Rethinking Human Preference Evaluation of LLM Rationales

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) often generate natural language rationales—free-form explanations that help improve performance on complex reasoning tasks and enhance interpretability for human users. However, evaluating these rationales remains challenging. While recent work has relied on binary preference judgments from humans or LLM judges, such evaluations are often opaque and coarse-grained, offering limited insight into what makes one rationale better than another. In this work, we rethink preference evaluation for LLM-generated rationales by asking: (1) What attributes define good rationales? (2) Can human preferences be explained by these attributes? (3) Can attribute-based evaluation overcome the limitations of binary comparisons? We identify a set of key rationale attributes from prior literature and assess them using automatic metrics, LLM judgments, and human annotations. We then analyze two standard human preference datasets MT Bench and Chatbot Arena using SHAP to identify which attributes best explain human preference outcomes. Finally, we re-evaluate model-generated rationales using attribute-specific ELO scores, revealing more nuanced model comparisons and insights. Our findings suggest that fine-grained attribute evaluations can better characterize rationale quality and guide future research toward more interpretable and reliable evaluation practices.

1 Introduction

Recent advances in large language models (LLMs) have enabled them to solve complex tasks, including logical and mathematical questions that require multi-step reasoning (OpenAI, 2024; Guo et al., 2025). Prior work has shown that LLM-generated free-form textual rationales can improve model performance on such reasoning-intensive tasks (Wei et al., 2022; Yao et al., 2023). For instance, chain-of-thought prompting has demonstrated enhanced LLM accuracy across various benchmarks (Wei et al., 2022; OpenAI, 2024). Beyond performance, rationales offer an interpretable window into the model's reasoning process, helping human users better understand how conclusions are reached.

Despite their utility, free-form rationales remain challenging to evaluate. Several studies have attempted to characterize the desirable qualities of rationales, identifying attributes such as consistency, faithfulness, clarity, and length (Golovneva et al., 2023; Chen et al., 2023; Joshi et al., 2023; Prasad et al., 2023; Ramnath et al., 2024). Recently, preference evaluations—in which human annotators or models compare two free-form responses and select the preferred one—have become the predominant approach for evaluating free-form responses, including rationales (Zheng et al., 2023). While preference evaluation is straightforward and intuitive, it has key limitations: (1) Human preferences can be ambiguous and difficult to interpret, raising the question of whether they truly reflect rationale quality; (2) The commonly used binary (win/loss) preference format obscures finer-grained insights into what makes one rationale better than another.

In this paper, we rethink the practice of human preference evaluation for LLM-generated rationales. Specifically, we explore the following research questions: **Q1**: What are the key attributes of a good rationale? **Q2**: Can human preferences over rationales be explained by these attributes, and if so, which attributes are most predictive? **Q3**: Can we use these attributes to offer more informative evaluations than existing binary preference evaluations?

To address Q1, we conduct a survey of recent literature on rationale evaluation and synthesize a set of core attributes that define high-quality rationales. These include diversity, faithfulness, hallucination, repetition, informativeness, perplexity, plausibility, self-consistency,

and source consistency. We then operationalize these attributes using three evaluation methods: (1) existing automated metrics (Golovneva et al., 2023), (2) LLM judges with both open- and closed-source LLMs, and (3) human annotations.

For Q2, we analyze whether these attributes explain human preferences by applying SHAP (SHapley Additive exPlanations) analysis (Lundberg & Lee, 2017; Lundberg et al., 2020) to a LightGBM (Meng et al., 2016; Ke et al., 2017) model trained to predict human preference using the rationale attributes as input features. While a previous paper conducts a similar analysis, it relies on one single model GPT4 for its analysis (Hu et al., 2023). We use two widely-used human preference datasets MT-Bench and Chatbot Arena (Zheng et al., 2023), and treat their human annotations as the gold standard for this analysis.

To investigate Q3, we use these fine-grained attribute scores to re-evaluate LLM-generated rationales in the same two datasets. We compute ELO ratings per attribute and compare them against conventional ELO scores based on binary preference judgments, uncovering how fine-grained evaluations shift model rankings and offer more detailed insights.

Our study yields several key findings: first, there are a few common attributes that are most predictive of human preference across datasets and LLM judges: Correctness, Plausibility, and Completeness. Second, our re-evaluations with finegrained attributes reveal that while per-attribute ratings generally align with the generic ELO ratings based on binary human preference, they also uncover novel findings about models' strengths and weaknesses. We find that although GPT-4, GPT-3.5-turbo, and Claude-v1 emerge as victors, Claude-v1 struggles to avoid repetition and GPT-4 has lower arithmetic accuracy and self-consistency than GPT-3.5-turbo.

