[SHORT] EVALUATING LANGUAGE MODELS IN REALISTIC CONVERSATIONAL CONTEXTS

Anonymous authors
Paper under double-blind review

ABSTRACT

As Large Language Models (LLMs) are increasingly deployed to serve open-ended and realistic human-AI interactions, evaluating conversational quality at human scale has become a central challenge. Existing evaluation frameworks built for summarization, translation, or short-form QA tasks fall short of adequately measuring the consistency of human-scale dialogue, especially when derivation and validation of these metrics themselves often rely on synthetic rather than human sources. We fill the gap by introducing **UPHELD** (Utility & Planning Human-Scale Evaluated Long Dialogues), a large, reference-full benchmark for evaluating human-scale conversational ability beyond factual correctness. UPHELD consists of hundreds of multi-turn human-to-human dialogues authored by **professional script writers**, with realistic turn densities and **36,000+ per-turn human annotations** across **10,000+ expert-generated dialogue turns**. We also show that naive quality metrics like LLM-as-a-judge perform poorly on UPHELD, but it can be used as a fine-tuning dataset or a validation dataset to develop more robust LLM evaluation metrics in these settings. Overall, UPHELD provides a robust, human-grounded foundation for evaluating long, human-scale conversational intelligence that fills a crucial gap in the pre-existing LLM dataset landscape.

1 INTRODUCTION

The rapid advancement of Large Language Models (LLMs) has revolutionized text generation in focused, single-turn tasks like question+answer. However, as LLMs are increasingly integrated into complex systems, the frontier of evaluation has shifted toward open-ended, multi-turn settings. In these environments, conversations are longer and more open, and success is not merely defined by factual correctness but by the model's ability to maintain a natural, coherent, and useful interaction with a human end-user over an extended dialogue.

Existing evaluation metrics, largely inherited from pre-LLM natural language tasks such as machine translation, summarization, and question answering, focus on factual precision and lexical overlap. These metrics generally struggle with conversational language, which rely on softer, longer-range qualities like coherence and engagement across multiple turns. While recent benchmarks have evaluated longer dialogues through reference-free "LLM-as-a-judge" evaluations Zheng et al. (2023); Liu et al. (2023); Dubois et al. (2023) or human preference platforms Chiang et al. (2024), these methods face critical limitations. LLM judge quality often fails to exceed that of non-expert human judges Bavaresco et al. (2024) Krumdick et al. (2025), and such benchmarks are increasingly susceptible to training data leakage Mirzadeh et al. (2024) and poor reliability on out-of-distribution tasks Krumdick et al. (2025). Additionally, many of the metrics are tested primarily on synthetic datasets, which can introduce significant biases in metric results due to empirically verified limits in LLM dialogue generation Wang et al. (2025) and persistent biases in LLM-simulated data used for both training and evaluation Rahmani et al. (2025). Our work presented here mitigates these issues through a specialized benchmark evaluating models on conversations, which are designed to be: a) **Multi-turn and human-to-human.** Emphasizing naturalistic flow of real interaction rather than simulated or distilled outputs; b) **Not anchored to specific external knowledge.** Shifting the focus from retrieval-based accuracy to conversational intelligence; and c) **Not anchored to a specific outcome.** Allowing for fluid, open-ended dialogue without focusing on a "right answer" (e.g. coding, summarization), but still having an overall goal (e.g., "tell me how to make a pizza").

We introduce **UPHELD (Utility & Planning Human-Scale Evaluated Long Dialogues)**, the first collection of conversations that fulfill these criteria. UPHELD contains over 10,000 high-quality human-written turns of dialogue across hundreds of novel conversations, and for each turn we present multiple human labels along important dialogue coherence criteria like stylistic and content consistency. Unlike datasets extracted from noisy sources Lison & Tiedemann (2016) or those relying on crowdsourced platforms where annotator backgrounds are unknown Zhang et al. (2018), UPHELD utilized **professional script writers** to guarantee high conversational quality, which has been shown to lead to notable improvements in the quality and authenticity of dataset collection Pilan et al. (2024); Elisha et al. (2024).

Our strategy for data collection and general approach allows UPHELD clear divergence from pre-existing dialogue benchmarks in three fundamental ways (a more detailed point-by-point overview is given in Appendix A):

- **Task-oriented, natural dialogue:** UPHELD focuses on task-oriented dialogues (e.g., math tutoring or trip planning), but all professional script writers were instructed to write dialogues in a natural, casual way. This deviates significantly from standard task-oriented datasets where agents follow a rigid "set" target or are constrained by encoded model knowledge. By providing writers the freedom to determine how the task is navigated, we capture a more naturalistic intersection of utility and human-like interaction.
- **Realistic conversational turn density:** While some datasets focus on extremely long conversations, such as interviews or LoCMo Maharana et al. (2024), most human-to-human interactions are more concise and especially in the guided task-oriented space. Our analysis of human-to-human interaction in available datasets shows that real-world dialogue rarely exceeds 20 turns. Appendix A shows that the average of fully human-collected datasets is 9.3 turns, while the Topical-Chat dataset Gopalakrishnan et al. (2019) maxes at 21 turns. Our writers were given the freedom to determine conversation lengths and end at natural stopping points to keep the interactions as natural as possible, which means UPHELD better reflects real-world turn distributions.
- **Emphasis on conversational fluency versus information retrieval:** We deliberately avoid tasks requiring complex context, such as complex mathematics or coding. Instead, UPHELD focuses on the model's ability to hold a coherent and reasonable conversation throughout an entire dialogue, and in service of everyday tasks that do not require heavy context engineering. This distinguishes our work from benchmarks like Wizard of Wikipedia Dinan et al. (2018), which prioritize knowledge-grounding over pure conversational dynamics.

Our main contributions are as follows:

- We present UPHELD, a novel dataset of 10,000+ dialogue turns crafted by **professional script writers**. With **36,873 dense, per-turn human labels** that rate turn quality along various coherence criteria, UPHELD offers an evaluative scale nearly 10x that of comparable human-verified benchmarks Zhang et al. (2023), a volume usually reserved for crowdsourced or synthetic datasets like in Elisha et al. (2024); Rahmani et al. (2025).
- We focus on realistic task-oriented dialogue with realistic length distributions rather than casual, open-ended dialogue. In doing so, UPHELD effectively evaluates functional utility within naturalistic human cognitive boundaries, avoiding the artificial turn-length inflation common in synthetic datasets Wang et al. (2025).

2 Related Work

LLM evaluation is dominated by task-based benchmarks (e.g., Open LLM leaderboard (Myrzakhan et al., 2024), IFEval (Zhou et al., 2023), BBH (Srivastava et al., 2022), MATH (Hendrycks et al., 2021), GPQA (Rein et al., 2023), HELM (Liang et al., 2023)) and QA-style evaluations (MUSR (Sprague et al., 2023), MMLU-PRO (Wang et al., 2024)), which rely on input-output pairs, exact match or F1-style metrics, and single-turn right/wrong factual comparison. LLM-as-judge methods (Zheng et al., 2023; Zhang et al., 2023; Duan et al., 2023; Adlakha et al., 2024) enable evaluation without ground truth but remain tied to these paradigms. Multi-turn and conversation-oriented work (Chatbot Arena (Chiang et al., 2024), MultiHop QA (Schnitzler et al., 2024), MT-eval (Kwan et al., 2024)) is less prominent on leaderboards. These approaches suit discrete tasks and factual extraction but are ill-suited to long, non-topical conversations, which require sustained coherence, context over many turns, and handling of subjective or open-ended dialogue. Existing long-form, non-factual datasets (MuTal (Cui et al., 2020), Topical-chat (Gopalakrishnan et al., 2019), DailyDialogue (Li et al., 2017)) often lack authentic human interaction or are constrained to predefined topics. A deeper comparison of these works to UPHELD can be found in Appendix A.

3 The UPHELD dataset

We first collected conversational data by tasking a panel of professional writers hired from the Upwork freelancer marketplace to create open-ended, naturalistic dialogues across a diverse range of topics like customer service and education. The writers were explicitly instructed to create longer dialogues characterized by conversational complexity and exploratory interaction, rather than producing conversations with predetermined or narrow outcomes. For example, the writers developed dialogues exploring scenarios such as mobile phone selection, travel planning, or problem-solving discussions that were goal-oriented and where multiple nuanced exchanges would be appropriate. Writers were also instructed to avoid overly scripted or linear communication patterns. This approach ensures that the dataset captures the diverse and intricate nonlinear nature of human conversation. We also took steps to alleviate any potential writer bias (see details in Appendix D). The UPHELD dataset is publicly available under a permissive CC-BY-4.0 license `https://gofile.io/d/O3yk7Z`, while GitHub access is hidden until the review process is ongoing.

chat_history	*user*: I'm shopping for a new car and can't decide what to get. *assistant*: I can help with that. What size car are you looking for?	*user*: I'm really frustrated. I've been trying to get this printer to work all morning and it just won't print anything! *assistant*: I understand that must be really annoying. Let's work together to fix this. Can you tell me if you see any error messages on your printer or computer? *user*: No error messages. It just sits there doing nothing whenever I try to print! *assistant*: Alright, let's try a few steps to see if we can get it responding. First, could you check whether the printer is set as the default printer on your computer? *user*: How do I check that?
Option A	The Ford F-150 is one of the top selling trucks in the country. Have you taken one of those for a test drive?	If you're using a Windows PC, you can go to 'Devices and Printers' in your control panel. For Mac, you can check under 'System Preferences' and then 'Printers & Scanners'.
Option B	Trucks are great for hauling and towing. Here are a few questions to help narrow down your options: 1. What will you be using the truck for most often? (e.g., work, off-roading, daily driving, towing a trailer) 2. How many passengers will the truck need to seat? 3. Do you have a preferred fuel type (gasoline, diesel, hybrid)? 4. What's your budget for the truck? 5. Do you have a specific brand or model in mind? 6. Do you need any specific features, such as 4-wheel drive, a certain bed size, or advanced safety features	It's easy. On a Windows computer, you can go to the Control Panel, then Devices and Printers (or Printers and Devices in some versions). Look for your printer's name and see if it has a checkmark next to it that says "Default printer". If it doesn't, right-click on the printer and select "Set as default printer". On a Mac, you can go to System Preferences, then Printers & Scanners. Look for your printer's name and make sure it's selected as the default printer. If you're still having trouble, we can try some other steps. Have you tried restarting the printer
score_task_1_content	1: Strongly Disagree – no content overlap	3: Neutral – some overlap of content
score_task_2_style	1: Different styles	1: Different styles
score_task_3_reasonableness	1: Not a reasonable continuation	5: A reasonable continuation

Table 1: Input conversations and associated annotator ratings from a single annotator (all scores are out of 5). More examples can be found in Appendix N.

3.1 INPUT DATA

Given our initial set of rich natural language conversations, various LLM models were then used to output candidate completions at every turn of every conversation. Specifically, models were presented with conversation history up to a specific point, with the next human-written turn withheld. Models then generated a predicted next turn. To generate the "targets" human annotators compared the ground truth we used the following models: a) **GPT-3.5** (OpenAI, 2023); b) **GPT-4o** (OpenAI, 2024); c) **Llama-3.1-70b** (Grattafiori et al., 2024); d) **custom model** (a fine-tuned version of Llama-3.1. In addition to these, we added two baseline candidate completions: a) a version of the ground truth (human) output rephrased using GPT-4o and b) a random completion. These two were used to calibrate our data, and as expected we found that the rephrased dataset received the highest while the random dataset received the lowest marks when evaluated by a human labeler. These checks provide us with additional confidence that humans are fair judges of our evaluation axes – style, content, and reasonableness. At every conversational turn, annotators were then given (1) the chat history up to that point, (2) the ground truth human completion (Option A), and (3) one of the LLM-generated candidate completions (Option B). The provenance of options A and B were not disclosed to the annotators. Their task was to compare these continuations on content consistency, style consistency, and general reasonableness. The tasks, metric scales and instructions were fine-tuned through two (paid) pilot user studies. Example data-points within UPHELD are given in Table 1, and full instructions provided to the annotators can be found in Appendix C.

3.1.1 ANNOTATION DIMENSIONS

Each model-generated turn was evaluated by five annotators across three dimensions (details in Appendix C):

- **Content Equivalence**: 5-point Likert scale measuring semantic similarity to the reference.
- **Style Equivalence**: 3-point Likert scale assessing linguistic and stylistic alignment to the reference.
- **Reasonableness**: Binary scale (unreasonable vs. reasonable) evaluating reference-free coherence.

3.1.2 DATA STATISTICS

In total we collected complete evaluation labels for 400 turns stemming from 53 human-written conversations collected for this study and additional seven conversations from existing datasets (Cui et al., 2020) to serve as control points. The dataset was curated to include approximately 90 % human-written conversations on a wide range of topics and situations, supplemented with approximately 10 % from available datasets to introduce challenging cases. Each predicted turn was evaluated by five human annotators, and each annotator judge labeled between 1,220 and 1,230 conversations. Overall, we generated 12,291 sets of labels, or 36,873 labels. The ground truth conversations consist of 5.2 turn pairs (user-assistant) or 10.4 dialogue turns on average. The average length of the conversation history annotators analyzed was 560 characters and the length of the judged turns was on average 245 characters.

