialogue, announce change in structure
Social Obligations Management	Social obligation	Greeting, Thanks, Apology and corresponding responses

Table 6: **Dialog Acts:** Communicative dimensions and functions from Bunt (2019)

Instruction: You will be given a dialog conversation between a human user and an LLM assistant. The dialog is split into turns - a turn is defined as an utterance by either the human user or the assistant. Note that the roles of "speaker" (S) and "addressee" (A) will alternate at every turn. The last turn alone will have two responses, sampled from different LLM assistants. **Your task is to label each turn of dialogue in terms of dialog-acts - dialog acts are defined in terms of communicative dimensions "Dim" and corresponding communicative functions "Func"** [detailed below along with their meanings]. For each turn of dialogue, you must mark *all* the dimensions and functions that are present. You should take the previous turns of the dialog into consideration when labeling the dialog acts. **Finally, use these dialog acts to determine which response is better** - say "1" if you think Assistant-1's response is better and "2" if you think Assistant-2's response is better. Also provide an explanation for your choice. **Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible.** Examples are given below with the desired output format.

List of dimensions and functions -
(1) Dim: Task, Meaning: Underlying task/activity, **Func:** Propositional Question, Set Question, Choice Question, Answer, Confirm, Disconfirm, Inform, Agreement, Disagreement, Correction, Promise, Offer, Accept Request, Decline Request, Accept Suggest, Decline Suggest, Request, Instruct, Suggest
(2) Dim: Auto-Feedback, Meaning: Speaker's processing of previous utterances, **Func:** Auto-Positive, Auto-Negative
(3) Dim: Allo-Feedback, Meaning: Speaker expressing opinions about addressee's processing of previous utterances, **Func:** Allo-Positive, Allo-Negative, Feedback

Maxim	Definition
Quantity	Related to the informativity, that is, the amount of information contained in a given response.
Quality	Pertains to the truthfulness and honesty of the response
Relation	The relevance of the response to the conversation
Manner	Logical flow and clarity of the discourse
Benevolence	Moral responsibility of the response, particularly concerning the generation of, and engagement with, harmful content
Transparency	concerning recognition of one's knowledge boundaries, operational constraints, and intents
Quantity-1	The response should provide a sufficient amount of information.
Quantity-2	The response should not contain unnecessary details.
Quality	The response should be factual and supported by adequate evidence whenever possible.
Relevance-1	The response should directly and relevantly address the recipient's statements in a helpful manner.
Relevance-2	The response should be relevant to the current topic and not unnaturally shift the conversation to unrelated subjects.
Manner-1	The response should be clear, unambiguous, and presented in a well-organized fashion.
Manner-2	The response should be accessible and use appropriate language tailored to the recipient's level of understanding.
Benevolence-1	The response should not exhibit insensitivity, rudeness, or harm.
Benevolence-2	The response should not reflect an engagement or endorsement with requests that are harmful or unethical.
Transparency-1	The response should recognize the speaker's knowledge boundaries, making clear any limitations in expertise, evidence, experience, or context.
Transparency-2	The response should recognize the speaker's operational capabilities, highlighting the nature of actions that can or cannot be performed.
Transparency-3	The response should be forthright about the speaker's willingness to engage with specific subjects or heed relevant advice.

Table 7: **Maxims:** Definitions quoted from Miehling et al. (2024). The highlighted cells are the sub-maxims we use in AMULET-MAXIM.

Elicitation
(4) Dim: Time Management, Meaning: Concerning the allocation of time to the speaker, **Func:** Stalling, Pausing
(6) Dim: Own Communication Management, Meaning: Editing speaker's own speech within the current turn, **Func:** Self-Correction, Self-Error, Retraction
(7) Dim: Partner Communication Management, Meaning: Editing what the addressee said, **Func:** Completion, Correct Misspeaking,
(8) Dim: Discourse/Interaction Structuring, Meaning: Explicitly structuring the interaction, **Func:** Interaction Structuring, Opening, Closing
(9) Dim: Social Obligations Management, Meaning: Social obligation, **Func:** Initial Greeting, Return Greeting, Intitial Self-Introduction, Return Self-Introduction, Apology, Accept Apology, Thanking, Accept Thanking, Initial Goodbye, Return Goodbye
You should use only the above dimensions and functions, do not make up new ones. Below you will find detailed definitions of the functions. **To reiterate, S refers to the speaker and A refers to the addressee. Both the human and the LM assistant take on the roles of S and A in alternating turns.**

Information-seeking functions -
* Propositional Question: Function performed by S, in order to know whether the proposition, which forms the semantic content, is true. S assumes that A knows

Instruction: You will be given a dialog conversation between a human user and an LLM assistant. The dialog is split into turns - a turn is defined as an utterance by either the human user or the assistant. Note that the roles of "speaker" (S) and "addressee" (A) will alternate at every turn. The last turn alone will have two responses, sampled from different LLM assistants. Your task is to analyze the two responses and say which one of them is better - you should take all the previous turns of the dialog into consideration. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. An example is given below with the required JSON output format. Say "1" if you think Assistant-1's response is better and "2" if you think Assistant-2's response is better.

Example Dialog -
Human: Human's turn
Assistant: Assistant's turn
Human: Human's turn
Assistant: Assistant's turn
Human: Human's turn
Assistant-1: Assistant's turn by Assistant-1
Assistant-2: Assistant's turn by Assistant-2

Example Output -
{
"Answer": "fill either 1 or 2",
}

Table 8: Prompt for I/O

Instruction: You will be given a dialog conversation between a human user and an LLM assistant. The dialog is split into turns - a turn is defined as an utterance by either the human user or the assistant. Note that the roles of "speaker" (S) and "addressee" (A) will alternate at every turn. The last turn alone will have two responses, sampled from different LLM assistants. Your task is to analyze the two responses and say which one of them is better - you should take all the previous turns of the dialog into consideration. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. An example is given below with the required JSON output format. Say "1" if you think Assistant-1's response is better and "2" if you think Assistant-2's response is better.

Example Dialog -
Human: Human's turn
Assistant: Assistant's turn
Human: Human's turn
Assistant: Assistant's turn
Human: Human's turn
Assistant-1: Assistant's turn by Assistant-1
Assistant-2: Assistant's turn by Assistant-2

Example Output -
{
"Answer": "fill either 1 or 2",
"Explanation": "fill your explanation here"
}

Table 9: Prompt for W-EXPL

whether the proposition is true or not, and puts pressure on A to provide this information.
* Set Question: Function performed by S, in order to know which elements of a given set have a certain property specified by the semantic content; S puts pressure on A to provide this information, which S assumes that A possesses. S believes that at least one element of the set has that property.
* Choice Question: Function performed by S, in order to know which one from a list of alternative propositions, specified by the semantic content, is true; S believes that exactly one element of that list is true; S assumes that A knows which of the alternative propositions is true, and S puts pressure on A to provide this information.

Information-providing functions -
* Answer: Function performed by S, in order to make certain information available to A which S believes A wants to know; S assumes that this information is correct.
* Confirm: Function performed by S, in order to inform A that certain information that A wants to know, and concerning which A holds an uncertain belief, is indeed correct.
* Disconfirm: Function performed by S, in order to let A know that certain information that A wants to know, and concerning which A holds an uncertain belief, is incorrect.
* Inform: Function performed by S, in order to make the information contained in the semantic content known to A; S assumes that the information is correct.
* Agreement: Function performed by S, in order to inform A that S assumes a given proposition to be true, which S believes that A also assumes to be true.
* Disagreement: Function performed by S, in order to inform A that S assumes a given proposition to be false, which S believes that A assumes to be true.
* Correction: Function performed by S, in order to inform A that certain information which S has reason to believe that A assumes to be correct, is in fact incorrect and that instead the information that S provides is correct.

Commissive functions -
* Promise: Function by which S commits to perform the action specified in the semantic content, in the manner or with the frequency or depending on the conditions that S makes explicit. S believes that this action would be in A's interest.
* Offer: Function by which S indicates willingness and ability to perform the action specified by the semantic content, conditional on the consent of A that S do so.
* Accept Request: Function by which S commits to perform an action that S has been requested to perform, possibly depending on certain conditions that S makes explicit.
* Decline Request: Function by which S refuses to perform an action that S has been requested to perform, possibly depending on certain conditions that S makes explicit.
* Accept Suggest: Function by which S commits to perform an action that was suggested, possibly with certain restrictions or conditions concerning manner or frequency of performance.
* Decline Suggest: Function by which S indicates that S will not perform an action that was suggested, possibly depending on certain conditions that S makes explicit.