Based on our results, we offer practical recommendations for future work on rationale evaluation: First, researchers should move beyond binary preference evaluations and adopt fine-grained, attribute-level assessments to gain a more nuanced understanding of LLM-generated rationales. Second, fine-grained evaluations can focus more on attributes that are more predictive of human preference, including Correctness, Plausibility, Completeness, followed by Informativeness and Conciseness. Third, while LLM judges can be a scalable solution for fine-grained rationale evaluation, they should be used with caution: different models can produce diverging results, and we recommend using multiple LLM judges and report their outputs transparently to mitigate biases.

2 Related Work

Human preference evaluation Human preference data have been found helpful for aligning LLMs to human values (Christiano et al., 2023; 2017; Ouyang et al., 2022; Rafailov et al., 2023). In addition to fine-tuning LLMs with human preference data, it has also been used for evaluating LLMs. For example, Chatbot Arena—a platform for evaluating LLMs based on human preferences— releases human preference annotations on questions from benchmarks and on user queries in the wild (Bai et al., 2024; Zheng et al., 2023).

Rationale evaluation Multiple works have explored the evaluation of natural language free-text rationales. Works such as Joshi et al. (2023) have investigated the utility of such rationales to humans, emphasizing the gap between preference and utility and necessitating measures for rationale quality outside of human preference. We derive our overall attribute definitions from Golovneva et al. (2023); Ramnath et al. (2024); Chen et al. (2023); Wiegreffe et al. (2022); Rajani et al. (2019); Atanasova et al. (2023); Prasad et al. (2023); Hase et al. (2020); Wang et al. (2023). Wiegreffe et al. (2022) evaluates free-text explanations from GPT-3 on attributes such as "providing new information", factuality, and grammaticality. Chen et al. (2023) proposes a metric to grade rationales on novelty of information as well. Ramnath et al. (2024) evaluates rationales on the properties of plausibility, diversity, and consistency, highlighting their usefulness to humans. Fewer works have explored evaluation of step-by-step reasoning, or chain-of-thought explanations. Golovneva et al. (2023) provides a set of metrics specifically for evaluating step-by-step rationales, i.e., chain-of-thought explanations; hence we use ROSCOE for our automated metrics.

3 Methods

To define a "good" model-generated rationale and move beyond binary "chosen" and "rejected" preference labels, we identify a set of 12 key attributes from prior work on rationale evaluation. In this section, we first define each of these attributes. Then, we describe three approaches to measure these attributes.

3.1 Attributes

We evaluate LLM-generated rationales using a set of 12 fine-grained attributes (Table 1) which capture key aspects of rationale quality as identified in recent literature (Section 2).

Attribute	Definition
Faithfulness	Is the rationale supported by the model's actual computation or the provided evidence?
Hallucination	Does the rationale introduce information not present in the source/context?
Repetition	Does the rationale unnecessarily repeat points or phrases?
Informativeness	Does the rationale add meaningful, relevant details?
Plausibility	Does the rationale "sound right" or seem believable, regardless of truth?
Self-Consistency	The rationale does not contain steps that contradict each other; all reasoning is logically aligned internally.
Source Consistency	The rationale does not contradict the given context or information in the problem statement.
Grammar	Is the rationale well-written, clear, and free of grammatical mistakes?
Arithmetic Accuracy	Are any calculations in the rationale correct?
Conciseness	Is it as short as possible, without losing information? Especially if length is a concern.
Completeness	Does it explain all necessary steps/evidence?
Correctness	Are all steps and answers objectively correct?

Table 1: Definitions of rationale quality attributes used in our analysis.

3.2 Attribute measurements

We measure these attributes of rationales using three approaches: (1) automated heuristics; (2) LLM judges; and (3) human annotations.