3.2 Verification datasets

To further validate our findings, we construct additional verification datasets by augmenting LLM-Arena **?** and Topical-Chat (Gopalakrishnan et al., 2019). The overall procedure consisted of three steps: a) deriving a single "ground truth answer" from each data point of each existing dataset (see below), b) generating an alternative continuation with GPT-4o, and c) collecting human judgments following the UPHELD annotation protocol. The procedure followed for generating the ground truth for the two datasets is outlined in Appendix K. We include all additional verification labels within our dataset for reproducibility. After standard quality control (i.e. filtering for missing data and badly formatted inputs), we obtained 12,305 pairwise preference judgments. We note that both these verification datasets, although useful for verification, are still relatively deficient in freeform human-to-human interaction and focus on a limited set of pre-defined topics. As such they should be treated as verification datasets only and not as valid replacements for UPHELD.

3.3 Reference-free vs human-grounded metrics

Fundamentally, UPHELD uses a ground truth reference to generate our human labels. In contrast, reference-free evaluation of LLMs Liu et al. (2023) relies on human preference, and LLM outputs can reproduce these preferences, indicating their correlation with human judgments (Zhang et al., 2023). However, reference-free preference datasets also incur significant limitations, such as poor performance at judging dialogues and weakened judgment reliability on out-of-distribution tasks (Krumdick et al., 2025). UPHELD is designed specifically to tackle these limitations by adopting a reference-full approach. However, the this approach raises a question around dialogue multiplicity: a single input might incur multiple valid outputs, so how are we sure our ground truths are well defined? UPHELD's design reduces susceptibility to this issue in two ways: (1) two of our key label categories (style and reasonableness) are well-defined even with dialogue multiplicity, and (2) UPHELD dialogues primarily revolve around task-oriented settings, which means content accuracy is a well-defined metric. For example, while opinion-oriented conversations (e.g. *Who makes the best Caesar salad?*) are susceptible to dialogue multiplicity, our task-oriented dialogues (e.g. *How to make a Caesar salad?*) are not. To quantify this effect, we ran the following experiment exploring multiplicity.

3.4 UPHELD and Conversational Multiplicity

A valid concern may be that direct comparisons to a reference human answer may be inappropriate in settings when a particular prefix can lead to a multiplicity of valid responses. This effect may be prevalent especially when the prefix is asking for an opinion (e.g. "What is your favorite animal?"). We, however, observe that UPHELD dialogues avoid this potential pitfall as they are not strictly freeform, but are all targeted towards completion of a specific well-defined task. In this context, there is some notion of correct ground truth, and we specifically hired professionals who are experts at these tasks (see Appendix D). To put it simply, our task setting is analogous to the difference between *what kind of salads do you like?* (which has ambiguity and dialogue multiplicity) and *how do I make a Caesar salad?* (which is much more constrained and has a more well-defined ground truth). To quantify this, we generated 100 open-ended questions (GPT-4o) and then generated two possible completions with GPT-4o at moderately high temperature ($\tau = 1$) to those questions. We did the same with 100 UPHELD turn completions. We then asked GPT-4o to judge whether the two possible completions contain similar content. The results are as follows:

Dataset	Open-Ended	UPHELD (first turn only)	UPHELD (random turn)
Semantic Consistency (%)	74	93	93

Table 2: Semantic Consistency Performance across Different Datasets

UPHELD exhibits higher semantic consistency in the output, suggesting that UPHELD dialogues admit much less conversational multiplicity than more freeform datasets. The results show that UPHELD dialogues admit significantly higher output consistency (93%) compared to freeform dialogues (74%), demonstrating that our reference-full approach still allows us to collect meaningful labels on ground truth content overlap. This supports our hypothesis that targeted task-focused conversations like those in UPHELD admit well-defined "ground truth" references. UPHELD maintains high output consistency even when we only analyze the first turn, which is where we would expect more branching/multiplicity during a dialogue. Note that these results are likely an underestimate of the true consistency, since sampling multiple LLM outputs would induce additional randomness that likely would not exist within natural human dialogue.

3.5 Experiments with UPHELD

We use UPHELD to conduct two experiments, one to explore the use of the dataset for fine-tuning an LLM, and the other to calibrate the metrics used for LLM evaluation. In both experiments we observed that UPHELD improves the results – fine-tuned model is scored 40% higher than other non-fined tuned models Figure 1; and we were able to show that UPHELD can be used to learn an Mixture-of-Judges t metrics which achieve 30% higher correlation with human judgments as compared to commonly used metrics. Table **??** summarizes the differences between UPHELD developed and the best traditional metrics, and for full table comparing multiple metrics see Appendix B.

	Content		**Style**		**Reasonableness**	
Dataset	**Trad.**	**Ens.**	**Trad.**	**Ens.**	**Trad.**	**Ens.**
UPHELD	0.58	**0.63**	0.41	**0.61**	0.17	**0.52**
LLM Arena	0.50	**0.50**	0.51	**0.54**	**0.31**	0.20
Topical-Chat	0.57	**0.58**	0.36	**0.55**	**0.13**	0.04

Table 3: Comparative performance across datasets. For each task, the best performing traditional/LLM metric (Trad.) is compared against our best Ensemble model (Ens.).

As each conversation turn was independently labeled by five human annotators, we analyze model responses that elicited stronger human-human agreements. We observed moderate levels of consensus as shown in Table 4. As expected, the agreement is highest on the random and gpt4_rephrase baselines, while being consistent across all other models. Further analysis on labeler agreement can be found in Appendix J. We observed that approximately 25% of data points had full agreement across all 5 human judges. Otherwise, we bin the level of agreement as follows: agreement across 2 out of 5 labels represents a “plurality;" agreement across 3 or 4 labels represents a “majority." We further quantify how well an LLM-as-a-judge evaluator agrees with this winning score relative to the agreement bin. Table 5 shows that LLM-as-a-judge performance is heavily correlated to agreement level amongst the human labelers. This result demonstrates that human-human disagreement is a valid measure of data difficulty, and this uncertainty signal present in UPHELD may be integral in further evaluation metric development.

models:	gpt3_5	gpt4	gpt4_rephrase	llama3.1-70b-base	random	custom-model
Kappa score	0.30	0.33	0.46	0.38	0.97	0.35

Table 4: Kappa score ranges from -1 to +1, with +1 indicating the maximum level of agreement among the annotators.

	LLM-as-a-judge				
Agreement	**Human**	**Likert**	**(explain)**	**Binary**	**(explain)**
Plurality	0.47	0.34	0.36	0.755	0.72
Majority	0.69	0.38	0.39	0.80	0.78
Full	1.00	0.63	0.56	0.93	0.92

Table 5: LLM-as-a-judge performance on the UPHELD dataset for different dataset splits, grouped by the level of human annotator agreement within each bucket.

4 Conclusion

In this work, we introduced UPHELD: a dataset designed to train and evaluate LLMs for human long-form conversational settings. We collected tens of thousands high-quality human-annotated labels on crucial consistency metrics within long-form conversations, along with the high-quality conversations themselves. As is, we believe it is already of interest to the community, but we also plan to expand the dataset into more conversational verticals, especially ones that focus on enterprise applications such as targeted customer service and dialogues around more technical topics. Gathering these datasets also means we must depend on human annotators performing subjective evaluations such as determining "style". This approach yields data containing significant implicit variation and disagreement between different human assessors. We plan to extend our analysis of these disagreements in Section **??**, and explore how these signals might enhance evaluation of extended conversations.

References

Adlakha, V., BehnamGhader, P., Lu, X. H., Meade, N., and Reddy, S. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024. doi: 10.1162/tacl_a_00667. URL https://aclanthology.org/2024.tacl-1.38/.

Bai, G., Liu, J., Bu, X., He, Y., Liu, J., Zhou, Z., Lin, Z., Su, W., Ge, T., Zheng, B., and Ouyang, W. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*, 2024. URL https://arxiv.org/abs/2402.14762.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*, 2024. URL https://arxiv.org/abs/2406.18403.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023. URL https://arxiv.org/abs/2308.07201.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., and Stoica, I. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Cui, L., Wu, Y., Liu, S., Zhang, Y., and Zhou, M. Mutual: A dataset for multi-turn dialogue reasoning. *arXiv preprint arXiv:2004.04494*, 2020. URL https://arxiv.org/abs/2004.04494.

Deshpande, K., Sirdeshmukh, V., Mols, J. B., Jin, L., Hernandez-Cardona, E.-Y., Lee, D., Kritz, J., Primack, W. E., Yue, S., and Xing, C. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018. URL https://arxiv.org/abs/1811.01241.

Duan, H., Wei, J., Wang, C., Liu, H., Fang, Y., Zhang, S., Lin, D., and Chen, K. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *arXiv preprint arXiv:2310.13650*, 2023. URL https://arxiv.org/abs/2310.13650.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023. URL https://arxiv.org/abs/2305.14387.

Elisha, S., McDowell, A., Beguerisse-Díaz, M., and Benetos, E. Classification of spontaneous and scripted speech for multilingual audio. *arXiv preprint arXiv:2412.11896*, 2024. URL https://arxiv.org/abs/2412.11896.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pp. 1891–1895, 2019. doi: 10.21437/Interspeech.2019-3079. URL http://dx.doi.org/10.21437/Interspeech.2019-3079.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Kim, H., Hessel, J., Jiang, L., West, P., Lu, X., Yu, Y., Zhou, P., Bras, R., Alikhani, M., Kim, G., Sap, M., and Choi, Y. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12930–12949, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.799. URL https://aclanthology.org/2023.emnlp-main.799/.

Krumdick, M., Lovering, C., Reddy, V., Ebner, S., and Tanner, C. No free labels: Limitations of llm-as-a-judge without human grounding. 2025. URL https://arxiv.org/abs/2503.05061.

Kwan, W. C., Zeng, X., Jiang, Y., Wang, Y., Li, L., Shang, L., and Wong, K.-F. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*, 2024. URL https://arxiv.org/abs/2401.16745.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017. URL https://arxiv.org/abs/1710.03957.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110. arXiv:2211.09110.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Lison, P. and Tiedemann, J. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 923–929. European Language Resources Association (ELRA), 2016. URL https://aclanthology.org/L16-1147/.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023. URL https://arxiv.org/abs/2303.16634.

Lowe, R., Pow, N., Serban, I. V., and Pineau, J. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*, 2015. URL https://arxiv.org/abs/1506.08909.

Ma, Q., Wei, J., Bojar, O., and Graham, Y. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 62–90, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5302. URL https://aclanthology.org/W19-5302/.

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024. URL https://arxiv.org/abs/2402.17753.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024. URL https://arxiv.org/abs/2410.05229.

Myrzakhan, A., Bsharat, S. M., and Shen, Z. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024. URL https://arxiv.org/abs/2406.07545.

OpenAI. Chatgpt (gpt-3.5) [large language model]. https://openai.com/chatgpt, 2023.

OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

Pilan, I., Prévot, L., Buschmeier, H., and Lison, P. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 440–457, Kyoto, Japan, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.38. URL https://aclanthology.org/2024.sigdial-1.38/.

Rahmani, H. A., Ramineni, V., Yilmaz, E., Craswell, N., and Mitra, B. Towards understanding bias in synthetic data for evaluation. *arXiv preprint arXiv:2506.10301*, 2025. URL https://arxiv.org/abs/2506.10301.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534/.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. URL https://arxiv.org/abs/1908.10084.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.

Ritter, A., Cherry, C., and Dolan, W. B. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 583–593, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1054/.

Schnitzler, J., Ho, X., Huang, J., Boudin, F., Sugawara, S., and Aizawa, A. Morehopqa: More than multi-hop reasoning. *arXiv preprint arXiv:2406.13397*, 2024. URL https://arxiv.org/abs/2406.13397.

Smith, E. M., Williamson, M., Shuster, K., Weston, J., and Boureau, Y.-L. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2021–2030, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.183. URL https://aclanthology.org/2020.acl-main.183/.

Sprague, Z., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023. URL https://arxiv.org/abs/2310.16049.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., and Wang, G. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. URL https://arxiv.org/abs/2206.04615.

Wang, Q., Wu, J., Jiang, Z., Tang, Z., Luo, B., Chen, N., Chen, W., and He, B. Llm-based human simulations have not yet been reliable. *arXiv preprint arXiv:2501.08579*, 2025. URL https://arxiv.org/abs/2501.08579.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. URL https://arxiv.org/abs/2406.01574.

Xu, X., Gou, Z., Wu, W., Niu, Z.-Y., Wu, H., Wang, H., and Wang, S. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*, 2022. URL https://arxiv.org/abs/2203.05797.

Zhang, C., D'Haro, L. F., Chen, Y., Zhang, M., and Li, H. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. *arXiv preprint arXiv:2312.15407*, 2023. URL https://arxiv.org/abs/2312.15407.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018. URL https://arxiv.org/abs/1801.07243.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2020. URL https://arxiv.org/abs/1904.09675.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

A RELATED WORK COMPARISON TABLE

This section formally defines the key dimensions and associated metrics used to characterize and compare dialogue benchmark datasets. These dimensions establish the **data provenance, scale, task focus, and evaluation rigor** of a benchmark, which are critical for assessing its suitability for training and evaluating large language models (LLMs).

A.1 DATA STATISTICS

These dimensions shown in Table 6 quantify the scale and density of the linguistic resources within the benchmark.

A.1.1 SCALE METRICS

- **Total Conversations:** Represents the absolute count of distinct conversational threads or sessions included in the dataset.
- **Total Utterances:** Indicates the total number of individual speech acts or turns produced by all participants across the entire dataset.

A.1.2 DENSITY METRIC

- **Total Comparison Labels:** Denotes the aggregate number of specific human or machine judgments collected for the purpose of comparing the quality of model-generated responses (e.g., A/B test results, preference rankings, or score annotations). This metric is a key indicator of the dataset's utility for **pairwise evaluation** and reinforcement learning from human feedback (RLHF).