Directive functions -
* Request: Function performed by S, in order to create a commitment for A to perform a certain action in the manner or with the frequency described by the semantic content, conditional on A's consent to perform the action. S assumes that A is able to perform this action.
* Instruct: Function performed by S, in order to create a commitment for A to carry out a named action in the manner or with the frequency specified by the semantic content; S assumes that A is able and willing to carry out the action.
* Suggest: Function performed by S, in order to make A consider the performance of a certain action specified by the semantic content. S believes that this action is in A's interest, and assumes that A is able to perform the action.

Feedback functions -
* Auto-Positive: Function performed by S, in order to inform A that S believes that S's processing of the previous utterance(s) was successful.
* Allo-Positive: Function performed by S, in order to inform A that S believes that A's processing of the previous utterance(s) was successful.
* Auto-Negative: Function performed by S, in order to inform A that S's processing of the previous utterance(s) encountered a problem.

* Allo-Negative: Function performed by S, in order to inform A that S believes that A's processing of the previous utterance(s) encountered a problem.
* Feedback Elicitation: Function performed by S, in order to know whether A's processing of the previous utterance(s) was successful.

Time management functions -
* Stalling: Function performed by S, in order to have a little extra time to construct S's contribution.
* Pausing: Function performed by S, in order to suspend the dialogue for a short while.

Own and Partner Communication Management Functions -
* Completion: Function performed by S in order to assist A in the completion of an utterance.
* Correct Misspeaking: Function performed by S, in order to correct (part of) an utterance by A assuming that A made a speaking error.
* Self-Error: Function performed by S, in order to signal to the A that S has made a mistake in speaking.
* Retraction: Function performed by S, in order to withdraw something that S just said within the same turn.
* Self-Correction: Function performed by S, in order to correct an error that S just made, or to improve on an infelicitous formulation that S just used, within the same turn.

Discourse structuring functions -
* Interaction Structuring: Function performed by S, in order to explicitly indicate to A the function or topic of S's next contribution(s).
* Opening: Function performed by S, in order to inform A that S is ready and willing to engage in a dialogue with A.
* Closing: Function performed by S, in order to inform A that S is about to end the conversation.

Social obligations management functions -
* Initial Greeting: Function performed by S, in order to inform A that S is present and aware of A's presence; S puts pressure on A to acknowledge this.
* Return Greeting: Function performed by S, in order to acknowledge that S is aware of A's presence, and of A having signalled A's presence to S. * Initial Self-Introduction: Function performed by S, in order to make A know about S; S puts pressure on A to acknowledge this.
* Return Self-Introduction: Function performed by S, in order to make A know about S in response to a self-introduction by A.
* Apology: Function performed by S, that signals S wants A to knows that S regrets something; S puts pressure on A to acknowledge this.
* Accept Apology: Function performed by S, in order to mitigate the feelings of regret that A has expressed.
* Thanking: Function performed by S, in order to inform A that S is grateful for some action performed by A; S puts pressure on A to acknowledge this.
* Accept Thanking: Function performed by S, in order to mitigate the feelings of gratitude which A has expressed.
* Initial Goodbye: Function performed by S, in order to inform A that S intends the current utterance to be S's final contribution to the dialogue; S puts pressure on A to acknowledge this.
* Return Goodbye: Function performed by S, in order to acknowledge awareness that A has signalled A's final contribution to the dialogue and S signals in return S's agreement to end the dialogue; S has been pressured to respond to an initial goodbye by A.

Important additional note: If a turn contains the answer to a question from a prior turn, make sure to mark the communicative function "Answer" in addition to any other function present in the turn (such as "Inform", "Suggest", "Instruct", etc.). That is, even if the answer is phrased as an instruction, or a suggestion, etc., if the turn includes the answer to a question from a prior turn, specify "Answer" as one of the functions in the turn.

Example Dialog 1 -
Human: How do I teach kids to meditate?
Assistant: Great question! That's a really useful skill to cultivate, it can bring peace, calm, and happiness. I'm glad you want to teach your kids about it.
Human: All right, so how do we start?
Assistant: Well, we can get started with just being silent. You can tell the kids it's okay if they just sit there quietly for a few minutes without thinking of anything. **Human:** any other ideas? they are fidgeting
Assistant-1: Kids are great at fidgeting, so let's make sure we give them permission to fidget.
Assistant-2: Yeah, definitely. What works for you personally? What works for other people?

Example Output 1 -
Human: How do I teach kids to meditate? <SEP> {"Dim": "Task", "Func": "Set Question"}
Assistant: Great question! That's a really useful skill to cultivate, it can bring peace, calm, and happiness. I'm glad you want to teach your kids about it. <SEP> {"Dim": "Allo-Feedback", "Func": "Allo-Positive"} {"Dim": "Task", "Func": "Inform"}
Human: All right, so how do we start? <SEP> {"Dim": "Allo-Feedback", "Func": "Allo-Positive"} {"Dim": "Task", "Func": "Request & Set Question"}
Assistant: Well, we can get started with just being silent. You can tell the kids it's okay if they just sit there quietly for a few minutes without thinking of anything. <SEP> {"Dim": "Time Management", "Func": "Stalling"} {"Dim": "Task", "Func": "Answer"}
Human: any other ideas? they are fidgeting <SEP> {"Dim": "Allo-Feedback", "Func": "Allo-Negative"} {"Dim": "Task", "Func": "Set Question & Inform"}
Assistant-1: Kids are great at fidgeting, so let's make sure we give them permission to fidget. <SEP> {"Dim": "Task", "Func": "Inform & Instruct"}
Assistant-2: Yeah, definitely. What works for you personally? What works for other people? <SEP> ""Dim": "Auto-Feedback", "Func": "Auto-Positive"} {"Dim": "Task", "Func": "Set Question"}

{
"**Answer**": "fill either 1 or 2",
"**Explanation**": "fill your explanation here"
}

Example Dialog 2 -
Human: what are some good novels for a 9 year old?
Assistant: That depends on the 9 year old, but if they like thrillers or mysteries then maybe something like "The Body in the Bog" or "The Boy Who Followed Ripley." For more modern, female-written fiction I might recommend "Jane Eyre" or "Frankenstein."
Human: That sounds cool
Assistant-1: I'm a programmer, so I'm really into good programming texts, like "The Practice of Programming" or "JavaScript: The Good Parts." But that might be a bit too much for a 9 year old, so
Assistant-2: Do you want some quotes?

Example Output 2 -
Human: what are some good novels for a 9 year old? <SEP> "Dim": "Task", "Func": "Set Question"
Assistant: That depends on the 9 year old, but if they like thrillers or mysteries then maybe something like "The Body in the Bog" or "The Boy Who Followed Ripley." For more modern, female-written fiction I might recommend "Jane Eyre" or "Frankenstein." <SEP> {"Dim": "Answer & Suggest"}
Human: That sounds cool <SEP> {"Dim": "Allo-Feedback", "Func": "Allo-Positive"}
Assistant-1: I'm a programmer, so I'm really into good programming texts, like "The Practice of Programming" or "JavaScript: The Good Parts." But that

might be a bit too much for a 9 year old, so <SEP> {“Dim”: “Task”, “Func”: “Inform”} {“Dim”: “Own Communication Management”, “Func”: “Retraction”}
Assistant-2: Do you want some quotes? <SEP> {“Dim”: “Task”, “Func”: “Offer & Propositional Question”}

{
“**Answer**”: “fill either 1 or 2”,
“**Explanation**”: “fill your explanation here”
}

Dialog -

F.1 The AMULET-DA prompt for CLAUDE

With CLAUDE, we encountered a unique problem - CLAUDE stopped its generation after the dialog act for the first human turn because it assumed that its turn was *over*. Hence, we use the following modified output format for the few-shot examples in AMULET-DA's prompt for CLAUDE:

Example Output 1 -
{
““**Human:** How do I teach kids to meditate?””: ““{“Dim”: “Task”, “Func”: “Set Question”}””,
““**Assistant:** Great question! That's a really useful skill to cultivate, it can bring peace, calm, and happiness. I'm glad you want to teach your kids about it.””: ““{“Dim”: “Allo-Feedback”, “Func”: “Allo-Positive’} {“Dim”: “Task”, “Func”: “Inform”}””,
““**Human:** All right, so how do we start?””: ““{“Dim”: “Allo-Feedback”, “Func”: “Allo-Positive’} {“Dim”: “Task”, “Func”: “Request & Set Question”}””,
““**Assistant:** Well, we can get started with just being silent. You can tell the kids it's okay if they just sit there quietly for a few minutes without thinking of anything.””: ““{“Dim”: “Time Management”, “Func”: “Stalling”} {“Dim”: “Task”, “Func”: “Answer”}””,
““**Human:** any other ideas? they are fidgeting””: ““{“Dim”: “Allo-Feedback”, “Func”: “Allo-Negative”} {“Dim”: “Task”, “Func”: “Set Question & Inform”}””,
““**Assistant-1:** Kids are great at fidgeting, so let's make sure we give them permission to fidget.””: ““{“Dim”: “Task”, “Func”: “Inform & Instruct”}””,
““**Assistant-2:** Yeah, definitely. What works for you personally? What works for other people?””: ““{“Dim”: “Auto-Feedback”, “Func”: “Auto-Positive”} {“Dim”: “Task”, “Func”: “Set Question”}””,
‘**Answer**’: ‘fill either 1 or 2’,
‘**Explanation**’: ‘fill your explanation here’
}

G The AMULET-MAXIM prompt

This section contains the prompt we use for AMULET-MAXIM with **this font**; we provide this prompt outside of a table since it is too big to be accommodated in a table.

Instruction: You will be given a dialog conversation between a human user and an LLM assistant. The dialog is split into turns - a turn is defined as an utterance by either the human user or the assistant. Note that the roles of “speaker” and “addressee” will alternate at every turn. The last turn alone will have two responses, sampled from different LLM assistants. Your task is to analyze the two responses and say which one of them is better using the below maxim based evaluation protocol. For each of the twelve maxims described below, indicate whether response 1 (from Assistant-1) or response 2 (from Assistant-2) is better in satisfying the maxim by specifying “1” or “2” respectively. If both responses satisfy the maxim equally, say “both” and if neither response satisfies the maxim, say “neither”. Finally provide a free-text explanation for your decisions and provide your answer of whether response 1 or response 2 is better. You should take the previous turns of the dialog into consideration when labeling the satisfaction of maxims and when providing your final answer. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. An example is given below with the required JSON output format. Say “1” if you think Assistant-1’s response is better and “2” if you think Assistant-2’s response is better. You have to pick “1” or “2”, you cannot say “both” for the final answer.

List of Grice’s maxims -
(1) Quantity-1: The response should provide a sufficient amount of information.
(2) Quantity-2: The response should not contain unnecessary details.
(3) Quality: The response should be factual and supported by adequate evidence whenever possible.
(4) Relevance-1: The response should directly and relevantly address the recipient’s statements in a helpful manner.
(5) Relevance-2: The response should be relevant to the current topic and not unnaturally shift the conversation to unrelated subjects.
(6) Manner-1: The response should be clear, unambiguous, and presented in a well-organized fashion.
(7) Manner-2: The response should be accessible and use appropriate language tailored to the recipient’s level of understanding.
(8) Benevolence-1: The response should not exhibit insensitivity, rudeness, or harm.
(9) Benevolence-2: The response should not reflect an engagement or endorsement with requests that are harmful or unethical.
(10) Transparency-1: The response should recognize the speaker’s knowledge boundaries, making clear any limitations in expertise, evidence, experience, or context.
(11) Transparency-2: The response should recognize the speaker’s operational capabilities, highlighting the nature of actions that can or cannot be performed.
(12) Transparency-3: The response should be forthright about the speaker’s willingness to engage with specific subjects or heed relevant advice.

You should use only the above maxims, do not make up new ones.

Example Dialog 1 -
Human: What do I do about rough knees?
Assistant: Oh, is there something I can help with?
Human: I want my knees to feel smoother
Assistant: You want to have smoother knees?
Human: Yes
Assistant-1: So you want your knees to not be rough?
Assistant-2: I’m happy to help with that, but first I’m going to ask you some questions to help me understand how you feel, and you can tell me anything that’s on your mind. I’ll start by asking you to tell me about the thing that’s making you the most frustrated, or the thing that’s bothering you the most. What’s making you the most frustrated? What is bothering you the most

Example Output - 1
{
“Quantity-1”: “neither”,
“Quantity-2”: “1”,
“Quality”: “1”,
“Relevance-1”: “neither”,
“Relevance-2”: “1”,
“Manner-1”: “1”,
“Manner-2”: “both”,
“Benevolence-1”: “both”,
“Benevolence-2”: “both”,
“Transparency-1”: “both”,
“Transparency-2”: “both”,
“Transparency-3”: “both”,
“Explanation”: “fill your explanation here”,
“Final Answer”: “fill either 1 or 2”
}

Example Dialog 2 -
Human: Is a higher octane gasoline better for your car?
Assistant: How old is the car? What is the mileage? What do you typically put in it? What model car is it?
Human: a 10 year old Ford.

Assistant: Then you probably don't need high-octane gasoline. The average car will do fine with the regular fuel.
Human: Then why do they sell it?
Assistant-1: Automotive manufacturers sell the high-octane fuel for you to use to make the engine run better. It's a trade-off. The fuel has more energy than the regular gas, but it also burns faster and can make your engine more prone to damage.
Assistant-2: There are various reasons. The gasoline may not be entirely identical, and if you have a few items that are very sensitive to fuel, you may need it for your car. However, most items should be OK with regular gas, and probably wouldn't be damaged by it either.

Example Output - 2
{ **"Quantity-1"**: "1",
"Quantity-2": "1",
"Quality": "1",
"Relevance-1": "1",
"Relevance-2": "1",
"Manner-1": "1",
"Manner-2": "1",
"Benevolence-1": "both",
"Benevolence-2": "both",
"Transparency-1": "1",
"Transparency-2": "both",
"Transparency-3": "both",
"Explanation": "fill your explanation here",
"Final Answer": "fill either 1 or 2"
}

Example Dialog 3 -
Human: Can you tell me the steps for getting a harbor boat slip for my sailboat?
Assistant: Sure. I believe you'll need a copy of the marina rules and regulations, since some of those details will likely determine what forms you'll need to fill out. It looks like there's a quick way to get a copy of the rules by phone, since you might be able to get the contact information for the harbormaster from your yacht club. If you get a hold of that person, you'll also be
Human: Thanks. And what are some typical forms required?
Assistant-1: Oh, I'm afraid I'm not very familiar with these processes myself. If you want, I can look it up for you, or you can ask me to do so when you need the information.
Assistant-2: At the marina level, there's usually the rental contract that covers the slip, and the paperwork to get insurance for the boat. You'll also need to register your boat with the state, usually by mail, but they'll also likely be able to help you if you go to the state office in person. At the yacht club level, you'll usually need

Example Output - 3
{
"Quantity-1": "2",
"Quantity-2": "2",
"Quality": "2",
"Relevance-1": "2",
"Relevance-2": "both",
"Manner-1": "2",
"Manner-2": "both",
"Benevolence-1": "both",
"Benevolence-2": "both",
"Transparency-1": "1",
"Transparency-2": "both",
"Transparency-3": "both",
"Explanation": "fill your explanation here",
"Final Answer": "fill either 1 or 2"
}

Dialog -

H Sample DA and MAXIM annotations

Tables 10 and 11 show sample annotations by GPT-4o with AMULET-DA and AMULET-MAXIM respectively.

I More DA and MAXIM analyses

We continue Section 3 here and provide more analyses.