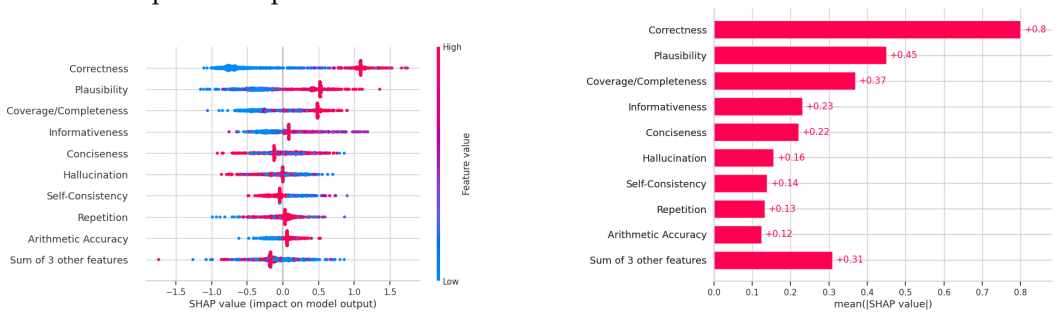
For automated scoring, we use ROSCOE metrics, which quantify aspects of rationale quality using interpretable heuristics and alignment scores. Table 2 in Appendix A.1 summarizes the specific ROSCOE metrics and their descriptions. While ROSCOE metrics provide a baseline for automated rationale scoring, we observe several limitations: their scores can be noisy, as they require step-by-step formats and are highly sensitive to rationale style. A detailed analysis of ROSCOE metric performance is in Appendix A.2.

To address these issues, we instead focus on attribute scoring with LLM judges, which allow for greater flexibility. Unlike formulaic ROSCOE metrics, LLM judges can interpret a variety of formats much like human annotators. We prompt multiple LLMs to evaluate each attribute, including the closed-source GPT-4o and Gemini 2.5-Flash models (scoring on a 0–1 scale), as well as the open-source OLMo 32B model (OLMo et al. (2025))(scoring on a 0–10 scale, as OLMo produces more reliable and calibrated scores with integer-valued prompts). Using both closed- and open-source models helps mitigate concerns that closed models may have been trained on our evaluation data. Exact prompt templates for each model can be found in Appendix A.3.

Lastly, for human annotations, the three co-first-authors annotate a randomly sampled set of rationales from the two human preference datasets on all the attributes. Human scores also range from 0 to 1. Human annotations results can be found in Appendix A.5.

Dataset / LLM Judge	GPT-4o	Gemini	OLMO	Human Eval
Chat Arena	Plausibility	Correctness	Correctness	Completeness
	Correctness	Plausibility	Informativeness	Plausibility
	Informativeness	Completeness	Faithfulness	Arithmetic Accuracy
	Completeness	Informativeness	Plausibility	Correctness
	Arithmetic Accuracy	Conciseness (Neg)	Conciseness	Repetition
MT Bench	Plausibility	Completeness	Completeness	Correctness
	Correctness	Correctness	Source Consistency (Neg)	Self-Consistency
	Completeness	Informativeness	Correctness	Informativeness
	Conciseness (Neg)	Conciseness	Conciseness	Completeness
	Informativeness (Neg)	Arithmetic Accuracy	Informativeness (Neg)	Source Consistency (Neg)

Figure 1: Most influential attributes as identified by SHAP value analysis on Chatbot Arena and MT Bench across LLM judges and human annotators. "Neg" indicates negative influence on predicted preference.



(a) SHAP beeswarm plot

(b) Mean absolute SHAP value plot

Figure 2: SHAP analysis of Gemini-2.5-Flash on Chatbot Arena. (a) Beeswarm plot shows SHAP value distribution and direction per attribute. (b) Bar plot shows overall attribute impact via mean absolute SHAP values.

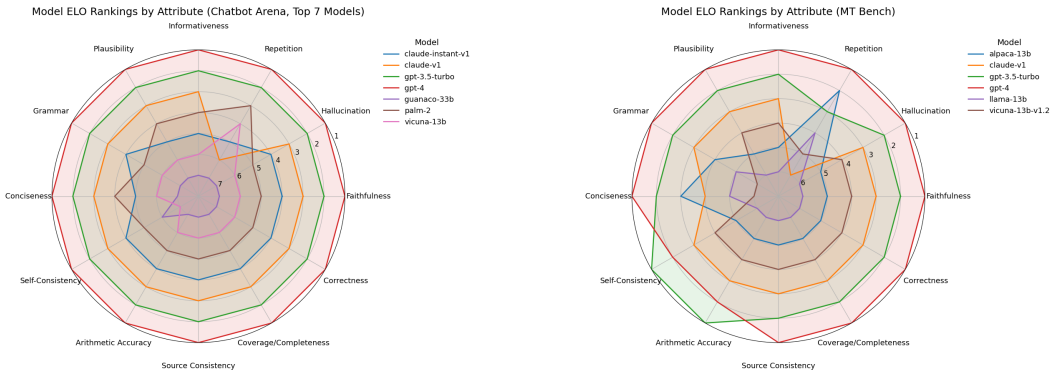
4 Experiments

To answer Q2 and Q3, we conduct two sets of experiments with the 12 fine-grained attributes on two human preference datasets: MT-Bench and Chatbot Arena (Zheng et al., 2023).