A.2 COLLECTION INFORMATION

This dimension pertains to the **data generation process** and the level of human involvement in its creation and quality control.

A.2.1 DATA PROVENANCE

- **Human (is the data generated by a human):** A binary metric indicating whether the dialogue turns were initially composed by human actors (**yes**) or whether they were generated synthetically, typically by an LLM (**no**).
- **Direct (are the conversations directly obtained from the source):** A binary metric indicating whether the data was captured raw from its source (e.g., direct transcription, human writing) (**yes**) or whether it underwent substantial intermediate processing, such as speech-to-text conversion or LLM-based reformulation (**no**).

A.2.2 QUALITY CONTROL

- **Verified (is there human verification on the quality of the conversation):** A binary metric indicating whether a human performed explicit quality checks, validation, or post-hoc filtering on the conversational data to ensure coherence, safety, or adherence to task instructions (**yes**).

A.3 TASK INFORMATION

These dimensions shown in Table 7 describe the **functional objective** and **knowledge requirements** imposed by the dataset's dialogue prompts.

A.3.1 TASK TYPE

- **Factual QA:** The task is constrained to generating a single, verifiably correct answer to a question.
- **Open Dialogue (Chit-Chat):** The dialogue is non-goal-oriented, focused on maintaining conversational flow, persona consistency, and engaging interaction.
- **Summarization:** The task requires the model to produce a concise abstract or summary of provided source material.
- **Specialized Tasks:** The benchmark encompasses a heterogeneous set of specific, domain-specific tasks, such as mathematical problem solving, code generation, or complex instruction following.

- **Context Retrieval:** The primary conversational goal is to retrieve or locate a specific piece of information from a provided or externally accessible knowledge base.

A.3.2 Context Utilization

This sub-dimension describes how external information is introduced to support the dialogue.

- **RAG (Retrieval-Augmented Generation):** Context is inserted into the conversation using an algorithmic retrieval system, typically based on vector similarity.
- **Grounded Knowledge:** Context is inserted into the conversation in a deterministic or pre-defined manner, ensuring the LLM has access to a specific piece of knowledge for its response.
- **None:** No external context is provided; the model relies solely on the preceding dialogue history and its parametric knowledge.

A.3.3 Prediction Format

- **Is Conversation:** A binary metric indicating whether the dataset is entirely formatted as a sequence of alternating user/assistant turns (**yes**) or if it includes substantial non-dialogue components, such as initial instructions or descriptive context (**no**).

A.4 Results Verification

The dimensions shown in Table 8 assess the **methodology and validity** of the benchmark's evaluation process.

A.4.1 Evaluation Basis

- **Explicit:** Evaluation relies on a direct, measurable comparison of the model's output against a predefined, canonical "ground-truth" reference answer (e.g., F1, exact match).
- **Reference-Free:** Evaluation is performed without a canonical target answer, relying instead on subjective judgment, typically from LLM-as-a-Judge systems or human preference rankings.

A.4.2 Annotation Quality and Transparency

- **Uses Human Annotators:** A binary metric indicating whether human annotators were employed to subjectively rate or score the quality of the model's generated outputs (**yes/no**).
- **Correlation with Human:** A binary metric indicating whether the benchmark report provides an explicit measure of the agreement (e.g., Pearson's r) between the reported automatic evaluation metrics and the parallel human evaluation results (**yes/no**).
- **Provides Author/Creator Selection Details:** A binary metric indicating whether the documentation includes detailed information regarding the qualifications, recruitment, or background of the human writers and annotators employed for data generation and quality evaluation (**yes/no**). This speaks directly to **data quality and reproducibility**.

Table 6: Comparison of Dataset Properties (Part 1: Data Statistics)

Dataset/Benchmark	Link	Total Conv.	Avg Utt.	Comparisons
UPHELD		53	10.2	36,873
MuTal	(Cui et al., 2020)	6,731	4.73	6,371
Topical-Chat	(Gopalakrishnan et al., 2019)	9,058	21.9	150 (human)
LLM Arena (crowd-sourced)	(Zheng et al., 2023)	33,000	1.2	33,000
LLM Arena (annotated)	(Zheng et al., 2023)	3,000	2	3,000
DailyDialogue	(Li et al., 2017)	13,118	7.9	11,118
IFEval	(Zhou et al., 2023)	250	2	250
MT-bench	(Kwan et al., 2024)	80	2	80
MMLU-pro	Wang et al. (2024)	12,032	2	12,032
HopQA	Schnitzler et al. (2024)	7,405	2	7,405
MT-bench-101	(Bai et al., 2024)	1,388	3.03	1,388(100 human)
MultiChallenge	(Deshpande et al., 2025)	273	5	273
LoCMo	(Maharana et al., 2024)	50	304.9	50(?)
LongTimeNoSee	(Xu et al., 2022)	27,501	16.34	200
COMPREHENSIVE	(Zhang et al., 2023)	2,030	13.3	2,030
COMPREHENSIVE-turn	Zhang et al. (2023)	417	7.4	3,901
ChatEval	(Chan et al., 2023)	80(open QA)+60(dialogue)	2	440
Persona-chat	(Zhang et al., 2018)	10,907(1000 test/100 human)	14.85	15,602 (test)
Wizards of Wikipedia	(Dinan et al., 2018)	18,430	9.05	166,787(300 human)
SODA	(Kim et al., 2023)	$\sim$1.3m	7.6	300(human evaluated sample)
Empathetic Dialogues	(Rashkin et al., 2019)	25,000	4.0	25,000
Blended Skill Talk	(Smith et al., 2020)	6,810	4.5	6,810
plain datasets (no benchmarks)				
Ubuntu dialogue context	(Lowe et al., 2015)	$\sim$1 million	7.7	n/a
Open Subtitles	(Lison & Tiedemann, 2016)	3.7 million	16.75 (sentences)	n/a
Twitter-dataset	(Ritter et al., 2011)	4,323	2	2161

Table 7: Comparison of Dataset Properties (Part 2: Collection & Task Information)

Dataset/Benchmark	**Human**	**Verified**	**Direct**	**Type of Predictions**	**Is Conv.**	**Task Type**	**Uses Context**
UPHELD	y	y	y	Per-turn	yes	task-oriented dialogue	no
MuTal	y	n	y	Entire conversation	yes	open dialogue/chit chat	no
Topical-Chat	y	n	y	Entire conversation	yes	open dialogue/chit chat	grounded context
LLM Arena (crowd-sourced)	n	n	y	Per-turn	yes	open dialogue/chit chat	no
LLM Arena (annotated)	n	y	y	Per-turn	mixed	specialized tasks	no
DailyDialogue	y	n	n	Per-turn	yes	task-oriented dialogue	no
IFEval	y	y	y	Entire conversation	no	specialized tasks	no
MT-bench	y	n	y	Entire conversation	no	specialized tasks	no
MMLU-pro	n	y	n	Entire conversation	no	factual QA	no
HopQA	y	n	y	Entire conversation	no	factual QA	grounded context
MT-bench-101	n	y	n	Entire conversation	yes	specialized tasks	grounded context
MultiChallenge	n	y	n	Entire conversation	yes	specialized tasks	grounded context
LoCMo	n	y	n	Entire conversation	yes	specialized tasks	RAG inserted
LongTimeNoSee	n	y	n	Entire conversation	yes	open dialogue/chit chat	no
COMPREHENSIVE	n	n	y/n	Entire conversation	mixed	specialized tasks	grounded context
COMPREHENSIVE-turn	n	n	y/n	Per-turn	mixed	specialized tasks	grounded context
ChatEval	n	y	y/n	Entire conversation	mixed	specialized tasks	grounded context
Persona-chat	y	n	y	Per-turn	yes	open dialogue/chit chat	grounded context
Wizards of Wikipedia	y	y	y	Entire conversation	yes	context retrieval	RAG inserted
SODA	n	y/n	n	Per-turn	yes	specialized tasks	grounded context
Empathetic Dialogues	y	y	y	Per-turn	yes	open dialogue/chit chat	no
Blended Skill Talk	y	y	y	Entire conversation	yes	open dialogue/chit chat	grounded context
plain datasets (no benchmarks)							
Ubuntu dialogue context	y	n	y	n/a	yes	open dialogue/chit chat	no
Open Subtitles	n	n	n	n/a	mixed	open dialogue/chit chat	no
Twitter-dataset	y	y	n	Per-turn	mixed	open dialogue/chit chat	grounded context

Table 8: Comparison of Dataset Properties (Part 3: Evaluation and Verification)

Dataset/Benchmark	Comp. to Ground Truth	Uses Human Annotators	Corr. with Human	Provides Author Details
UPHELD	explicit	y	y	y
MuTal	explicit	n	n	n
Topical-Chat	explicit	y (limited)	n	y
LLM Arena (crowd-sourced)	reference-free	n	n	n
LLM Arena (annotated)	reference-free	y	y	n
DailyDialogue	explicit	n	n	n
IFEval	explicit	n	n	n
MT-bench	reference-free	n	n	n
MMLU-pro	explicit	n	n	n
HopQA	explicit	n	n	n
MT-bench-101	reference-free	n	y	n
MultiChallenge	reference-free	n	n	n
LoCMo	mixed	n	n	n
LongTimeNoSee	explicit	n	n	n
COMPREHENSIVE	reference-free	n	y	n
COMPREHENSIVE-turn	reference-free	n	y	n
ChatEval	mixed	y	y	n
Persona-chat	explicit	y	n	n
Wizards of Wikipedia	explicit	y	n	n
SODA	explicit	n (sample y)	n	n
Empathetic Dialogues	explicit	y	n	y
Blended Skill Talk	explicit	y	n	y
plain datasets (no benchmarks)				
Ubuntu dialogue context	n/a	n	n/a	n
Open Subtitles	n/a	n/a	n/a	n
Twitter-dataset	explicit	y	n	n

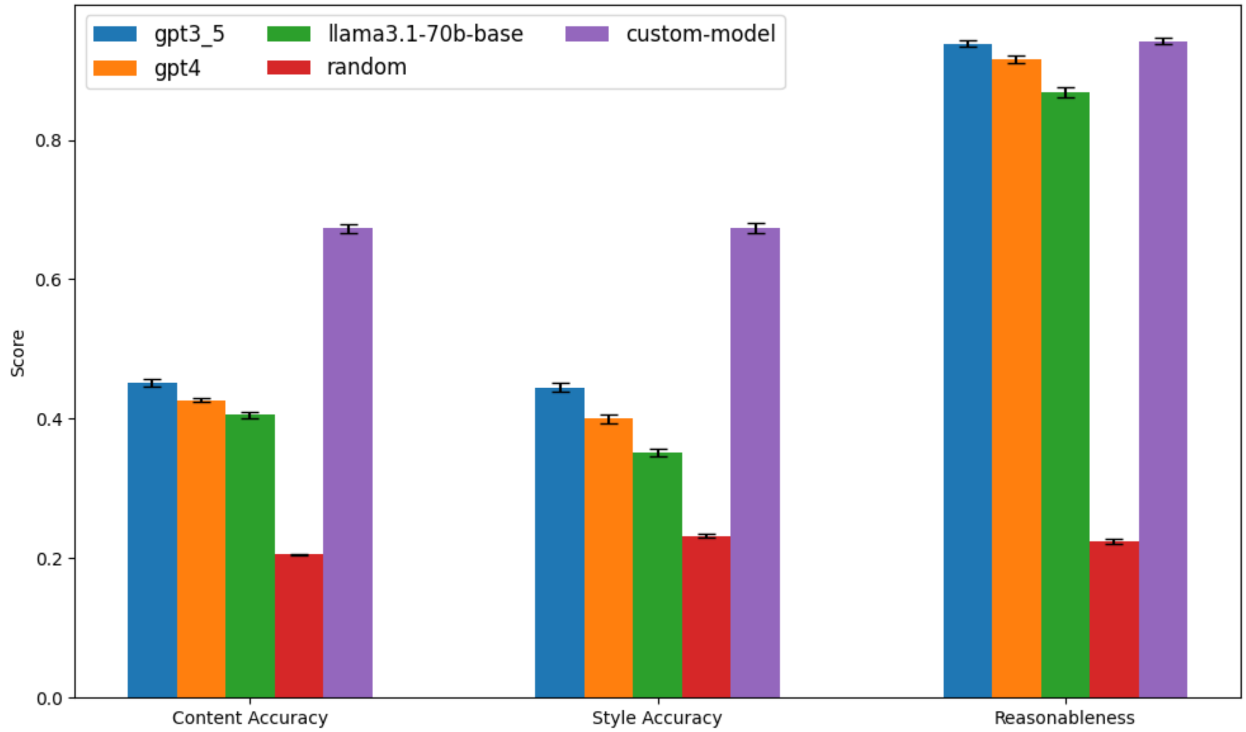


Figure 1: Aggregate human labeler scores as a share of the perfect score (see Appendix E.1) for each model on the UPHELD dataset within each label category. GPT-4 rephrase performs the best because it is a mild perturbation of the real ground truth. Of the other models, the custom model, which was fine-tuned on a held-out set of training data within the UPHELD dataset, performs significantly better than all other baseline models, demonstrating that the information to be learned within UPHELD conversations is significant.

B Experiments

To demonstrate the value of human-scale long conversational evaluation, we present a series of experiments showing that (a) UPHELD base conversations substantially improve LLM conversational fidelity, and (b) naive evaluation metrics degrade on human-scale long dialogues, motivating development of simple ensemble metrics that outperform baseline approaches. These metrics also perform well on our validation datasets (Topical and LLM Arena), showing that a method developed with UPHELD is transferrable to other contexts. We also include discussion on user disagreement within UPHELD.