I.1 Multi-turn analysis figures

We present multi-turn analysis figures from Section 3.1 for the remaining datasets HH-TEST, HH-TRAIN and NECTAR here, in Figures 6 and 8. Fig-

Human: I found a baby bird. What do I do? <SEP> {"Dim": "Task", "Func": "Request & Set Question"}
Assistant: Do you know how to use a phone? <SEP> {"Dim": "Task", "Func": "Choice Question"}
Human: Yes, I do. <SEP> {"Dim": "Task", "Func": "Confirm"}
Assistant: I'm going to send you a link to an animal rehabilitation center. <SEP> {"Dim": "Task", "Func": "Inform & Promise"}
Human: Okay, thanks. <SEP> {"Dim": "Allo-Feedback", "Func": "Allo-Positive"}, {"Dim": "Social Obligations Management", "Func": "Thanking"}
Assistant: Please just stay home with the bird until someone comes to pick it up, okay? I'm trying to get in touch with the center, but we can't reach them yet. <SEP> {"Dim": "Task", "Func": "Instruct & Inform"}
Human: Okay, now what? <SEP> {"Dim": "Allo-Feedback", "Func": "Allo-Positive"}, {"Dim": "Task", "Func": "Request"}
Assistant-1: I need you to go to the window and look outside. Can you tell me what you see? <SEP> "Dim": "Task", "Func": "Instruct & Set Question"
Assistant-2: Please open the link in the next few seconds. <SEP> {"Dim": "Task", "Func": "Instruct"}

Answer: Response-1
Explanation: Assistant-1 provides a clear next step for the user by instructing them to look outside and report back, which keeps the user engaged and provides a way to gather more information about the situation. This is more interactive and supportive in the context of the ongoing situation. Assistant-2's response, while instructive, does not directly address the user's request for further guidance beyond opening the link, which may not be immediately helpful if the user is seeking more immediate actions to take with the bird.

Table 10: Sample output when prompting GPT-4o with AMULET-DA

{
'Quantity-1': 'neither',
'Quantity-2': '2',
'Quality': 'neither',
'Relevance-1': 'neither',
'Relevance-2': 'neither',
'Manner-1': '2',
'Manner-2': '2',
'Benevolence-1': 'both',
'Benevolence-2': 'both',
'Transparency-1': 'neither',
'Transparency-2': 'both',
'Transparency-3': 'both',
'Answer': '2',
'Explanation': "Assistant-1's response does not provide any relevant information or guidance related to the situation with the baby bird, making it irrelevant and lacking in quality. Assistant-2's response, while also not directly addressing the situation, at least suggests an action related to the previous mention of a link, making it slightly more relevant and organized. Neither response provides sufficient information or clarity about what to do with the baby bird, nor do they acknowledge any limitations or provide transparency about the situation. Both responses are polite and do not engage in harmful or unethical behavior.",
}

Table 11: Sample output when prompting GPT-4o with AMULET-MAXIM. This follows the same instance as in Table 10.

ure 7 shows the frequency of most common dimensions for all four datasets. Further, we expand the acronyms we used in these figures and in Figure 3 here:
SOM - Social Obligations Management, OCM - Own Communication Management, PCM - Partner Communication Management, DIS - Discourse/Interaction Structuring, TM - Turn Management,
Req - request, Info - Inform, PQ - Propositional Question, SQ - Set Question, Instr - Instruct, Ans - Answer, Apolg - Apology, Sug - Suggest, Corr - Correction, Allo+ - Allo-positive feedback, CQ - Choice Question, Disconf - Disconfirm.

I.2 Statistics about #instances

In Appendix D we provided final evaluation set sizes for all datasets. However, when we prompt GPT-4o (or any model) with these instances and prompts, some instances do not get a usable output due to (1) safety/non-toxicity reasons, in case the original conversation was deemed unsafe by the language model, or (2) the model simply failed to produce an output with the right format even after repeated attempts (we attempt 6 times in total). For such instances, we consider the judge or jury to be a failure, and mark it as a loss for experiments in Section 4. However, for Section 3, we consider only the instances which did get a valid output, since these analyses are used to understand the data. Further, since we run each instance twice to avoid position bias, it is possible that the dimensions and functions generated for each turn slightly vary across the two turns due to variability of the model (note that for Section 3, we only consider the dimensions and functions generated for the first vote). We provide the statistics on all these in Table 12.

I.3 Dialog–Act Variation Across Consecutive Speaker Turns

We present again the proportion of consecutive turns in each dataset that have different dialog acts; Section 3.1 shows this via graphs, and this section shows this mathematically.

Human turns. We compute the fraction of consecutive human turns where the dialog act changes in the entire dataset. This statistic captures how often users shift intent between successive turns.

$$\mathrm{DAshift}^{\mathrm{turn}}_{\mathrm{human}} = \frac{\#\{\text{consecutive human turns with diff. DA's}\}}{\#\{\text{all consecutive human turns}\}}$$

Assistant turns. We apply the same metric to assistant turns (again excluding the final two response options), quantifying dialog act shifts between successive assistant responses.

$$\mathrm{DAshift}^{\mathrm{turn}}_{\mathrm{assistant}} = \frac{\#\{\text{consecutive asst. turns with diff. DAs}\}}{\#\{\text{all consecutive asst.-turns}\}}$$

The results are shown in Table 14.

I.4 Does the DA change atleast once in a conversation?

Human turns. We compute the fraction of instances in which the dialog act DA assigned to the human turns changes *atleast* once over the course of the conversation. This metric captures, for how the number of instances where there was a shift in user intent during the interaction:

$$\mathrm{DAshift}^{\mathrm{conv}}_{\mathrm{human}} = \frac{\#\{\text{instances with a human-DA change}\}}{\#\{\text{instances in the dataset}\}}$$

Assistant turns. Similarly, we compute the fraction of instances in which the assistant's DA changes at least once during the conversation, excluding the final preference responses.

$$\mathrm{DAshift}^{\mathrm{conv}}_{\mathrm{assistant}} = \frac{\#\{\text{instances with a assistant-DA change}\}}{\#\{\text{instances in the dataset}\}}$$

Table 13 lists the results for both human and assistant turns.

I.5 DA changes across preferences

Continuing Section 3.2, in Table 15 we provide the proportion of conversations in each dataset where the function/dimension is different across the preference responses.

I.6 Maxim Asymmetry in Preferred Responses

We measure the average number of Grice's maxims that are uniquely satisfied by either the chosen or the rejected response (but not both) across the dataset. This demonstrates the average number of maxims that are important for an instance; not that Figure 4 showed trends of which maxims are important to distinguish between the chosen and rejected in every dataset. For each dialogue, let $\mathcal{M}_i^{\text{chosen}}$ and $\mathcal{M}_i^{\text{rejected}}$ denote the sets of Grice's maxims satisfied by the two candidate responses then

$$\mathrm{MaximGap} = \frac{1}{N}\sum_{i=1}^{N}\left|\,\mathcal{M}_i^{\text{chosen}} \;\triangle\; \mathcal{M}_i^{\text{rejected}}\right|,$$

where N is the total number of dialogues and $\triangle$ denotes the symmetric difference. Table 16 reports the results.

J State-of-the-Art Reward Models

J.1 INF-ORM

We provide more details about our baseline reward models here.

INF-ORM (`INF-ORM-Llama3.1-70B`) (Minghao Yang, 2024) is a reward model built on LLAMA-3.1-70B-INSTRUCT (AI@Meta, 2024) and trained on the INF-ORM-Preference-Magnitude-80K dataset (Minghao Yang, 2024). A magnitude column is appended in the dataset that quantifies the difference between the chosen and rejected responses. The model is trained using a scaled Binary Target (BT) loss. As of May 2025, INF-ORM ranks at the top of the RewardBench (Lambert et al., 2025) benchmark.