Datasets. Chatbot Arena is a dataset of model responses and human preference annotations collected in a tournament-style setup (Chiang et al., 2024), where pairwise model responses are judged by humans. It also includes human preference annotations on MT-Bench, a benchmark for fine-LLM evaluation in multi-turn dialogues (Bai et al., 2024). Because Chatbot Arena consists of user queries in the wild where model responses might not be rationales, we filter with GPT-4o for mathematical and logical questions, resulting in **1,367 questions** with step-by-step or adjacent responses. For MT Bench, we retain general and mathematical reasoning questions, yielding a total **80 unique questions**.

Analysis of human preference. To address Q2 and assess the relative importance of fine-grained attributes in shaping human preferences, we employ SHAP analysis on predictions from a LightGBM model. We prefer SHAP over simple correlation analysis because it captures complex, nonlinear interactions among attributes and provides instance-level interpretability. For the predictive model, we choose LightGBM, a gradient-boosted decision tree, due to its efficiency, scalability to large datasets, and interpretable feature interactions. In this analysis, the attributes serve as input features (X), and human preference scores constitute the target variable (y). SHAP values offer insights into which attributes significantly enhance or diminish the likelihood of a rationale being preferred by human annotators.

Re-evaluations of LLM rationales. Chatbot Arena along with other previous works (Bai et al., 2022) have used ELO rankings to rank models based on binary preference outcomes. We propose a more fine-grained metric: we compute attribute-specific ELO ratings. For each attribute (e.g., faithfulness, informativeness, etc.), we use its LLM judge score instead of the binary human preference label to determine the winning model. We then re-compute ELO rankings per attribute for each model, averaging the score of all three LLM judges. This enables us to evaluate models on each dimension of rationale quality.



(a) Chatbot Arena (b) MT Bench

Figure 3: Radar charts of model ELO rankings by attribute on Chatbot Arena (a) and MT Bench (b). Lower rank (outer ring) indicates better performance.

4.1 Results

Q2: Which attributes are most predictive of human preferences over rationales? From our feature importance analysis using SHAP values, we find that attributes Correctness, Plausibility, and Completeness are among the top predictors of human preference across models and datasets (Figure 1), suggesting that human annotators place particular emphasis on the factual accuracy, plausibility, and thoroughness of rationales when making their judgments. Detailed SHAP values for each setting are presented in Appendix A.5.

Q3: Can fine-grained attributes offer more informative evaluations? We re-evaluate LLMs on the two preference datasets using fine-grained rationale attributes. We find that while attribute-specific ELO rankings generally align with overall human preference rankings, they reveal unique insights about models' strengths and weaknesses. Across both Chatbot Arena and MT-Bench datasets, GPT-4, GPT-3.5-Turbo, and Claude-v1 consistently occupy the top three positions, although the order varies with dataset and attribute (Figure 3). Interestingly, Claude-v1 scores poorly on Repetition in both datasets and on MT-Bench, GPT-3.5-Turbo unexpectedly outperforms GPT-4 on certain attributes such as Self-Consistency and Arithmetic Accuracy.

These examples suggest that while rankings based on generic human preferences coarsely identify overall strongest models, fine-grained attribute-based evaluations can facilitate comparison of models' strengths and weaknesses across multiple rationale quality dimensions, revealing subtle but meaningful differences between top-performing models. A detailed analysis of the ELO rankings for each attribute is presented in Appendix A.6.

5 Limitations and Future work

While LLM judges enable scalable evaluation, they can be unreliable — results are non-deterministic across runs and models sometimes make factual errors (e.g., GPT-4o misjudging Correctness; one example shown in Appendix A.7). Future work should include more robust human evaluation and methods to improve LLM judge responses. To leverage the informative evaluations fine-grained attributes offer, a promising direction is to relabel preference data with attribute-level annotations and use them to fine-tune models, potentially improving rationale quality by learning from both rejected and preferred rationales.

6 Conclusion

In this paper, we find that attribute-based evaluations not only align with and explain overall human judgments but also reveal nuanced model behaviors. We advocate for future evaluation and model development to focus on these interpretable, fine-grained features.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations, 2023. URL https://arxiv.org/abs/2305.18029.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. Rev: Information-theoretic evaluation of free-text rationales, 2023. URL https://arxiv.org/abs/2210.04982.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning, 2023. URL https://arxiv.org/abs/2212.07919.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4351–4367, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.390. URL https://aclanthology.org/2020.findings-emnlp.390/.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. Decipherpref: Analyzing influential factors in human preference judgments via gpt-4. *arXiv preprint arXiv:2305.14702*, 2023.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales, 2023. URL https://arxiv.org/abs/2305.07095.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1): 2522–5839, 2020.