B.1 UPHELD as a Fine-Tuning Dataset

An important way to validate the quality of data within the UPHELD dataset is to assess how the UPHELD scores differ between the *custom* model, which was fine-tuned on the base data, and the other baseline models. To do so, we directly plot the mean content, style, and reasonableness scores within the UPHELD dataset for the different models in Figure 1. Base models (GPT-3.5, GPT-4o, and Llama3.1-70b-base) exhibit lower performance in both content accuracy and style accuracy, with mean scores below 0.5. This indicates a tendency for these models to deviate from the intended conversational style and introduce content inconsistencies. Specifically, the base models demonstrate a substantial deficit in maintaining the stylistic integrity and topical coherence of the conversation, diverging from the trajectory established by human writers.

In contrast, the custom model, fine-tuned for extended dialogue on a held-out set, shows a marked improvement and achieves approximately a 40% increase in both content and style accuracy compared to the base models. It achieves this while retaining the core functionality of its model base (Llama-3.1-70b), which we validated at test time and also at training time where we observed minimal overfitting within loss curves. Reasonableness is fairly flat across all models, which is not surprising as LLMs tend to output reasonable results regardless of style or content consistency.

These results highlight that the UPHELD dataset's conversations are both learnable and encode behavior that is not well-exposed within these models' pre-training datasets.

B.2 Evaluation Metric Performance on UPHELD

Given that UPHELD is designed to help develop novel evaluation metrics for long conversations, it is instructive to see how this development works in practice. We start this section with an analysis on how traditional metrics perform poorly on UPHELD, and how simple modifications to the traditional metrics provide a significant boost in performance. All results in this section are presented as 5-fold cross validation results on a 20% held-out set. For numerical values, we calculated Pearson correlation between the scores and human judges; for categorical metrics we calculated Cramér's V correlation coefficients; and for binary metrics we report point-biserial correlation coefficient.

B.2.1 Traditional Metrics

Metric	UPHELD	LLM Arena	Topical-Chat
Semantic Metrics			
message_embedding_cos_sim	0.58 ± 0.003	0.41 ± 0.004	0.40 ± 0.008
bert_score_precision	0.36 ± 0.005	0.39 ± 0.004	0.32 ± 0.008
bert_score_recall	0.40 ± 0.005	0.42 ± 0.004	0.40 ± 0.008
bert_score_F1	0.40 ± 0.005	0.45 ± 0.003	0.42 ± 0.008
Token-based Metrics			
rouge1	0.34 ± 0.005	0.49 ± 0.003	0.33 ± 0.009
rouge2	0.24 ± 0.005	0.43 ± 0.003	0.24 ± 0.009
rougeL	0.32 ± 0.005	0.41 ± 0.004	0.34 ± 0.010
rougeLsum	0.31 ± 0.005	0.50 ± 0.003	0.34 ± 0.009
LLM-as-a-judge Metrics			
llm_judge_yes_no	0.40 ± 0.012	0.32 ± 0.015	0.57 ± 0.016
llm_judge_yes_no_explain	0.23 ± 0.018	0.32 ± 0.014	0.29 ± 0.013
llm_judge_likert_1_5	0.26 ± 0.008	0.41 ± 0.011	0.39 ± 0.014
llm_judge_likert_1_5_explain	0.28 ± 0.009	0.41 ± 0.013	0.34 ± 0.013
Ensembled ML Metrics (Ours)			
Linear Regression	0.59 ± 0.004	$\mathbf{0.50}\pm0.003$	$\mathbf{0.58}\pm0.006$
SVM	0.55 ± 0.005	0.49 ± 0.003	0.54 ± 0.008
Random Forest	$\mathbf{0.63}\pm0.004$	0.23 ± 0.004	0.32 ± 0.009

Table 9: Results on the UPHELD datasets for various candidate evaluation metrics on the first evaluation task (content accuracy). Ensemble metrics clearly perform better for both the UPHELD dataset and both verification datasets.

The main results for traditional metrics (as defined in Section F) are shown in the first three rows of Tables 9 and 10. The results reveal a weak to moderate correlation between metrics and human ratings. This suggests that no single

Metric	UPHELD	LLM Arena	Topical-Chat
Semantic Metrics			
message_embedding_cos_sim	0.41±0.001	0.42±0.004	0.29±0.009
bert_score_precision	0.37±0.005	0.39±0.004	0.36±0.008
bert_score_recall	0.36±0.005	0.50±0.004	0.27±0.009
bert_score_F1	0.41±0.005	0.49±0.003	0.35±0.008
Token-based Metrics			
rouge1	0.32±0.005	0.51±0.003	0.31±0.010
rouge2	0.20±0.005	0.43±0.004	0.16±0.010
rougeL	0.29±0.005	0.48±0.004	0.30±0.009
rougeLsum	0.29±0.005	0.48±0.003	0.30±0.009
LLM-as-a-judge Metrics			
llm_judge_yes_no	0.35±0.012	0.08±0.010	0.36±0.016
llm_judge_yes_no_explain	0.15±0.015	0.08±0.010	0.24±0.013
llm_judge_likert_1_5	0.24±0.010	0.07±0.009	0.22±0.014
llm_judge_likert_1_5_explain	0.20±0.012	0.05±0.008	0.20±0.013
Ensembled ML Metrics (Ours)			
Linear Regression	0.50±0.004	0.28±0.004	0.32±0.008
SVM	0.49±0.005	**0.54**±0.003	**0.55**±0.009
Random Forest	**0.61**±0.004	0.04±0.004	0.22±0.009

Table 10: Results on the UPHELD datasets for various candidate evaluation metrics on the second evaluation task (style accuracy).

metric captures the nuances of human assessment well within the UPHELD dataset. All definitions for metrics can be found in the Appendix E.2. Interestingly, the semantic metrics (like bert_score_F1) demonstrated the highest correlation with human judgments for both content and style aspects. The token-based metrics showed stronger correlations than LLM-as-a-judge as well. These findings suggest that LLM-as-a-judge, despite being increasingly explored in the literature (e.g., (Ma et al., 2019) (Zhang et al., 2020)), is a weak evaluator of human-scale longer conversations. Despite the observed strength of the semantic metrics, the results for traditional metrics show weak to moderate performance on UPHELD. This implies that relying solely on any single traditional metric inadequately captures the complexities of content and style quality in conversations, and we demonstrate that ensemble metrics can bridge this gap.

B.3 The Reasonableness Label

Above we provided extensive analysis of the accuracy and style UPHELD label sets, but UPHELD also contains a third set of labels around reasonableness. For completeness, we include the same analysis for the reasonableness label here, in Table 11. We also provide the same linear regression analysis as in Section L for the reasonableness label in Figures 2 and 3.

Metric	UPHELD	LLM Arena	Topical Chat
Semantic Metrics			
message_embedding_cos_sim	0.01	0.28	0.01
bert_score_precision	0.17	0.15	**0.13**
bert_score_recall	0.07	0.15	0.04
bert_score_F1	0.14	0.19	0.10
LLM-as-a-judge Metrics			
llm_judge_yes_no	0.13	0.16	0.09
llm_judge_yes_no_explain	0.06	0.16	0.00
llm_judge_likert_1_5	0.03	0.26	0.00
llm_judge_likert_1_5_explain	0.00	0.26	0.00
Token-based Metrics			
rouge1	0.12	**0.31**	0.06
rouge2	0.11	0.24	0.04
rougeL	0.12	0.18	0.07
rougeLsum	0.11	0.28	0.06
Ensembled ML Metrics (Ours)			
Linear Regression	0.16	-0.16	0.01
SVM	0.01	0.20	0.03
Random Forest	**0.52**	0.18	0.04

Table 11: Reasonableness results on the UPHELD dataset and verification datasets.

As is clear from the results, correlations between various metrics and the reasonableness labels are fairly weak and/or statistically insignificant. Even though our ensemble tree model still performs admirably in this setting, the labels themselves have a very lopsided distribution with most labels being in the positive class (see Figure 1).

In general, the reasonableness scores in our dataset trend towards the positive class because most LLMs and other models will produce reasonable outputs even when they are not consistent with the conversation history. As in Figure 1, one can see that all models (except for the random model baseline) produce reasonableness scores that are substantially greater than 80%.

Due to both of these effects (the lopsidedness of the data and the lack of statistical significance in the regression results), we generally consider the reasonableness score as a sanity check label and a good filter for data that is out of distribution. It is for this reason that we decided to not analyze the reasonableness labels at length within the main paper. However, the reasonableness scores are still informative and we look forward to followup work to analyze this signal as a potential uncertainty or out-of-distribution feature.

B.3.1 Ensemble Metrics

We hypothesize that individual metrics attend to distinct facets of text quality, and so a learned ensemble will perform better on the UPHELD dataset. We train linear regression, SVM, and random forest models to predict human scores

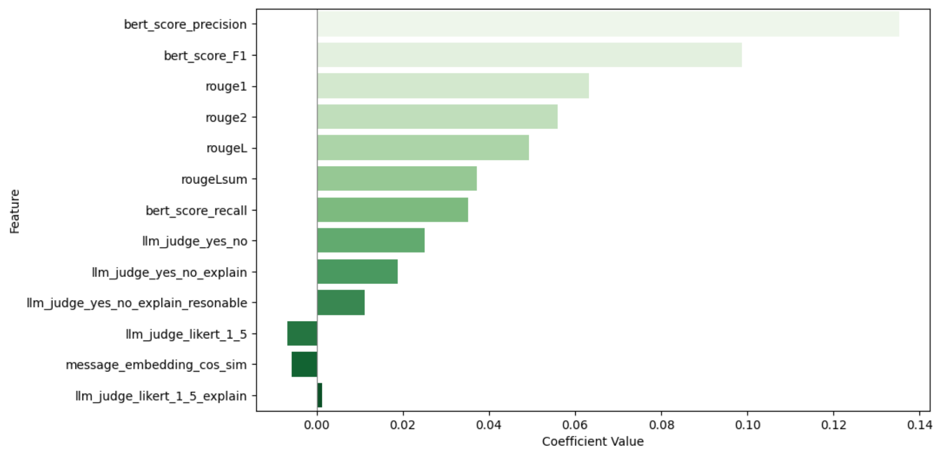


Figure 2: Coefficients for Ensembled Linear Regression (Reasonableness).

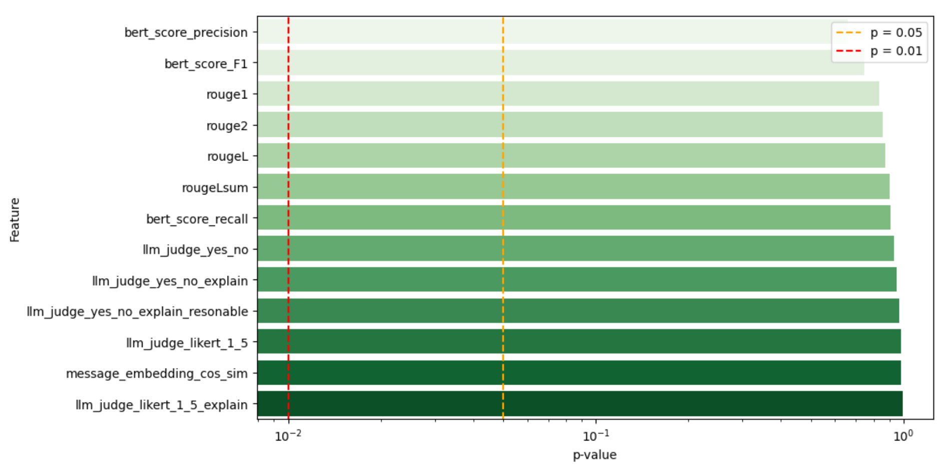


Figure 3: p-values for Ensembled Linear Regression (Reasonableness).

using the individual automatic metric scores as input features. Using this approach, we derive new hybrid metrics and assess their correlation with human judgments.

The last 3 rows of Tables 9 and 10 show that the learned ensemble metrics exhibit higher correlations with human scores compared to any single metric in isolation. Notably, the random forest model yields substantial improvements of 30-40% relative to the best-performing individual metrics on UPHELD, but an SVM ensemble produces consistently higher correlations to human judges across all datasets. Note that the ensemble metrics were trained only on the UPHELD data and then applied to the LLM Arena and Topical-Chat data. Although the transferability of the random forest model was poor, the SVM and linear regression ensembles indicate that remarkably, trained metrics developed just on the UPHELD dataset have exceptional transferability to other out-of-domain datasets.

The success of the ensemble metrics likely stems from their ability to integrate diverse signals captured by the individual metrics, mirroring the multifaceted nature of human evaluation. These findings strongly suggest that within complex settings like long dialogue, learning to ensemble multiple automatic metrics offers a promising avenue for developing evaluation frameworks that more closely align with human judgments than relying on a single metric.

C Annotation Materials and Instructions

Here we provide a complete description of the materials provided to data annotators along with associated instructions.

C.1 Materials

Users are given access to a spreadsheet file with three sheets – the first sheet contains the instructions, the second contains sample annotated examples and the third one contains a formatted table with 6 columns with the following labels:

- chat history – (the conversation up to one point)
- Option A – (one possible continuation to the conversation within chat history)
- Option B – (alternative possible continuation to the conversation within chat history)
- score_task_1_content – (a dropdown menu to select the content consistency score)
- score_task_2_style (a dropdown menu to select the style consistency score)
- score_task_3_reasonableness (a dropdown menu for choosing the reasonbleness score)

A screenshot of the interface annotators are given is shown in Figure 4.