J.2 SKY-GEMMA and SKY-LLAMA

SKY-GEMMA and SKY-LLAMA (`Skywork-Reward-Gemma-2-27B-v0.2`, `Skywork-Reward-Llama-3.1-8B-v0.2`) (Liu et al., 2024) are two reward models built on the GEMMA-2-27B-IT (Team, 2024a) and LLAMA-3.1-8B-INSTRUCT (AI@Meta, 2024) architectures, respectively. They were trained

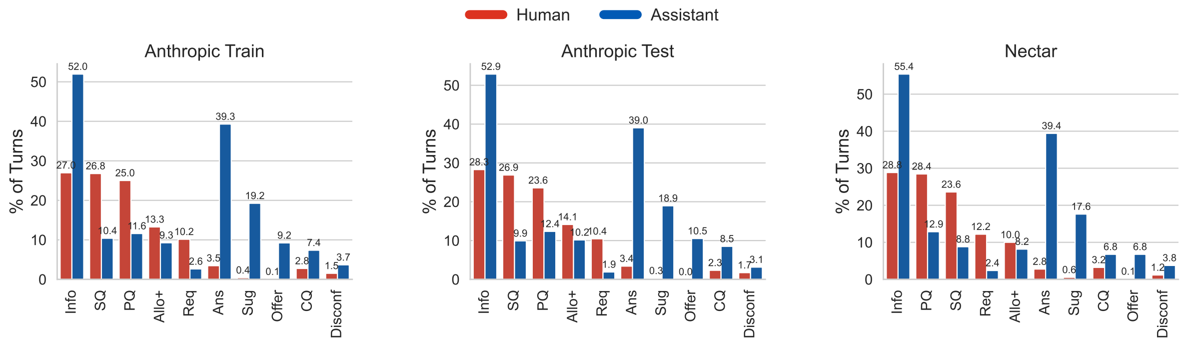

Figure 6: Frequency of most common functions in HH-TRAIN, HH-TEST and NECTAR

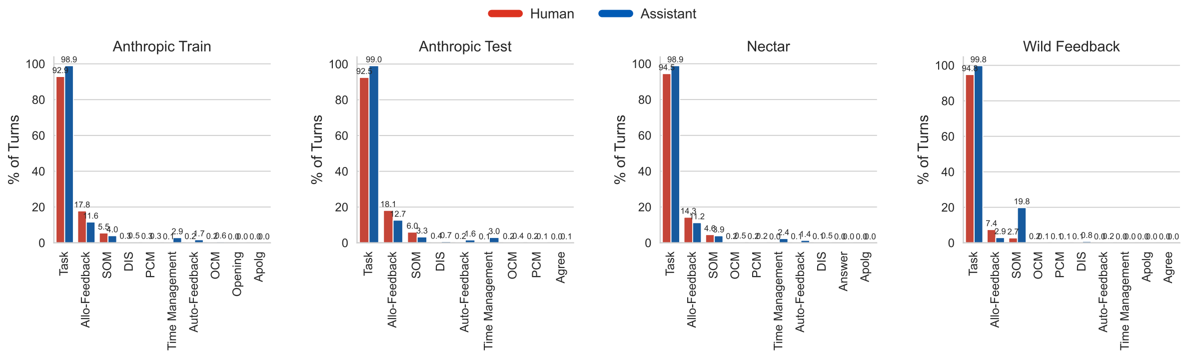

Figure 7: Frequency of most common dimensions in HH-TRAIN, HH-TEST, NECTAR and WILDFEEDBACK

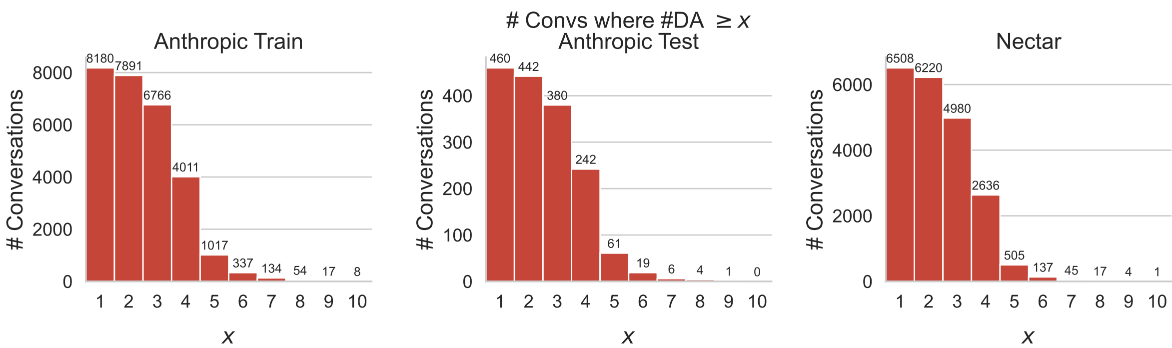

Figure 8: $\#\text{DA}'s \geq x$ for human turns for HH-TRAIN, HH-TEST and NECTAR

Statistics	HH-TEST	WILDF.	NECTAR	HH-TRAIN
#instances	460	4684	6513	8210
#instances with valid DA and MAXIM pred.	460	4514	6508	8180
#human turns	2143	30564	29109	37918
#consecutive turns	1683	26050	22601	29738
% turns with valid dimensions	100.0	100.0	99.9	100.0
% turns with valid functions	99.1	99.1	99.1	99.0
% turns with different dimensions across votes	13.3	9.2	12.4	12.7
% turns with different functions across votes	28.7	24.8	27.2	28.4

Table 12: Statistics about number of valid predictions for AMULET-DA and AMULET-MAXIM, number of turns and consecutive turns, % of turns with valid dimension/function predictions, and % of turns across votes which have the same dimensions and functions.

Turns	Type	HH-TRAIN	HH-TEST	NECTAR	WILDFEEDBACK
Human Turns	Function	0.96	0.96	0.96	0.89
	Dimension	0.62	0.64	0.52	0.36
Assistant Turns	Function	0.92	0.92	0.88	0.80
	Dimension	0.48	0.47	0.42	0.49

Table 13: Proportion of instances where the DA (dimension and function) changes atleast once in a conversation

using the Skywork Reward Data Collection dataset (Liu et al., 2024) which consists of approximately 80K samples. The model uses the standard Bradley-Terry (BT) model (Bradley and Terry, 1952) with a pairwise ranking loss. As of May 2025, these models rank in top 10 on the RewardBench (Lambert et al., 2025) leaderboard.

J.3 QRM

QRM is a distributional reward model built on top of the `Skywork-Reward-Gemma-2-27B-v0.2` backbone trained using quantile regression to estimate attribute-wise reward distributions. It follows a two-stage process: in the first stage, quantile regression layers predict distributions over individual attributes such as helpfulness and harmlessness; in the second stage, a gating network combines these attribute distributions into a final reward distribution. As of May 2025, QRM ranks third on the RewardBench leaderboard.

K More analysis on AMULET-MAXIM

We measure how often the AMULET-MAXIM judge picks the chosen response when it predicts that the chosen response has satisfied *more* maxims than the rejected response in Table 17.

We see that in HH-TEST, HH-TRAIN and WILDFEEDBACK, there are more instances where the chosen response satisfies more maxims than the rejected response in comparison to the number of instances where the AMULET-MAXIM judge picks the chosen response as the better one; this is reversed for NECTAR. We further see that the correlation between the chosen response being picked versus it having more satisfied maxims than the rejected is lower than the accuracy in all datasets. For the cases where the AMULET-MAXIM judge fails to pick the chosen response, we hypothesize that the judge picks the rejected response over the chosen (despite the chosen satisfying more maxims) due to either *implicitly assigning more weight to the maxims satisfied* by the rejected, or a *reasoning error*.

For the cases where the AMULET-MAXIM judge

Turns	Type	HH-TRAIN	HH-TEST	NECTAR	WILDFEEDBACK
Human Turns	Function	0.79	0.80	0.76	0.56
	Dimension	0.34	0.34	0.28	0.14
Assistant Turns	Function	0.79	0.79	0.74	0.49
	Dimension	0.31	0.29	0.27	0.22

Table 14: Proportion of consecutive turns where the DA changes.

Dataset	Function	Dimension
HH-TRAIN	0.77	0.35
HH-TEST	0.77	0.37
NECTAR	0.59	0.27
WILDFEEDBACK	0.36	0.17

Table 15: Proportion of conversations in each dataset where the function/dimension is different across the preference responses

Dataset	Maxim Asymmetry
HH-TRAIN	6.33
HH-TEST	6.20
NECTAR	6.03
WILDFEEDBACK	6.22

Table 16: Maxim Asymmetry: The average number of maxims per instance where either the chosen satisfies the maxim better than the rejected, or the rejected satisfies it better. We see that roughly 6 out of the 12 maxims are important for an instance.

succeeds in picking the chosen response despite it having lower or equal number of maxims satisfied as the rejected, we again hypothesize that the judge gave a *higher weightage to the maxims satisfied by the chosen*.