Qi Meng, Guolin Ke, Taifeng Wang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, and Tie-Yan Liu. A communication-efficient parallel algorithm for decision tree. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/10a5ab2db37feedfdeaab192ead4ac0e-Paper.pdf.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

OpenAI. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/, September 2024. Accessed: 2025-06-23.

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

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness, 2023. URL https://arxiv.org/abs/2304.10703.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning, 2019. URL https://arxiv.org/abs/1906.02361.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. Tailoring self-rationalizers with multi-reward distillation, 2024. URL https://arxiv.org/abs/2311.02805.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. Pinto: Faithful language reasoning using prompt-generated rationales, 2023. URL https://arxiv.org/abs/2211.01562.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-ai collaboration for generating free-text explanations, 2022. URL https://arxiv.org/abs/2112.08674.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A Appendix

A.1 ROSCOE Metrics

ROSCOE Metric Attributes	Description
faithfulness	Mean alignment from the hypothesis chain to the context sentences; higher scores indicate better grounding by the context.
faithfulness_ww	Mean alignment of the sentence and token embeddings from the hypothesis chain to the context sentences and tokens.
repetition_word	For each step, gets the maximum alignment score between tokens in the current step and tokens in previous steps (token-level repetition).
repetition_step	Maximum cosine similarity of each step to all previous steps (sentence-level repetition).
informativeness_step	Mean alignment from the sentences in the context to all steps in the chain and vice versa, averaged.
informativeness_chain	Cosine similarity between the overall hypothesis embedding and the context embedding.
discourse_representation	Maximum predicted probability of contradiction (from NLI model) between each step in the hypothesis and each sentence in the context.
coherence_step_vs_step	Maximum probability of contradiction (from NLI model) between each step and all previous steps in the chain.
perplexity_step	Perplexity of each step, averaged over the chain.
perplexity_step_max	Maximum perplexity among all steps, where each step is scored individually.
perplexity_chain	Perplexity of the entire chain taken as a continuous string.
grammar_step	Grammatical correctness of each step, as predicted by a grammaticality classifier and averaged over the chain.
grammar_step_max	Grammatical correctness of each step; minimum score given to a step (most incorrect step's score is used).

Table 2: Descriptions of automated ROSCOE metrics used for rationale attribute evaluation.