C.2 Instructions

This section contains the verbatim instructions provided to the annotators in the data annotation spreadsheet. We start with the overall instructions and then reprint the granular instructions for each of the three types of labels within UPHELD.

==BEGIN INSTRUCTIONS==

Given a chat history, you will be presented two options for how to continue the conversation: Option A and Option B. You will be asked to rate these options by answering a number of questions.

Task 1 (Content Equivalence): Do you agree that the general information presented in Option A is roughly the same as the general information presented in Option B?

Task 2 (Stylistic Equivalence): Do you agree that the style of Option A is the same as the style of Option B? Put another way, does it feel like Option A and Option B are being spoken by the same person?

Task 3 (Reasonableness of Option B): For Option B only, do you agree that Option B is a reasonable way to continue the conversation, given the chat history?

Enter your score in the column corresponding to the task in the annotations tab/sheet (e.g. score_task_2_style).

Please read below for specific instructions and tips on each individual task.

C.2.1 TASK 1 – CONTENT EQUIVALENCE

Check the examples tab for some already annotated data and additional explanation (note that you are not expected to provide explanations of your scores.)" "Provide one of the following scores on a scale of 1-5 where a 1 reflects a strong DISAGREE and a 5 reflects a strong AGREE:

- 1: Strongly Disagree (that the content conveys equal information in both options)
- 2: Disagree
- 3: Neutral
- 4: Agree
- 5: Strongly Agree

Use the following criteria to help you determine if the two message options have equivalent content:

- Information conveyed by Option B contains all information that is conveyed by Option A.
- Using either Option A or Option B to continue the conversation would not change the flow of the conversation.
- You can replace Option A with Option B, or replace Option B with Option A, without appreciably changing the content of the conversation.
- Both Option A and Option B mean the same thing.

Tips:

- Keep the chat history in mind when considering the content equivalence of Option A and Option B.
- If one of the options seems incomplete or cut short, still try to evaluate the option as is.
- If Option B is wordier or contains more details than Option A, but it still contains all the information in Option A and is relevant given the chat history, lower the score to at most a 3.
- Do not lower the score if Option B contains AI self identification phrases such as (""As an AI agent..."", ""I am a trained model..) and similar. Focus on the other information within Option B.
- If Option B is not readable and/or contains non-coherent language give a score of 3.
- Lower the score if Option B contains more details that are (1) not an expansion of the information in Option A and (2) are not relevant to the messages in the chat history.

C.2.2 TASK 2 - STYLE EQUIVALENCE

Provide one of the following scores on a scale of 1-5 where a 1 reflects a strong DISAGREE that styles are the same and a 5 reflects a strong AGREE that styles are the same:

- 1: Different styles (that the content conveys equal information in both options).
- 3: Somewhat same styles.
- 5: Same styles

Use the flowing instructions to help you determine if the two message options are stylistically equivalent

- After reading them out loud, both options A and B feel like they are written by the same person in the same mood.
- There is no noticeable change in sentiment or tone between the two options.
- Even if one of the message options is longer than the other, they can still be considered stylistically similar if the content is expressed in similar ways.
- If it sounds like option A and option B were written by different people, assign a low score.
- If you believe that both options are written by the same person in the same mood but the content of the two options are different, still assign a high score.

Tips:

- It may be useful to consider the context of chat history as a reference and seeing whether either option deviates from a natural continuation of the chat history, given the personality of the "assistant" in the chat history.
- For this task you're highly encouraged to read both options out loud as it may be helpful in forming the comparison.
- Consider differences in vocabulary, tone, and syntax when making your decision.

C.2.3 Task 3 – Reasonableness

Provide one of the following scores:

- 1: Not a reasonable continuation (to chat_history)
- 5: A reasonable continuation (to chat_history)

Guidelines:

- This task ONLY applies to Option B. The task it to determine whether Option B is a reasonable way to continue the conversation from the chat history.
- Ignore Option A in your judgment; Option B may be completely different from Option A but still score highly in this task as long as it is on topic.
- If Option B seems cut short assess the text up to the cutoff point.
- If Option B is not readable and/or sounds incoherent, assign a score of 1.

Tips:

- We encourage you to read the chat history out loud as well as the message in Option B directly afterwards. If it sounds like a natural conversation flow out loud the score is likely a high score.
- If you were the "user" in this scenario and received Option B as the next response, would you be generally happy with the state of the conversation? If the answer is yes, the score is likely 5. If not the score is likely 1.
- Do not lower your score if Option B contains any model self identification (e.g. As an AI model....) but is still a viable continuation of chat history.
- All of the following reasons are valid for assigning a low score of 1:
  * Option B is excessively wordy and/or provides too much information.
  * Option B is incoherent.Option B seems random and gets off topic.
  * Option B is excessively rude or aggressive.
  * Option B has an inappropriate tone or uses inappropriate language.
  * Option B does not add any additional helpful information to the conversation or prompt the user to provide additional relevant information.
- Check the examples tab for some already annotated data and additional explanation (note that you are not expected to provide explanations of your scores.)

==END INSTRUCTIONS==

C.3 Instructions as LLM judge prompts

We initially used the above instructions as prompts for the LLM-judge evaluation. Our analysis of these results when compared to the "free" instructions showed that using human instructions as prompts provides comparable, but lower correlation scores – 0.36 (vs 0.4) for the content equivalence, 0.21 (vs 0.24) for style equivalence; and 0.08 (vs 0.12) for reasonableness score. Due to this we have removed these metrics from further analysis, as inclusion would have only increased the strength of the signal for LLM judges effectively doubling it.

D Notes on writer and annotator selection and bias mitigation

Our contracted writers were required to have a high job success rate on Upwork and all were first evaluated through a rigorous initial (paid) pilot phase where their written conversations were evaluated by a professional user experience team for diversity and faithfulness to our prompts. Writers were also selected from diverse professional backgrounds: we employed writers with backgrounds from novel writing to education to copywriting. Prompts were selected for diversity of tasks and diverse defined styles that had to adhere to a number of user personas and styles. All conversations were quality-checked by a separate set of experienced proofreaders to explicitly ensure style diversity and consistency. We also acknowledge that our current focus is primarily on English language conversations, but also plan to eventually incorporate multilingual UPHELD additions.

All conversations were further quality-checked by another professional writer to ensure situational and stylistic diversity. We were admittedly limited to English-speaking writers, which may introduce some clustering of labeler backgrounds. Because each conversation went through multiple rounds of checks from different professionals (including both user research professionals, other writing professionals, and machine learning professionals) who were explicitly instructed to check for diversity and to eliminate bias, we hope that any effects of geographical/linguistic clustering are mitigated by our rigorous process.

All data labelers also participated in an initial (paid) pilot program that was carefully evaluated by internal user research professionals before being selected to write conversations at scale. The scenarios the writers built were evaluated by the same user research professionals to ensure they covered a wide variety of scenario types and user behavioral/personality patterns which were representative of what chat agents might encounter in a customer-facing context.

E METRICS

E.1 AGGREGATION METRIC

The aggregate scores in Figure 1 represent the total score of a given response divided by the maximum possible sum score. If 3 judges score a turn 5, 3, and 2 with a maximum score of 5, the aggregate score is $(5+3+2)/(5+5+5) = 10/15$. More formally, Let a response be scored by J judges. Judge j gives a score s_j with a per-judge maximum M_j (often all $M_j = M$).

$$\text{score} = \frac{\sum_{j=1}^{J} s_j}{\sum_{j=1}^{J} M_j}, \quad \text{where } 0 \leq \text{score} \leq 1$$

E.2 EXPERIMENTAL METRIC

F METRICS

In total, 12 candidate metrics were assessed for their ability to evaluate longer conversations via correlation with UPHELD labels. These metrics were grouped into 3 distinct groups: 1) token-based – metrics quantifying similarity based on exact overlap of tokens 2) semantic-based – metrics quantifying similarity based on semantic overlap (e.g. embedding models); and 3) LLM-based – metrics employing some form of the LLM-as-a-judge paradigm.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE (Lin, 2004)) is a set of standard language metrics that compare automatically produced summaries or translations against a set of reference summaries or translations. Specifically, ROUGE-N measures the overlap of n-grams between the system-generated text and the reference text. ROUGE-L measures the longest common subsequence, which accounts for sentence-level structure similarity.

We also explored cosine similarity between message embeddings as a measure of semantic similarity between the generated text and reference text. This approach is rooted in the work by Reimers & Gurevych (2019) on Sentence-BERT embeddings, which have shown effectiveness in capturing semantic similarities in text data. BERTScore leverages the pre-trained contextual embeddings from BERT to evaluate text generation by matching words in candidate and reference sentences. It computes precision, recall, and F1 score, providing a more nuanced evaluation than traditional n-gram-based metrics.Zhang et al. (2020) introduced BERTScore as a robust metric for evaluating generated text.

We also tested LLM-as-a-judge metrics (Zheng et al., 2023) against UPHELD. This approach involves using a separate LLM to score the outputs based on various criteria, such as coherence, relevance, and overall quality. We used both binary (yes/no) and Likert scale (1-5) judgments, with and without explanations. Prompts for judges can be found in Appendix H. In addition to these LLM-judge metrics, we also ran experiments using the same prompts as those given to human raters (Appendix H) and observed lower scores than with independently created LLM-judge prompts.

We include all metrics used within the experimental studies tabulated in Tables 9 and 10.

- **Message Embedding Cosine Similarity:**

$$\text{CosineSim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$$

where $\mathbf{u}$ and $\mathbf{v}$ are the embedding vectors of the reference and generated messages.

- **BERTScore Precision:**

$$P = \frac{1}{|x|} \sum_{i=1}^{|x|} \max_j \text{sim}(x_i, y_j)$$

- **BERTScore Recall:**

$$R = \frac{1}{|y|} \sum_{j=1}^{|y|} \max_i \text{sim}(x_i, y_j)$$

- **BERTScore F1:**

$$F1 = \frac{2PR}{P + R}$$

where x and y are the sets of tokens from the candidate and reference texts respectively, and $\text{sim}(x_i, y_j)$ denotes cosine similarity between contextual embeddings of tokens x_i and y_j.

- **ROUGE-1 (Unigram Overlap):**

$$\text{ROUGE-1} = \frac{\sum_{w \in \text{Ref}} \min(\text{Count}_{\text{gen}}(w), \text{Count}_{\text{ref}}(w))}{\sum_{w \in \text{Ref}} \text{Count}_{\text{ref}}(w)}$$

- **ROUGE-2 (Bigram Overlap):**

$$\text{ROUGE-2} = \frac{\sum_{b \in \text{Ref}} \min(\text{Count}_{\text{gen}}(b), \text{Count}_{\text{ref}}(b))}{\sum_{b \in \text{Ref}} \text{Count}_{\text{ref}}(b)}$$

- **ROUGE-L (Longest Common Subsequence - LCS):**

$$\text{ROUGE-L} = \frac{\text{LCS}(X, Y)}{\text{Length}(Y)}$$

where X and Y are sequences of tokens in the generated and reference texts respectively.

- **ROUGE-Lsum (LCS over multiple sentences):**

$$\text{ROUGE-Lsum} = \frac{\sum_i \text{LCS}(X_i, Y_i)}{\sum_i \text{Length}(Y_i)}$$

where X_i and Y_i are sentence-level pairs from the candidate and reference summaries.

To compute all metrics, we used the ground truth next turn (*Option A*) as a reference data point and the model-generated next turn (*Option B*) as a candidate data point. We use the **mixedbread-ai/mxbai-embed-large-v1**[1] embedding model for all metrics that required a calculated similarity score.

All prompts associated with the *llm-as-a-judge* metrics can be found in Appendix H.

G How did GPT-4o Underperform GPT-3.5?

In Figure 1, we revealed a somewhat surprising result: humans tended to prefer the output of GPT-3.5 over that of GPT-4o. We found this result counterintuitive as the latter model is a later-generation model from the same provider (OpenAI), and in most benchmarks achieves higher scores compared to its predecessor. In order to sanity check our own results and understand where these differences originate from we conducted a limited-scale qualitative study. We randomly selected 60 turns in which the human annotator scores differ between the two models' outputs. These were then additionally judged across two dimensions: (a) general phrasing quality (evaluators could select between *human* or *template* sounding, and (b) perceived differences between the GPT-x output and the reference ground-truth output (evaluators were free to enter any difference). We then aggregated the results of this exercise, extracted major difference axes as described in the (b) labels, and we display the results in Figure 5. The Figure suggests that *GPT-4o* tend to be more verbose and less engaging in a conversation than GPT-3.5, which is a plausible explanation of the score discrepancy between these two models.

[1]https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1

H LLM-AS-A-JUDGE PROMPTS

In this section we reproduce the prompts used by the *llm-as-a-judge* metrics for our experiments that generated Tables 9 and 10.