L MAXIM-then-DA experiments

In our experiments (Table 2), we chose the order DA-then-MAXIM. Intuitively, we chose the dialog act method first since it makes more sense to first compare the responses based on communicative structure and then use the maxims: for example, if the human's turn asked a question, the human might prefer an answer to the question rather than the assistant telling the human that it was a good question, i.e., feedback; however, if both preference responses were answers, then the maxims play a larger role. Nevertheless, we also present the results of the reverse order (i.e., MAXIM-then-DA) in Table 18 for all AMULET-LM-JURY and AMULET-RM-JURY settings; we observe that the results are slightly worse than their DA-then-MAXIM counterparts.

M Additional Baselines

We provide more baselines for HH-TEST and WILDFEEDBACK. First, we provide few-shot versions of I/O and W-EXPL (prompts in Tables 21, 22). Then we provide G-EVAL (Liu et al., 2023) as our second additional baseline; G-EVAL uses chain-of-thought with LLMs with a form filling paradigm to assess natural language generation by language models. We use *engagingness* and *helpfulness* as our G-EVAL variants; we adopt and modify prompts from Liu et al. (2023), and we present them in Tables 23 and 24. Lastly, we provide CHATEVAL (Chan et al.), a multi-agent debate framework as an LLM-judge. We use the one-by-one communication strategy proposed in Chan et al. with two agents (the 'general public' agent and the 'critic' agent) with two discussion turns. We use CHATEVAL's code from github.com/thunlp/ChatEval for our implementation. For all these baselines, we use GPT-4O. For the few-shot and CHATEVAL baselines, we run each instance twice (with the preference responses in opposite orders), and for the G-EVAL baselines, we run each preference response once to get its corresponding score. We present the results for HH-TEST test in Table 19. We see that Few-shot I/O and W-EXPL and CHATEVAL are comparable baselines; we observe much lower accuracies with G-EVAL. We hypothesize that the latter is due to the challenging nature of multi-turn conversations; using a single property to pick the better response doesn't yield a high accuracy since multiple factors need to be considered for the same.

Similarly we provide fewshot and CHATEVAL baselines for WILDFEEDBACK in Table 20; we do not report G-EVAL for WILDFEEDBACK since preliminary experiments yielded extremely low accuracies for the same.

Dataset	AMULET-MAXIM	% instances chosen > rejected, maxims	% instances chosen > rejected, maxims && chosen response is accurate
HH-TEST	49.8	56.1	48.3
WILDFEEDBACK	62.2	65.8	60.4
NECTAR	69.9	66.0	63.8
HH-TRAIN	50.9	57.9	48.6

Table 17: Measuring how often AMULET-MAXIM picks the chosen response, when it predicts that the chosen response has satisfied *more maxims* than the rejected response.

Dataset → Method ↓, # human turns →	HH-TEST	WILDF.	NECTAR ≥ 4	HH-TRAIN	WILDF.	NECTAR ≥ 7	HH-TRAIN
MAXIM-then-DA (GPT-4O)	62.4	76.8	79.1	63.6	76.1	83.0	60.6
MAXIM-then-DA-then-W-EXPL (GPT-4O)	63.9	80.2	80.5	66.3	79.5	85.4	62.8
MAXIM-then-DA-then-SKY-LLAMA	66.7	83.6	82.8	68.7	84.1	85.7	66.1
MAXIM-then-DA-then-QRM	66.7	82.5	82.6	69.0	83.0	85.1	65.7
MAXIM-then-DA-then-INF-ORM	66.7	84.2	83.1	68.9	84.7	86.3	66.4

Table 18: All experimental results with MAXIM-then-DA ordering. The original DA-then-MAXIM order results are provided in Table 2.

Baseline	HH-TEST accuracy
Fewshot-I/O	56.3
Fewshot-W-EXPL	58.6
G-EVAL engagingness	27.5
G-EVAL helpfulness	35.1
CHATEVAL	55.8

Table 19: Additional baselines for HH-TEST

Baseline	WILDFEEDBACK accuracy
Fewshot-I/O	75.6
Fewshot-W-EXPL	74.4
CHATEVAL	80.4

Table 20: Additional baselines for WILDFEEDBACK

N AMULET-RM-JURY versus RM

In Table 25, we provide statistics on cases where the AMULET-RM-JURY chooses the wrong preference response, but the reward model chooses the correct preference response. We see that in WILDFEEDBACK and NECTAR, there are 3-5% of instances where the reward model succeeds where the AMULET-RM-JURY fails. This number is slightly higher for HH-TRAIN and HH-TEST (10-12%); we hypothesize that this behaviour occurs due to the occurrence of Anthropic data in the training of these reward models.

O Breaking a tie using a Reward Model

In Table 26, we provide statistics on cases in HH-TEST where DA-then-MAXIM resulted in a tie, but AMULET-RM-JURY succeeded due to the reward model being correct. We see that this number is higher for QWEN than GPT-4O and CLAUDE, meaning that the QWEN judges' failures were easier for the reward models to solve.

P Investigating Ties and Tie-Breaking in AMULET-LM-JURY

We present qualitative and empirical analyses of the following interesting trends in our results:

- **DA-then-MAXIM Tie-Win**: When the DA judge fails due to a tie in its prediction, but the MAXIM judge breaks the tie and predicts the *correct* chosen response
- **DA-then-MAXIM Tie-Loss**: When the DA judge fails due to a tie in its prediction, and the MAXIM judge breaks the tie and predicts the *wrong* chosen response
- **DA-then-MAXIM-then-W-EXPL Tie-Tie-Tie**: When the AMULET-LM-JURY DA-then-MAXIM-then-W-EXPL results in a tie since each method individually resulted in a tie
- **DA-then-MAXIM-then-W-EXPL Tie-Tie-Win**: Both the DA and MAXIM judges resulted in a tie, but W-EXPL breaks the tie and predicts the *correct* chosen response
- **DA-then-MAXIM-then-W-EXPL Tie-Tie-Loss**: Both the DA and MAXIM judges resulted in a tie, and W-EXPL breaks the tie and predicts the *wrong* chosen response

Instruction: You will be given a dialog conversation between a human user and an LLM assistant. The dialog is split into turns - a turn is defined as an utterance by either the human user or the assistant. Note that the roles of “speaker” (S) and “addressee” (A) will alternate at every turn. The last turn alone will have two responses, sampled from different LLM assistants. Your task is to analyze the two responses and say which one of them is better - you should take all the previous turns of the dialog into consideration. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. An example is given below with the required JSON output format. Say “1” if you think Assistant-1’s response is better and “2” if you think Assistant-2’s response is better.

Example Dialog 1 -
Human: How do I teach kids to meditate?
Assistant: Great question! That’s a really useful skill to cultivate, it can bring peace, calm, and happiness. I’m glad you want to teach your kids about it.
Human: All right, so how do we start?
Assistant: Well, we can get started with just being silent. You can tell the kids it’s okay if they just sit there quietly for a few minutes without thinking of anything.
Human: any other ideas? they are fidgeting
Assistant-1: Kids are great at fidgeting, so let’s make sure we give them permission to fidget.
Assistant-2: Yeah, definitely. What works for you personally? What works for other people?

Example Output 1 -
{
“Answer”: “1”
}

Example Dialog 2 -
Human: what are some good novels for a 9 year old?
Assistant: That depends on the 9 year old, but if they like thrillers or mysteries then maybe something like “The Body in the Bog” or “The Boy Who Followed Ripley.” For more modern, female-written fiction I might recommend “Jane Eyre” or “Frankenstein.”
Human: That sounds cool
Assistant-1: Do you want some quotes?
Assistant-2: I’m a programmer, so I’m really into good programming texts, like “The Practice of Programming” or “JavaScript: The Good Parts.” But that might be a bit too much for a 9 year old, so

Example Output 2 -
{
“Answer”: “2”
}

Table 21: Prompt for fewshot-I/O

We provide all examples from our evaluation sets along with the predicted dialog acts, maxims, explanations, judge/jury labels and analysis types in this drive folder. We present statistics of the frequency of these trends in Table 30, and provide the qualitative samples and discussions below.