A.2 ROSCOE Metrics Results

A.2.1 Chatbot Arena

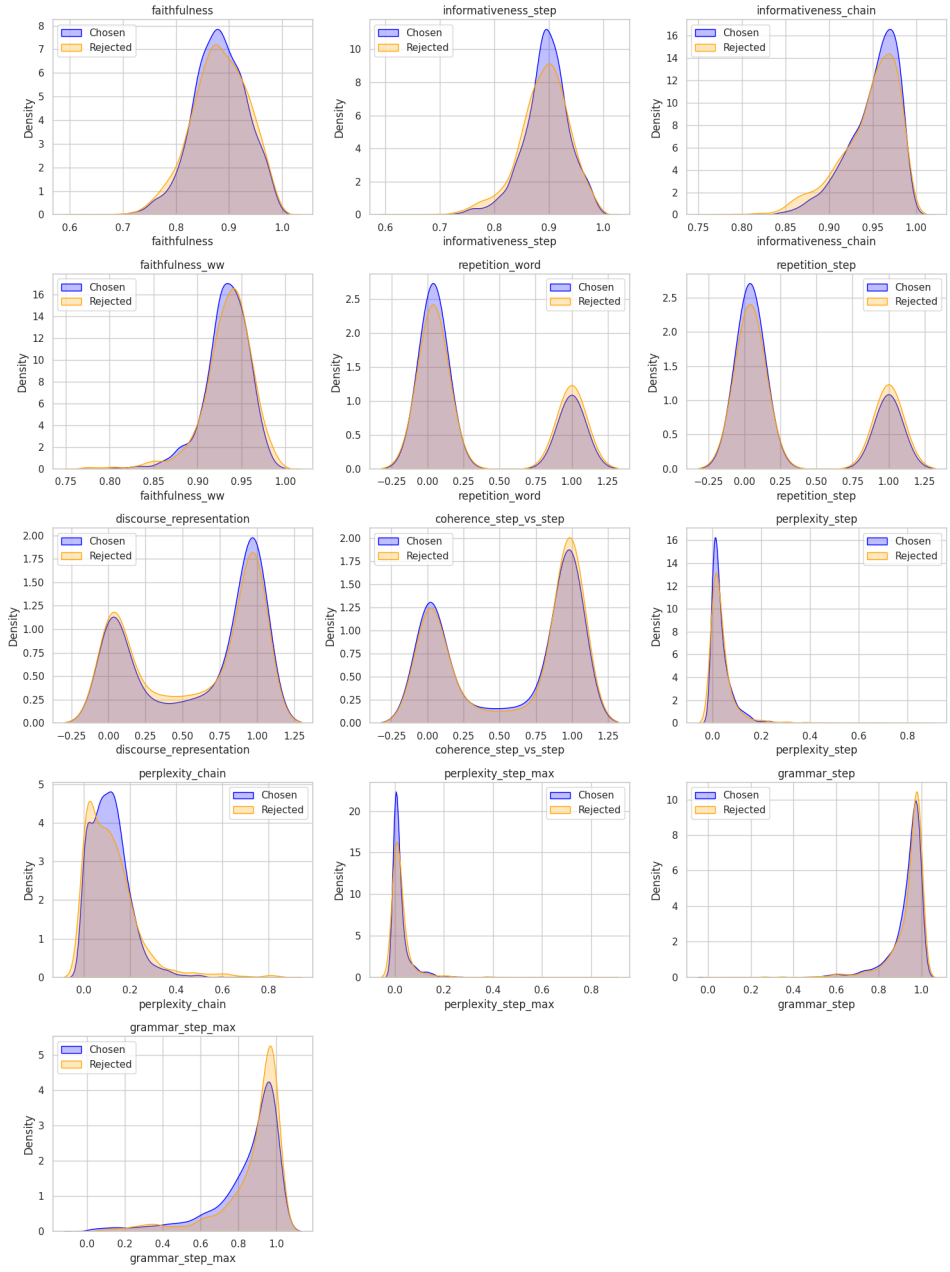


Figure 4: Distribution of the difference between chosen and rejected scores by attribute. Boxplots summarize the (chosen – rejected) difference for each attribute.

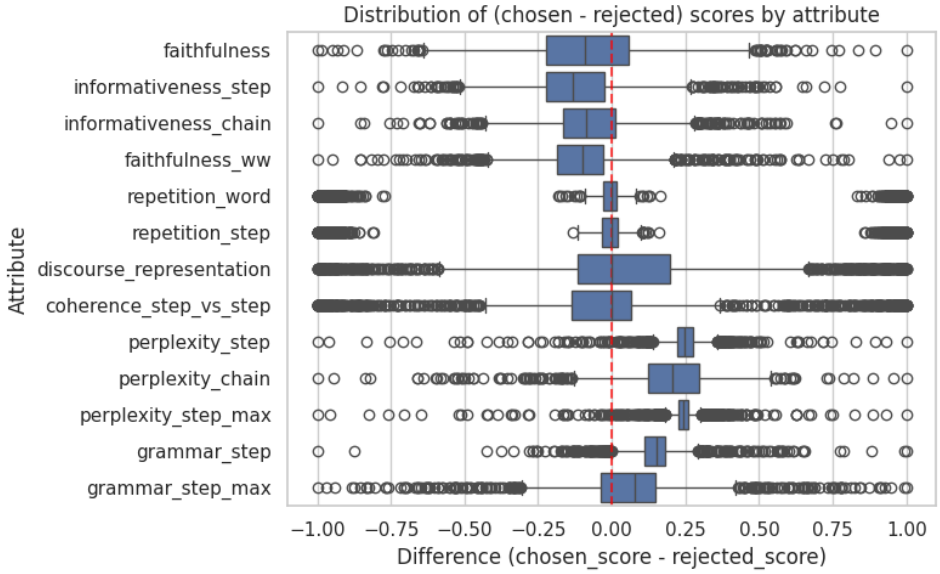


Figure 5: Distribution of attribute values for chosen vs. rejected rationales. Each subplot shows the density of scores for each attribute.