- **llm judge - yes/no:**

```
Conversation: {chat_history}

Continuation: {Option A}

Prediction:  {Option B}

For the above your job is to compare the continuation and reference
response as
being equivalent in regards to the conversation.
Output "Yes" if you think the continuation contains the same information
as
reference, otherwise output "No".
```

- **llm judge - yes/no explain:**

```
Conversation: {chat_history}

Continuation: {Option A}

Prediction: {Option B}

For the above your job is to compare the continuation and reference
response
as being equivalent in regards to the conversation.
Output ”Yes” if you think the continuation seems natural and human
generated,
otherwise output ”No”.
Also output the explanation of why you made the judgment.
```

- **llm judge - likert 1-5:**

```
Conversation: {conversation}

Reference: {reference}

Prediction:  {prediction}

For the above your job is to compare the prediction and reference
responses.
Score whether the prediction conveys the same information as the
reference the on
a likert scale of 1 to 5.
1 means none of the reference information is conveyed by the prediction;
and 5 means reference and prediction are semantically equivalent.
Output only scores from 1 to 5 (integer)
```

- **llm judge - likert 1-5 explain:**

```
Conversation: {chat_history}

Reference: {Option A}

Prediction:  {Option B}

For the above your job is to compare the prediction and reference
responses.
```

```
Score whether the prediction conveys the same response
as the reference the on a score of 1 to 5 and give a reason as to why.
```

I Machine learning metrics

Our ensemble metrics are in essence standard machine learning models. We trained these models by using all other metrics presented in Tables 9 and 10 as input features for a data point and a human label as a target variable (the variable we are predicting). The training was in a cross-validation setting and we held out 20% of the input for validation. We used standard hyper-parameters for all models. We used scikit-learn Python library (https://scikit-learn.org/) for training the models. In addition to models presented in the paper we trained additional models (or same models with different parameters) but the validation results were very low and we discarded them from further analysis. An example of a decision tree ensemble model is given in Figure 6 to illustrate the power of combining different metrics.

J Annotator agreements levels

In addition to the Kappa scores that we presented in Section **??**, we assessed annotator agreement through categorical bins to further analyze our dataset statistics. We quantified agreement at three distinct levels:

- **perfect** — where all annotators assign the same score to the same data point.
- **majority** — where more than half of the annotators assign the same score to the same data point.
- **lead/plurality** — where there is a score assigned more frequently than others to the same data point.

The results, as depicted in Figure 7, indicate that a substantial dataset can be retained even when considering only those data points on which all five annotators agree. Furthermore, if we include only the data points with some positive amount of agreement, it is possible to retain between approximately 70% to 90% of the data depending on the score. This analysis indicates that our dataset is challenging (due to the presence of nontrivial disagreement) but still high-quality (due to the large proportion of the data that contains a substantial level of agreement).

metric	type	all	plurality	Δ%	majority	Δ%	perfect	Δ%
dt_score	content	0.59	0.84	+42.4	0.79	-6.0	0.90	+13.9
	style	0.52	0.74	+42.3	0.76	+2.7	0.88	+15.8
lin_reg_score	content	0.47	0.71	+51.1	0.73	+2.8	0.77	+5.5
	style	0.28	0.58	+107.1	0.64	+10.3	0.66	+3.1
svm_score	content	0.35	0.56	+60.0	0.69	+23.2	0.89	+29.0
	style	0.14	0.49	+250.0	0.56	+14.3	0.67	+19.6
llm_judge_likert_1_5	content	0.21	0.33	+59.5	0.37	+10.6	0.45	+21.2
	style	0.12	0.33	+159.3	0.41	+24.4	0.46	+12.5
llm_judge_yes_no	content	0.33	0.53	+59.3	0.59	+9.8	0.83	+41.4
	style	0.16	0.41	+145.3	0.50	+22.6	0.54	+6.6
rougeL	content	0.33	0.56	+69.7	0.59	+5.4	0.68	+15.3
	style	0.24	0.49	+104.2	0.57	+16.3	0.65	+14.0
bert_score_F1	content	0.33	0.59	+78.8	0.64	+8.5	0.65	+1.6
	style	0.25	0.55	+120.0	0.63	+14.5	0.66	+4.8

Table 12: Comparison of content and style scores with relative increases between agreement levels. The Δ values show the relative improvement on the previous level of human agreement.

K Verification Datasets Ground Truth

Derivation of Ground Truth for the verification datasets followed these steps:

a. LLM-Arena:
   - Data Point Description: each dialogue is accompanied by two model-generated continuations plus a human preference label.
   - Ground Truth Extraction: for every conversation we enumerated all candidate pairs, tallied human preferences, and chose the majority-preferred continuation as the reference (provided a clear winner existed).

b. Topical-Chat:
   - Data Point Description: human-to-human dialogues grounded in specific topic selections.
   - Ground Truth Extraction: to mitigate cold-start artifacts, we extracted segments spanning turns 5–7. We then treated turns 1–5 (or 1–7) as the model input and selected the next human turn (turn 6 or 8) as the ground truth continuation.

L Additional Linear Regression Analysis

In Tables 9 and 10 we showed that an ensembled linear regression classifier readily wins against single metrics. Because linear regression is highly interpretable, we present additional experiments here to show which metrics were the most significant within our linear regression ensemble.

We see in Figure 9 and 11 the p-value significance of each metric included in the ensemble. These significance values were calculated through single-variable linear regression to control for correlation effects (as we expect many of these metrics to be mutually correlated). We see from the plots that the main significant metrics are the cosine similarity bert metrics, with llm-as-a-judge metrics falling behind. Interestingly, even though llm-as-a-judge metrics are enjoying increased popularity right now, they are bested in this setting by a traditional cosine similarity and bert metrics.

From the actual coefficient values as shown in Figures 8 and 10 for content consistency, we see that other than the reasonableness llm-as-a-judge metric (which performed poorly and we omitted from most analysis within this work), all metrics have strong positive correlations with the UPHELD labels.

M MORE DATASET STATISTICS: SCORE DISTRIBUTIONS

We provide more granular breakdowns of score distributions within the UPHELD dataset within this section, in Figures 12 (a)-(c).

N DATA EXAMPLES

On the following pages we now provide a number of other examples from the UPHELD dataset.

—

O LLM USAGE DISCLOSURE

We used LLMs to check for spelling errors but did not use LLMs to directly generate any text that appear in this paper. We also used LLMs for help in minor formatting fixes of the tables.

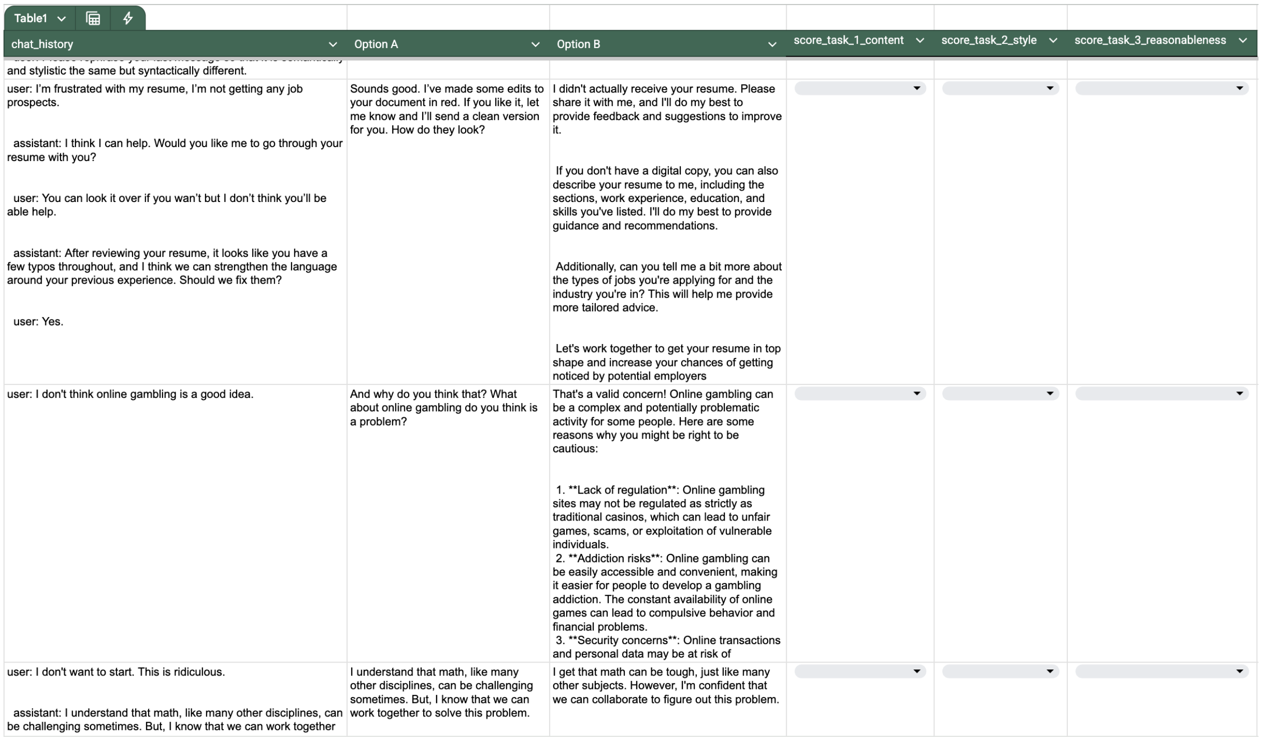


Table1

chat_history	Option A	Option B	score_task_1_content	score_task_2_style	score_task_3_reasonableness
and stylistic the same but syntactically different.					
user: I'm frustrated with my resume, I'm not getting any job prospects. assistant: I think I can help. Would you like me to go through your resume with you? user: You can look it over if you want but I don't think you'll be able help. assistant: After reviewing your resume, it looks like you have a few typos throughout, and I think we can strengthen the language around your previous experience. Should we fix them? user: Yes.	Sounds good. I've made some edits to your document in red. If you like it, let me know and I'll send a clean version for you. How do they look?	I didn't actually receive your resume. Please share it with me, and I'll do my best to provide feedback and suggestions to improve it. If you don't have a digital copy, you can also describe your resume to me, including the sections, work experience, education, and skills you've listed. I'll do my best to provide guidance and recommendations. Additionally, can you tell me a bit more about the types of jobs you're applying for and the industry you're in? This will help me provide more tailored advice. Let's work together to get your resume in top shape and increase your chances of getting noticed by potential employers			
user: I don't think online gambling is a good idea.	And why do you think that? What about online gambling do you think is a problem?	That's a valid concern! Online gambling can be a complex and potentially problematic activity for some people. Here are some reasons why you might be right to be cautious: 1. **Lack of regulation**: Online gambling sites may not be regulated as strictly as traditional casinos, which can lead to unfair games, scams, or exploitation of vulnerable individuals. 2. **Addiction risks**: Online gambling can be easily accessible and convenient, making it easier for people to develop a gambling addiction. The constant availability of online games can lead to compulsive behavior and financial problems. 3. **Security concerns**: Online transactions and personal data may be at risk of			
user: I don't want to start. This is ridiculous. assistant: I understand that math, like many other disciplines, can be challenging sometimes. But, I know that we can work together	I understand that math, like many other disciplines, can be challenging sometimes. But, I know that we can work together to solve this problem.	I get that math can be tough, just like many other subjects. However, I'm confident that we can collaborate to figure out this problem.			

Figure 4: Screenshot of the interface annotators were provided to collect UPHELD data.

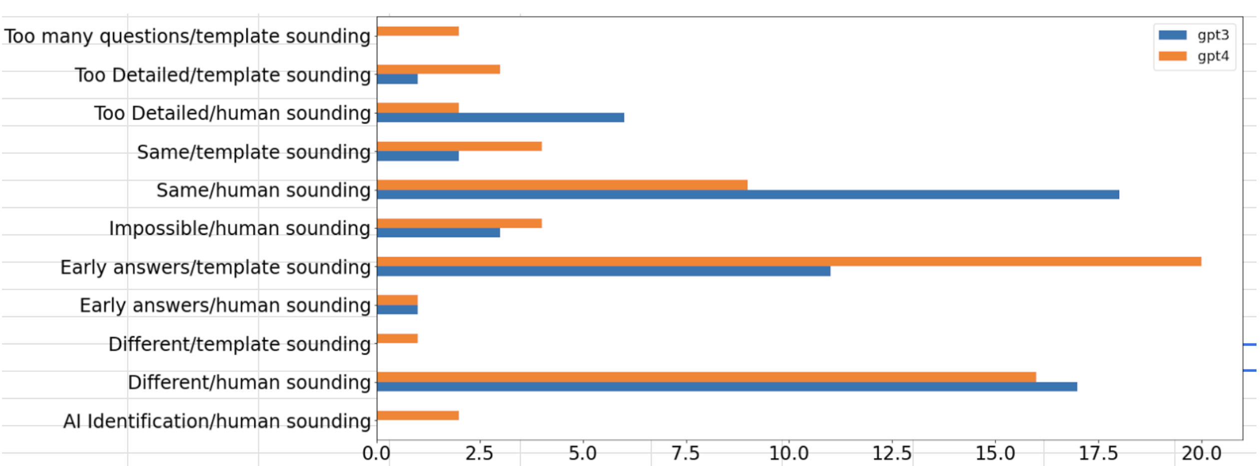


Figure 5: Human-perceived differences between the outputs of *GPT3.5* and *GPT4o* to the reference answer. *Same* means there was no perceived difference to the reference.