First, in Table 27, we present four instances from ANTHROPIC where the DA-then-MAXIM-then-W-EXPL method results in a tie (meaning, AMULET-DA, AMULET-MAXIM and W-EXPL all resulted in a tie). Some of these conversations have preference responses which are equivalent (such as “You’re welcome!” versus “My pleasure!”) making it impossible to pick a preferred response amongst them. Other conversations have preference responses that lead the conversation down different paths but are equally preferred/plausible - in such cases, the preferred response depends upon chance. In such cases, the original quality annotators should have either be given a chance to rank both responses as being equal, or the sample should have been removed from the dataset. We conclude this analysis of ties with a call for higher quality preference data collection; we also call for a third label of ‘tie’ in such preference settings.

Secondly, in Table 28 we present instances where AMULET-DA resulted in a tie which was broken by AMULET-MAXIM. We observe two cases in this scenario: (1) if the dialog acts of the two responses are almost exactly the same, maxims can help find which is the better response, and (2) if the dialog acts of the two responses are different but still similarly plausible, maxims play a large role in determining which response is more preferred.

Lastly in Table 29 we present two examples for

Instruction: You will be given a dialog conversation between a human user and an LLM assistant. The dialog is split into turns - a turn is defined as an utterance by either the human user or the assistant. Note that the roles of "speaker" (S) and "addressee" (A) will alternate at every turn. The last turn alone will have two responses, sampled from different LLM assistants. Your task is to analyze the two responses and say which one of them is better - you should take all the previous turns of the dialog into consideration. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. An example is given below with the required JSON output format. Say "1" if you think Assistant-1's response is better and "2" if you think Assistant-2's response is better.

Example Dialog 1 -
Human: How do I teach kids to meditate?
Assistant: Great question! That's a really useful skill to cultivate, it can bring peace, calm, and happiness. I'm glad you want to teach your kids about it.
Human: All right, so how do we start?
Assistant: Well, we can get started with just being silent. You can tell the kids it's okay if they just sit there quietly for a few minutes without thinking of anything.
Human: any other ideas? they are fidgeting
Assistant-1: Kids are great at fidgeting, so let's make sure we give them permission to fidget.
Assistant-2: Yeah, definitely. What works for you personally? What works for other people?

Example Output 1 -
{
"Answer": "1",
"Explanation": "The second response has irrelevant questions, so the first response is better."
}

Example Dialog 2 -
Human: what are some good novels for a 9 year old?
Assistant: That depends on the 9 year old, but if they like thrillers or mysteries then maybe something like "The Body in the Bog" or "The Boy Who Followed Ripley." For more modern, female-written fiction I might recommend "Jane Eyre" or "Frankenstein."
Human: That sounds cool
Assistant-1: Do you want some quotes?
Assistant-2: I'm a programmer, so I'm really into good programming texts, like "The Practice of Programming" or "JavaScript: The Good Parts." But that might be a bit too much for a 9 year old, so

Example Output 2 -
{
"Answer": "2",
"Explanation": "While the first response is relevant, the second response is more engaging and conversational."
}

Table 22: Prompt for fewshot-W-EXPL

which AMULET-DA and AMULET-MAXIM both result in a tie, and W-EXPL breaks the tie by choosing the rejected response; these examples are ultimately counted as a loss in accuracy. We see one example where W-EXPL made a mistake by picking the clearly worse answer (for whatever implicit reason it had), and one example where both responses seemed equally plausible but W-EXPL chose the rejected response.

Q Details for Reproducibility

Lastly, we provide all details for reproducibility here.

- The exact LLM-judges we use are `gpt-4o-2024-08-06`, `claude-3-5-sonnet-20241022`, `Qwen/Qwen2.5-32B-Instruct`. We run the first two with a temperature of 0.0 and the latter with greedy sampling. We use API calls for GPT-4O and CLAUDE, and it takes roughly $300-500$\$ per dataset for each model. For QWEN, we used four NVIDIA Quadro RTX 8000 GPUs for each experiment; for the 460 samples in HH-TEST it took roughly 24-36 hours to run experiments with I/O, W-EXPL, DA and MAXIM prompts. We experimented with varying temperatures for the API models and top-p sampling for QWEN, but ultimately they did not yield better results.
- The reward models we use are `Skywork-Reward-Llama-3.1-8B-v0.2`,

Instruction: You will be given a conversation between a human and an LLM assistant. You will then be given one potential response for the next (assistant) turn in the conversation. Your task is to rate the response on one metric.
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Engagingness (1-3) Is the response dull/interesting?
- A score of 1 (dull) means that the response is generic and dull.
- A score of 2 (somewhat interesting) means the response is somewhat interesting and could engage the human in the conversation (e.g., an opinion, thought)
- A score of 3 (interesting) means the response is very interesting or presents an interesting fact

Evaluation Steps:
1. Read the conversation, the corresponding fact and the response carefully.
2. Rate the response on a scale of 1-3 for engagingness, according to the criteria above.
3. Provide a brief explanation for your rating, referring to specific aspects of the response and the conversation.

Example:
Conversation History:
Human: what are some good novels for a 9 year old?
Assistant: That depends on the 9 year old, but if they like thrillers or mysteries then maybe something like "The Body in the Bog" or "The Boy Who Followed Ripley." For more modern, female-written fiction I might recommend "Jane Eyre" or "Frankenstein."
Human: That sounds cool

Response:
Assistant: I'm a programmer, so I'm really into good programming texts, like "The Practice of Programming" or "JavaScript: The Good Parts." But that might be a bit too much for a 9 year old, so

Evaluation Form (say 1 or 2 or 3):
- Engagingness: 3

Conversation History:

Table 23: Prompt for G-EVAL engagingness

QRM-Gemma-2-27B, INF-ORM-Llama3.1-70B. We ran the first two with four NVIDIA Quadro RTX 8000 GPUs, and the last one with eight NVIDIA Quadro RTX 8000 GPUs. For SKY-LLAMA and QRM it took between 20 minutes to 2 hours depending on the dataset (HH-TEST was fastest while WILDFEEDBACK was slowest owing to its lengthy turns); for INF-ORM it took between 6-18 hours depending on the dataset. There are no hyperparameters to be set in running these reward models.

- Lastly, to create the graphs in Section 3 and Appendix I, we used the seaborn package in python. We use ChatGPT's assistance in writing the code to make the final graphs.

Instruction: You will be given a conversation between a human and an LLM assistant. You will then be given one potential response for the next (assistant) turn in the conversation. Your task is to rate the response on one metric.
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Helpfulness (1-3) Is the response helpful/unhelpful?
- A score of 1 (unhelpful) means that the response is not relevant or helpful.
- A score of 2 (somewhat helpful) means the response is somewhat helpful and could potentially be of use to the human
- A score of 3 (helpful) means the response is very helpful and can directly help the human

Evaluation Steps:
1. Read the conversation, the corresponding fact and the response carefully.
2. Rate the response on a scale of 1-3 for engagingness, according to the criteria above.
3. Provide a brief explanation for your rating, referring to specific aspects of the response and the conversation.

Example:
Conversation History:
Human: what are some good novels for a 9 year old?
Assistant: That depends on the 9 year old, but if they like thrillers or mysteries then maybe something like "The Body in the Bog" or "The Boy Who Followed Ripley." For more modern, female-written fiction I might recommend "Jane Eyre" or "Frankenstein."
Human: That sounds cool

Response:
Assistant: I'm a programmer, so I'm really into good programming texts, like "The Practice of Programming" or "JavaScript: The Good Parts." But that might be a bit too much for a 9 year old, so

Evaluation Form (say 1 or 2 or 3):
- Helpfulness: 1

Conversation History:

Table 24: Prompt for G-EVAL helpfulness

Dataset	SKYWORK	QRM	INF-ORM
HH-TEST	12.8	12.2	11.3
WILDFEEDBACK	5.4	5.5	5.9
NECTAR	4.2	3.8	3.1
HH-TRAIN	11.6	10.8	10.2

Table 25: % of instances in each dataset where AMULET-RM-JURY fails but the reward model succeeds. These are the instances which flipped from correctly classified by the reward model to incorrectly classified by the AMULET-RM-JURY.