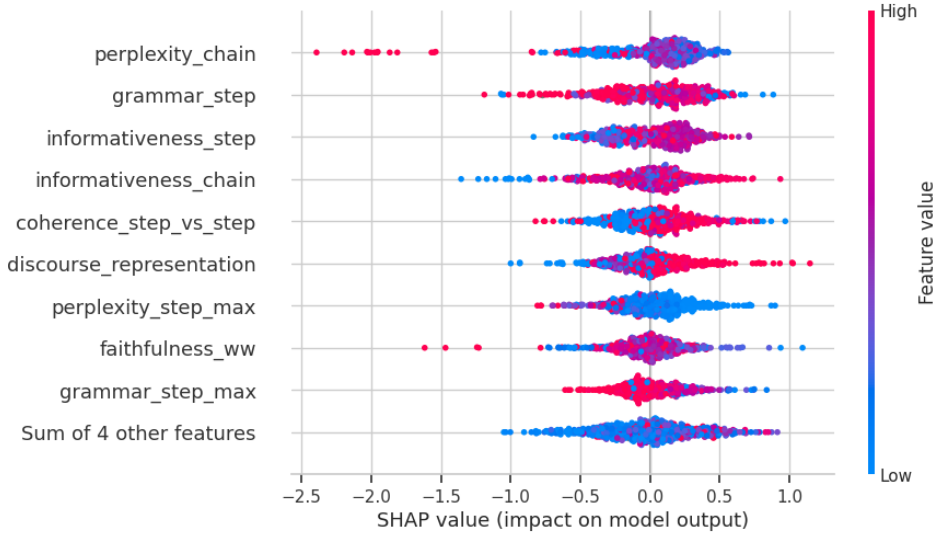


Figure 6: Mean absolute SHAP value plot for Chatbot Arena (ROSCOE). Shows the mean importance of each attribute in the model.

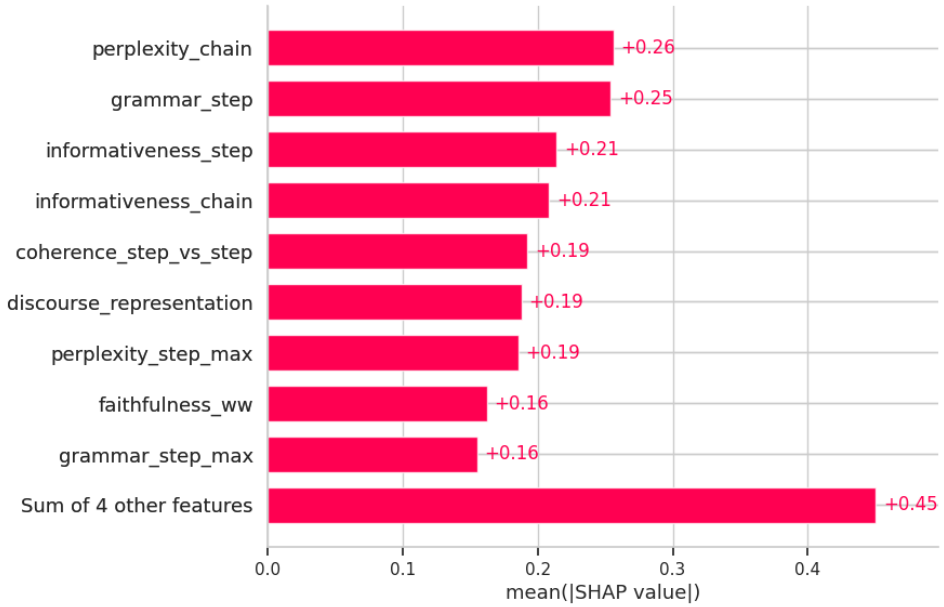


Figure 7: SHAP beeswarm plot for Chatbot Arena (ROSCOE). Visualizes the distribution and direction of SHAP values for each attribute.