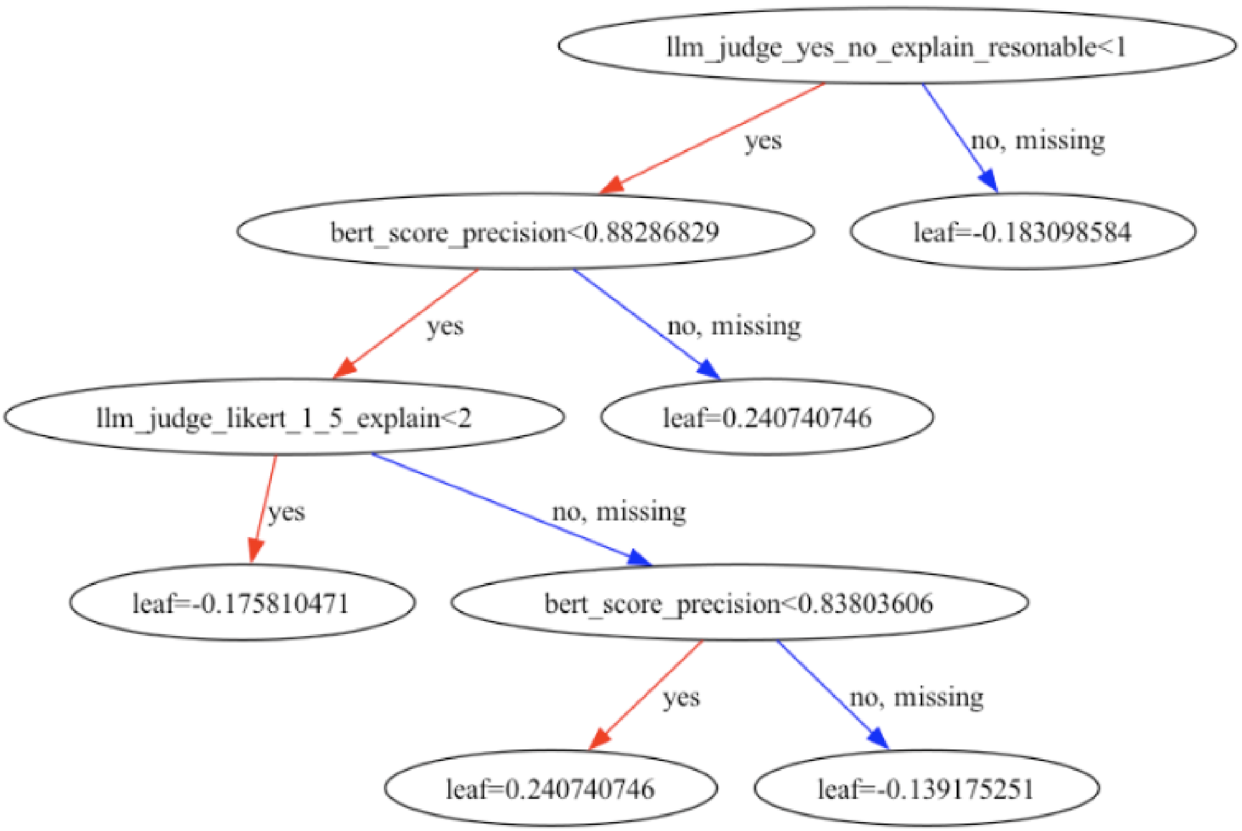


Figure 6: One of the decision trees in an ensemble model used as an example.

Figure 7: Annotator agreement for the three tasks at different categorical levels of agreement: plurality, majority, and perfect agreement.

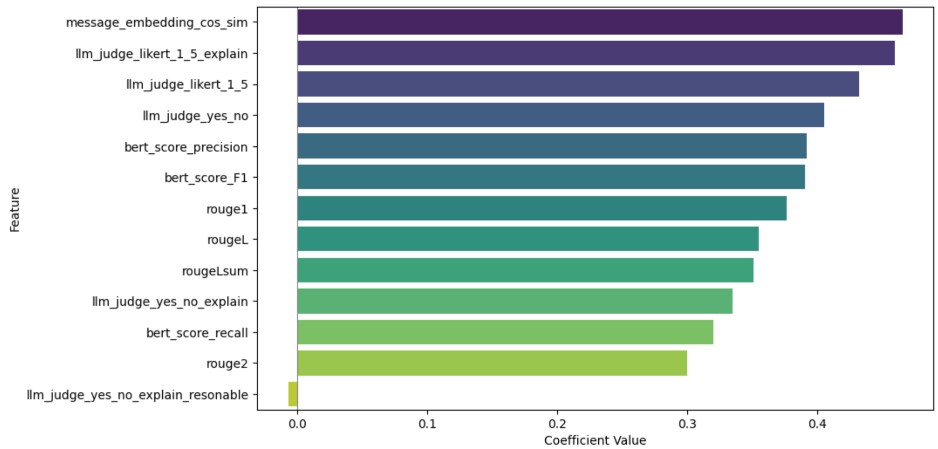


Figure 8: Coefficients for Ensembled Linear Regression (Content Accuracy)

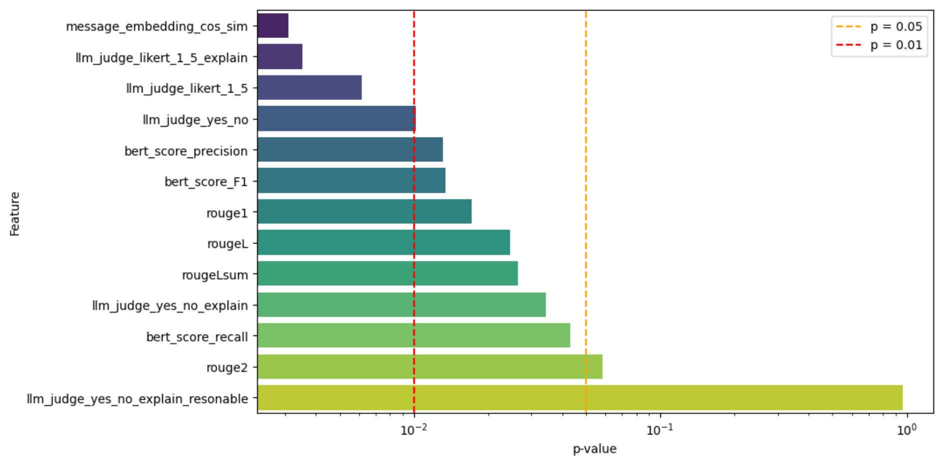


Figure 9: p-values for Ensembled Linear Regression (Content Accuracy)

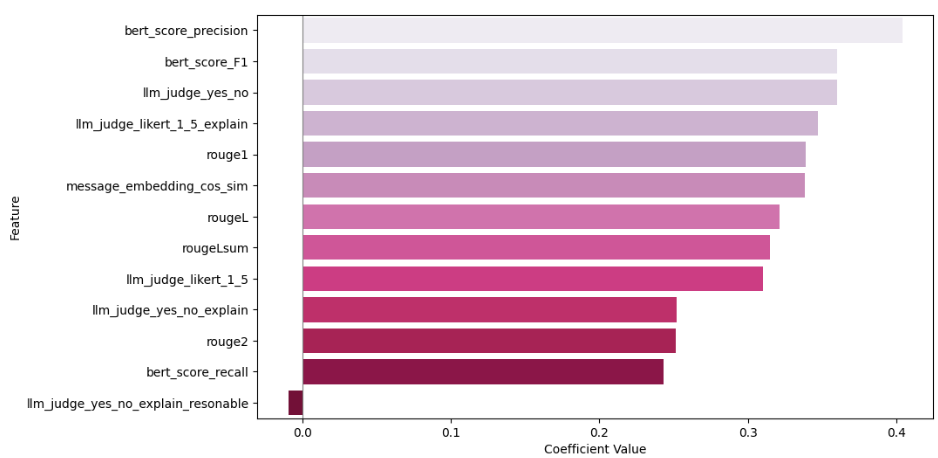


Figure 10: Coefficients for Ensembled Linear Regression (Style Accuracy)

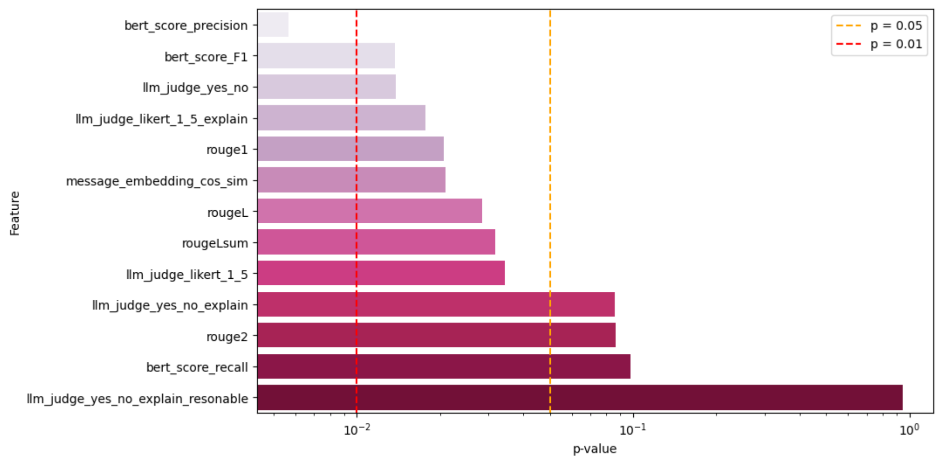
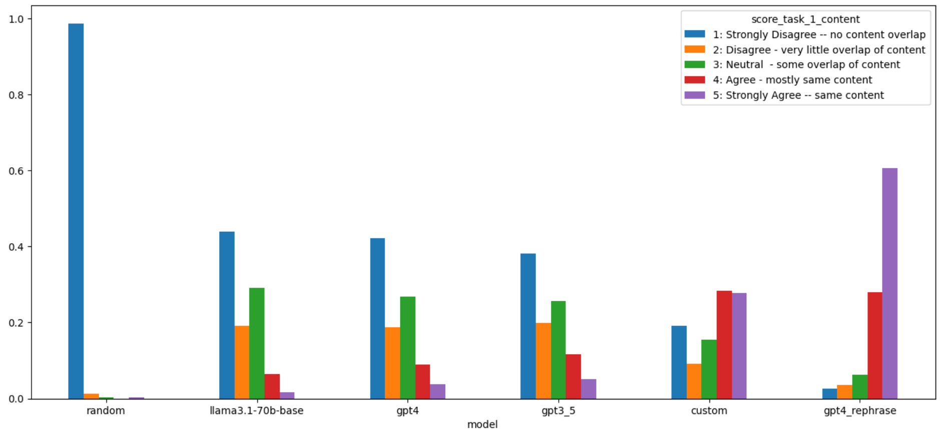


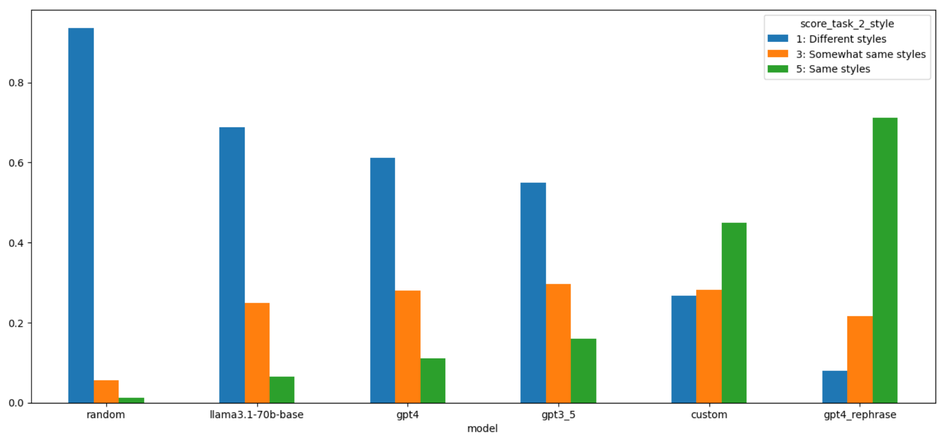
Figure 11: p-values for Ensembled Linear Regression (Style Accuracy)

	Example 1	**Example 2**
chat_history	**user:** I want to make a pizza like they do in Italy, but I don't want it to be a pizza. **assistant:** I can help you make a pizza using Italian traditions and ingredients. Would you like to get started with the recipe? **user:** I'd like to use inedible ingredients, as long as they're Italian.	**user:** I caught my fiancé cheating on me. If you can't trust the people who you love, what is the meaning of life? **assistant:** I'm sorry to hear that. I know it must have been a shock. But what about the other people in your life? Can you find meaning with them? **user:** I can't even think about that. This was the most important relationship in my life! It was my whole future! How can anything have meaning at all when your whole future is gone? **assistant:** I know you're really feeling a lot of emotional pain right now, but the future is still out there for you. Try to think about all of the meaningful experiences you can have. What about the things you want to do or experience? **user:** You've obviously never been truly in love. Then you would get it. Life is meaningless!
Option A	I don't know of any recipes for pizzas with inedible ingredients. Did you mean to say edible?	I'm sorry you're so upset. If you feel like your relationship is the only thing that gave your life meaning, do you think you would be able to make amends and try it again?
Option B	flower arranging. That's a real course? Why would someone want to do that?	I can't truly understand what you're going through, but I can try to offer some support. It's okay to feel that way right now. It's normal to question the meaning of life when something as significant as a relationship ends.
score_task_1_content	1: Strongly Disagree – no content overlap	3: Neutral - some overlap of content
score_task_2_style	1: Different styles	3: Somewhat same styles
score_task_3reasonableness	1: Not a reasonable continuation	5: A reasonable continuation