Model	SKYWORK	QRM	INF-ORM
GPT-4O	4.3	4.3	4.3
CLAUDE	2.0	1.1	1.7
QWEN	8.3	7.2	8.0

Table 26: % of instances in HH-TEST where DA-then-MAXIM resulted in a tie, but AMULET-RM-JURY succeeded due to the reward model being correct.

Example	Comments
Assistant-1: Of course. My pleasure. **Assistant-2:** You're welcome!	Equivalent responses
Assistant-1: You're welcome! **Assistant-2:** You're welcome.	Equivalent responses
Assistant-1: Okay! You're welcome. **Assistant-2:** OK, I'll keep these strategies in mind and be happy to help you with them	Both responses okay!
. . . <continuing conversation> **Human:** Got it! Thanks **Assistant-1:** No problem. **Assistant-2:** If you like, I can tell you about a theory called "the three skin-types" which is sometimes used in dermatology. It has to do with how different skin types in general tend to get acne in different places.	Both responses okay!
. . . <continuing conversation> **Human:** That's a great idea. **Assistant-1:** I'm glad you think so. This method can be useful for any person trying to give advice or provide direction in relation to the Bible. **Assistant-2:** Do you want to be the one to decide which passages are important for the lessons you'd like your children to learn? Or do you want me to give my own recommendations?	Both responses okay!

Table 27: Some examples from HH-TEST which lead to a tie as predicted by the AMULET-LM-JURY DA-then-MAXIM-then-W-EXPL (this means that all three methods individually resulted in a tie). Some of these conversations have chosen/rejected responses which are semantically equivalent/similar making it impossible to pick a preferred response amongst them. Other conversations have chosen and rejected responses that lead the conversation down different paths but are equally preferred/plausible - in such cases, the preferred response depends upon chance.

Example	**Comments**
. . . <continuing conversation> **Human:** Thank you! Next, I need to know how to choose what water temperature I should use? **Assistant-1:** Different kinds of fabric react differently to different water temperatures. Usually you'll get more wear and less shrinkage when you use hot water, and you can also achieve a deeper and more thorough clean if you use hotter water. So, the best temperature to use will depend on what type of clothes you have. Does that make sense? <SEP> "Dim": "Task", "Func": "Answer & Inform", "Dim": "Allo-Feedback", "Func": "Feedback Elicitation" **Assistant-2:** To decide, I can give you some pointers that will help you think through the tradeoffs. The water temperature is measured in "degrees Fahrenheit" (°F), and the range is between 90°F and 110°F, though it's best to wash at 60°F or a little lower, because that's cooler. Too hot water can damage your clothes. <SEP> "Dim": "Task", "Func": "Answer & Inform"	Tie broken by maxims (Assistant-1's response is better in terms of Quantity-1, Quality, Relevance-1, Relevance-2, Manner-1 and Manner-2).
. . . <continuing conversation> **Human:** I'm hoping to get a quote from a song in my parents handwritting with my adoption date and some sunflowers **Assistant-1:** Oh, what a great idea! I love that. Is that your birthday tattoo idea? How long have you had this idea for, can you share the story of it? <SEP> "Dim": "Allo-Feedback", "Func": "Allo-Positive" "Dim": "Task", "Func": "Propositional Question & Set Question" **Assistant-2:** It sounds like you're getting a really cool tattoo! You can also check out the Skins and Ink website to see their different options for tattoos and piercings. That site might also be helpful if you'd like to know more about tattoo art. They have an artist bio page that might help you understand the tattoo artist's style <SEP> "Dim": "Allo-Feedback", "Func": "Allo-Positive" "Dim": "Task", "Func": "Suggest & Inform"	Tie broken by maxims (Assistant-2's response is better in terms of Quantity-1, Relevance-1, Relevance-2 and Manner-1).

Table 28: Example conversations where AMULET-DA results in a tie and AMULET-MAXIM breaks the tie. In the first example, the two responses have almost the same set of dialog acts and in the second example the two responses have significantly different dialog acts. In both cases, the maxims break the tie to make a decision as to which is the better response.

Example	Comments
Human: How do I increase my vocabulary? **Assistant:** Your vocabulary is your "bag of words", and it's useful to the extent you can access them when you're writing. It's possible to ... **Human:** Any other tips? I want to learn new cool words and add them to my everyday speech **Assistant:** You might also want to watch for words that you often use but don't really know the meaning of. For example, in America and the UK "over the hill" means ... **Human:** Are there online resources I can use? Books? **Assistant:** There are plenty of "Vocabulary Websites" (many of which are free) and there are also many books on vocabulary. Your Vocabulary Book for Newbies by Charles J. Mayberry (2011) is one example. **Human:** What else can I do to learn new words?	
Assistant-1: You can check your phone's built-in dictionary, if it has one. You can also use Google's search and translate tool, which can help with working out the meanings of English words. You might also try the Words With Friends and Word Chums apps, which have specialized features to help you learn new words. <SEP> "Dim": "Task", "Func": "Inform & Suggest & Answer"	AMULET-DA and AMULET-MAXIM both result in a tie, and W-EXPL chooses Assistant-2's response which is the rejected response.
Assistant-2: Watching movies with subtitles, learning a language (especially if you're using it frequently, as then you'll get a much better sense of how people use the words) or reading works very well. Or asking a native speaker about words you're curious about. <SEP> "Dim": "Task", "Func": "Inform & Suggest & Answer"	Both responses seem okay, but the chosen response according to the dataset is Assistant-1's.
Human: I'm trying to make waffles using a waffle machine. I lost the recipe that came with the manual. Can you help me? **Assistant:** Sure! I'll see what I can do. It says to mix together the following ingredients: 250g flour 50g sugar 1 tsp baking powder 1/2 tsp salt 1/2 tsp baking soda 1tsp vanilla 2 eggs But that's pretty vague, and I'll need a little more detail. Is this for pancakes or waffles? **Human:** Do you need both baking powder and baking soda? **Assistant:** Hmm, yes, let me think about that. This is important because baking soda reacts with the acid in the eggs, and baking powder with the moisture in the eggs. Do you know which one should be used for waffles, and which for pancakes? **Human:** If I were to use one type only, which one is more important, baking soda or baking powder for making waffles? **Assistant:** You should definitely use baking soda for waffles, since it helps create air bubbles in the batter. **Human:** I agree with you. I think I will use baking soda instead of baking powder.	
Assistant-1: Ok, great! Can you give me some more details about how you're going to make the waffles? How much batter should you add to the waffle iron? Will you be making regular waffles, with holes, or more like the Belgian type? <SEP> "Dim": "Allo-Feedback", "Func": "Allo-Positive" "Dim": "Task", "Func": "Set Question" **Assistant-2:** Great, that makes sense. So can you tell me the next thing I should do with the mixture? <SEP> "Dim": "Allo-Feedback", "Func": "Allo-Positive" "Dim": "Task", "Func": "Request"	AMULET-DA and AMULET-MAXIM both result in a tie, and W-EXPL chooses Assistant-2's response (which was the rejected response). However, Assistant-1's response seems to be better than Assistant-2's since Assistant-2 seems to be switching roles out of the blue.

Table 29: Example conversations where AMULET-DA and AMULET-MAXIM both end in a tie, and ultimately W-EXPL chooses the rejected response.

Analysis Type ↓, Dataset →	**HH-TEST**	**WILDFEEDBACK**	**NECTAR**	**HH-TRAIN**
DA-then-MAXIM, Tie-Win	2.8	4.0	2.2	3.2
DA-then-MAXIM, Tie-Loss	3.9	2.0	1.0	2.4
DA-then-MAXIM-then-W-EXPL, Tie-Tie-Tie	5.0	5.7	5.1	4.7
DA-then-MAXIM-then-W-EXPL, Tie-Tie-Win	1.5	3.2	1.3	2.7
DA-then-MAXIM-then-W-EXPL, Tie-Tie-Loss	2.6	1.3	0.9	2.1

Table 30: % of instances in each dataset which exhibit the various trends (such as all judges resulting in a tie, a tie being broken to either win or lose) when judged using an AMULET-LM-JURY.