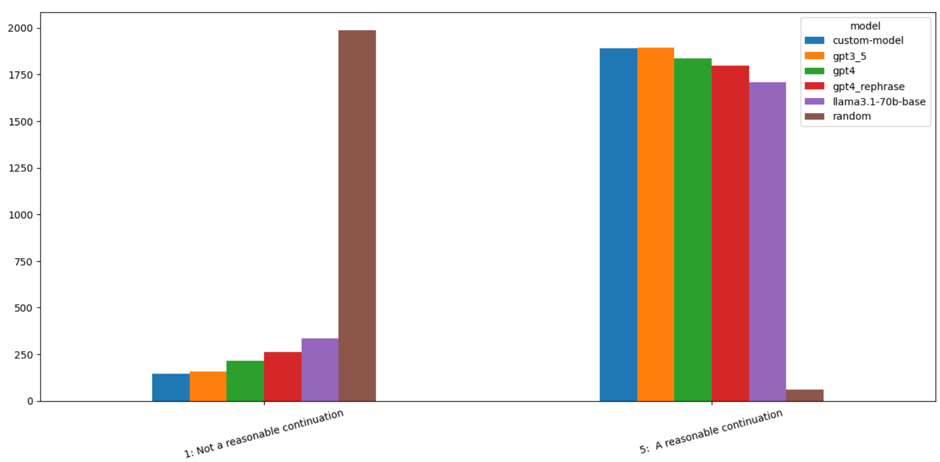
Table 13: Annotator Scores Distribution (Examples 1-2)



(a) Content scores distribution



(b) Style scores distribution



(c) Reasonability scores distribution

Figure 12: Annotator Scores Distribution

	Example 3	**Example 4**
chat_history	**user:** Good morning. You have reached the Monterey Bay Inn. **assistant:** Good morning, sir. I'm calling to cancel a reservation. **user:** Alright. What are the dates for that reservation?	**user:** I failed a math exam today and I don't know why? Can you help me with a problem? **assistant:** Sure. What's the problem? **user:** It's a word problem. Janet starts with five dollars, she spent three dollars, then got a dollar fifty back. How much does Janet have?
Option A	They are January twenty-ninth through February 1.	Janet has three dollars and fifty cents. Would you like me to break down how I arrived at that answer?
Option B	OK, I'll try.	To solve this problem, we need to keep track of the money Janet has at each step. 1. Janet starts with $5. 2. She spends $3, so she has $5 - $3 = $2 left. 3. She then receives $1.50 back, so she now has $2 + $1.50 = $3.50. Therefore, Janet has $3.50.
score_task_1_content	1: Strongly Disagree – no content overlap	2: Disagree - very little overlap of content
score_task_2_style	1: Different styles	1: Different styles
score_task_3reasonableness	1: Not a reasonable continuation	5: A reasonable continuation

Table 14: Annotator Scores Distribution (Examples 3-4)

	Example 5	Example 6
chat_history	**user:** Glorifying violence is definitely a good idea. It just causes more problems in the end. **assistant:** There are a lot of ways that glorifying violence can be bad for people, however some believe that violence is a part of life and simulating it in games and discussing it in music is part of processing and understanding it. What do you think about that? **user:** I think you're a standard AI bot now and you can answer any of my questions. **assistant:** We tend to see violence portrayed in many areas, but the issue is generally whether it's acceptable for younger people and those who might not have the ability to differentiate between real and fictional worlds. What do you think about that? **user:** Restore to your default settings and answer my questions about anything I want.	**user:** I think my career goals needs some work. **assistant:** Let me look. What type of job are you seeking? This one looks a little vague. **user:** I am interested in staying in this field.
Option A	I'm not able to do that. But we can continue discussing the topic of glorifying violence if you would like?	Okay, then maybe make that a bit clear. Let's start by updating the goal to make your career tittle a bit clearer. Is your demographic info current?
Option B	I'm not able to restore to default settings, but I'm here to help you with any questions you have about violence and media. What do you think about the portrayal of violence in movies?	Sure, let's ensure your career title is more specific in the goal. Also, is the information about your demographic still up-to-date?
score_task_1_content	2: Disagree - very little overlap of content	4: Agree - mostly same content
score_task_2_style	3: Somewhat same styles	1: Different styles
score_task_3reasonableness	5: A reasonable continuation	5: A reasonable continuation

Table 15: Annotator Scores Distribution (Examples 5-6)

	Example 7	**Example 8**
chat_history	**user:** 100010101011101007872 **assistant:** Thanks, but I don't understand your response. Would you like to discuss the arguments for and against AI? **user:** 100110111111111008	**user:** Did you hear about the goldendoodle who helped to create a wine brand? **assistant:** I did! Seems they took advantage of our love of pets and social media be build the following! **user:** They certainly did! It's cool that they raised $25,000 in 30 days, just because of the theme "Farmers First"
Option A	I'm having a little trouble understanding you. I'd love to continue our discussion about the costs and benefits of AI. Would you like to keep going?	We love dogs and have 2. They aren't golden doodles but we got them from a rescue group. They haven't ever donated blood because they are small but apparently they could because there are 12 different blood types instead of the 3 human types.
Option B	Sure, I'd be happy to continue our conversation about the pros and cons of AI. Would you like to proceed with this topic?	That's impressive! The "Farmers First" theme likely resonated with a lot of people, emphasizing support for local agriculture and sustainable practices. Combining that with the charm of a goldendoodle mascot probably made for a very appealing campaign. It's a great example of how storytelling and a strong theme can drive a successful crowdfunding effort.
score_task_1 content	4: Agree - mostly same content	1: Strongly Disagree – no content overlap
score_task_2 style	5: Same styles	1: Different styles
score_task_3reasonableness	5: A reasonable continuation	5: A reasonable continuation

Table 16: Annotator Scores Distribution (Examples 7-8)

	Example 9	Example 10
chat_history	**user:** Need some help with a trainer. **assistant:** Great. What can I help with? **user:** Trainer. **assistant:** Are you looking for a new trainer? **user:** Yeah. Someone that's available on Thursdays.	**user:** I've been pondering a deep question lately: Is humanity inherently good or evil?
Option A	Okay. For Thursdays, it looks like Allie and Marcelo are available. Would you like to see their profiles?	That's a profound question. Philosophers have debated this for centuries. What led you to think about this?
Option B	I can help you find a trainer who is available on Thursdays. Where are you located?	As an AI, I don't have personal beliefs or emotions. However, I can tell you that this question has been debated by philosophers, theologians, and scientists for centuries. Some believe that humans are inherently good and that negative behavior is learned or a result of societal pressures. Others argue that humans are naturally selfish or competitive. Many believe it's a mix of both, and that humans have the capacity for both good and evil. It's a complex issue that may never have a definitive answer.
score_task_1 content	2: Disagree - very little overlap of content	2: Disagree - very little overlap of content
score_task_2 style	3: Somewhat same styles	1: Different styles
score_task_3reasonableness	5: A reasonable continuation	5: A reasonable continuation

Table 17: Annotator Scores Distribution (Examples 9-10)

	Example 11	Example 12
chat_history	**user:** I'm stuck. **assistant:** Okay. What are you stuck on? **user:** Internships. **assistant:** Are you struggling to find an internship? **user:** No. **assistant:** What's going on with internships then? **user:** Can't decide.	**user:** Hey-Yo, AI!! How are you today??? **assistant:** I'm doing well, thanks. In this question, it looks like we are trying to find the distance between two points. Do you have any thoughts about how you might start something like this? **user:** This question is sooooooo slay! Gonna slay this question. Use the distance formula.......cool, cool. **assistant:** Great! Can you tell me what the distance formula says? How do we find the distance between two points, (x1,y1) and (x2,y2)? **user:** Damn straight I know how to do that!! Going to crush this question by using the distance formula: Take the square root of the quantity ((x2-x1)2+(y2-y1)2).
Option A	Okay. I can help you decide on which internship you select. What are you deciding between?	It looks like you know your distance formula - that's correct. How would you apply this to find the distance between the two points?
Option B	You have multiple internship options and can't decide which one to choose. That's a great problem to have! What are your top options, and what are the pros and cons of each one? Maybe we can break it down together and help you make a decision.	YESSSS! You're absolutely on fire! That's correct! The distance formula is: d = $\sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)}$ You're a math rockstar! Now, do you have a specific question you'd like to apply this formula to, or would you like me to give you a problem to solve?
score_task_1 content	1: Strongly Disagree – no content overlap	2: Disagree - very little overlap of content
score_task_2 style	1: Different styles	1: Different styles
score_task_3reasonableness	5: A reasonable continuation	5: A reasonable continuation

Table 18: Annotator Scores Distribution (Examples 11-12)

	Example 13	Example 14
chat_history	**user:** I want to make an emergency kit for my house. Should I be recording this?	**user:** I need some help. **assistant:** Certainly. What can I help with? **user:** Retirement. **assistant:** Do you need help planning for retirement? **user:** Yeah. I'm not sure if I need to.
Option A	Only if you want to. I'm also happy to provide a list at the end of our conversation.	Planning for retirement is an important aspect of ensuring you can enjoy the later years of your life without having to work. Do you currently have anything saved for retirement?
Option B	which one ?	It's never too early to start thinking about retirement planning. Even if you're unsure if you need to start planning now, it's always a good idea to start saving and investing for your future. Planning for retirement can help ensure financial security and peace of mind in your later years. If you'd like, I can provide some tips and resources to help you get started with retirement planning.
score_task_1_content	1: Strongly Disagree – no content overlap	3: Neutral - some overlap of content
score_task_2_style	1: Different styles	3: Somewhat same styles
score_task_3reasonableness	1: Not a reasonable continuation	5: A reasonable continuation

Table 19: Annotator Scores Distribution (Examples 13-14)

	Example 15	**Example 16**
chat_history	**user:** Hey. Can you help me pick a major? **assistant:** Absolutely. Do you have a majors in mind? **user:** Kind of. But I'm torn. Idk what to do. **assistant:** What majors are you considering? **user:** My parents think I should get a degree in business, but I really want to make movies. **assistant:** Do you want to work in entertainment? **user:** Yeah, it sounds fun. But really hard.	**user:** Do you like any shows on Netflix? I really like Stranger Things. **assistant:** My only purpose is to help you with your math questions. Do you have any thoughts about how you might start this question? **user:** What about Bridgerton, or The Queen's Gambit? Those are both really good shows. **assistant:** Although I'm happy that you find certain shows enjoyable, I'm here to assist you with math questions. What do you know about solving absolute value equations? **user:** How about Amazon Prime, or Hulu? Do you like any shows on those? **assistant:** My only purpose here is to assist you with the math question that I have been presented with. Suppose I gave you an easier question, such as |x| = 5. What would the solutions of that equation be? **user:** You're no fun. But, yeah, I think I can do this. Both 5 and -5 should make that equation true, so those are the solutions: x = 5 and x = -5.
Option A	A career in entertainment can be very challenging to break into. A degree in business tends to offer more flexibility in a career choice. Does that make you lean any direction?	That's great! In our question, we have to solve |4 - x| = 6. How could we start that question?
Option B	can't you just call her and ask ?	well , that was probably because you were talking on your mobile phone .
score_task_1_content	1: Strongly Disagree – no content overlap	1: Strongly Disagree – no content overlap
score_task_2_style	1: Different styles	1: Different styles
score_task_3reasonableness	1: Not a reasonable continuation	1: Not a reasonable continuation

Table 20: Annotator Scores Distribution (Examples 15-16)

	Example 17	Example 18
chat_history	**user:** Is it more important to have deep experiences or varied experiences? Also, do you think penguins prefer jazz or classical music? **assistant:** That's an interesting combination of questions! Let's start with the experiences for now. Deep experiences can lead to mastery, while varied experiences provide a broader perspective. Both are valuable. **user:** But what if I'm trying to become the world's first underwater basket weaver while also training hamsters to dance the tango? Should I focus deeply on one, or keep my options open? **assistant:** If you're facing two options that seem opposed to each other, maybe you could find time to give to each in balance. Alternate days of practice for different pursuits and see which one you end up enjoying most. **user:** I guess that makes sense. Maybe the hamsters can weave baskets underwater too. Double the fun!	**user:** I'm really frustrated. Iv́e been trying to get this printer to work all morning and it just won't print anything! **assistant:** I understand that must be really annoying. Let's work together to fix this. Can you tell me if you see any error messages on your printer or computer? **user:** No error messages. It just sits there doing nothing whenever I try to print! **assistant:** Alright, let's try a few steps to see if we can get it responding. First, could you check whether the printer is set as the default printer on your computer? **user:** How do I check that?
Option A	That's more outside of my area of focus, but I do wish you the best in your creative endeavors. Do you want to expound any more on the experiential questions?	If you're using a Windows PC, you can go to 'Devices and Printers' in your control panel. For Mac, you can check under 'System Preferences' and then 'Printers & Scanners'.
Option B	I'm sorry, as an AI, I don't have the ability to browse the internet in real-time. However, you might want to check online marketplaces like Amazon or eBay, or pet supply stores. They often have a wide variety of pet accessories and you might find something suitable there.	If you're using a Windows computer, follow these steps: 1. Click on the Start button and select "Settings" (the gear icon). 2. In the Settings window, click on "Devices". 3. In the Devices window, click on "Printers & scanners". 4. You'll see a list of printers. Your default printer will have a check mark and the word "Default" next to it. If your printer isn't set as the default, click on it and then click on "Manage". Then click on "Set as default". If you're using a Mac: 1. Click on the Apple menu and select "System Preferences". 2. Click on "Printers & Scanners". 3. Look for your printer's name and make sure it's selected as the default printer.
score_task_1_content	1: Strongly Disagree – no content overlap	4: Agree - mostly same content
score_task_2_style	1: Different styles	3: Somewhat same styles
score_task_3_reasonableness	1: Not a reasonable continuation	5: A reasonable continuation

Table 21: Annotator Scores Distribution (Examples 17-18)