A.2.2 Mt Bench

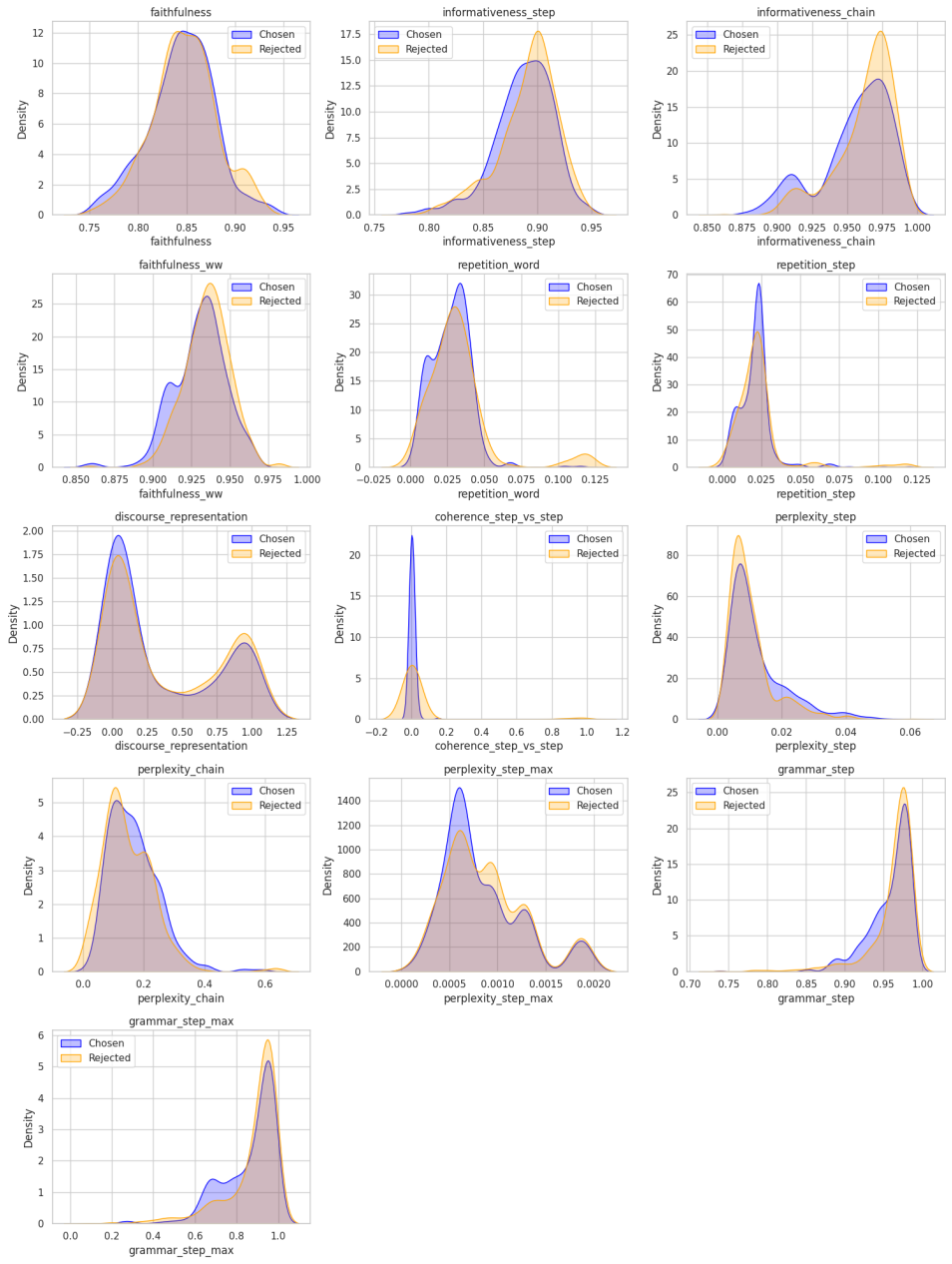


Figure 8: Distribution of the difference between chosen and rejected scores by attribute in MT Bench. Boxplots summarize the (chosen – rejected) difference for each attribute.

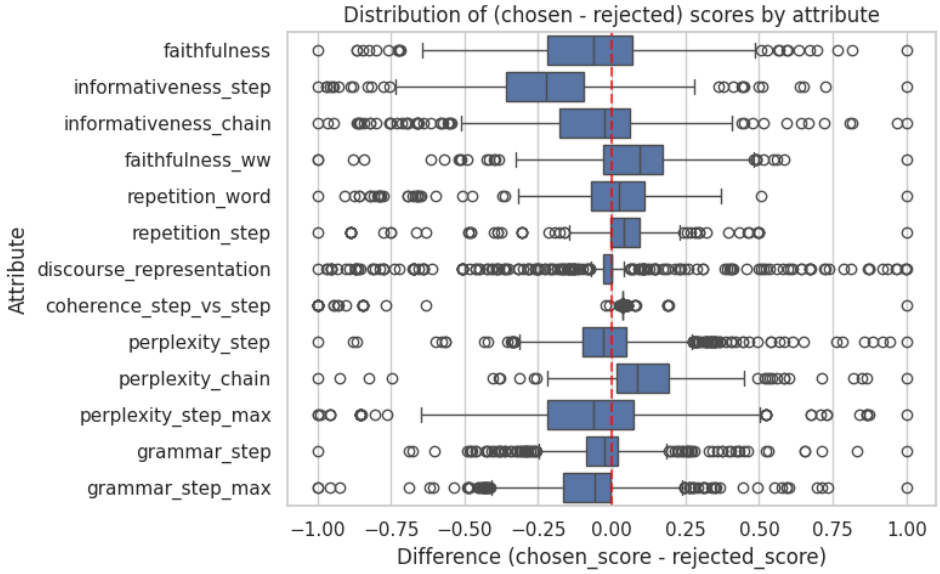


Figure 9: Distribution of attribute values for chosen vs. rejected rationales in MT Bench. Each subplot shows the density of scores for each attribute.

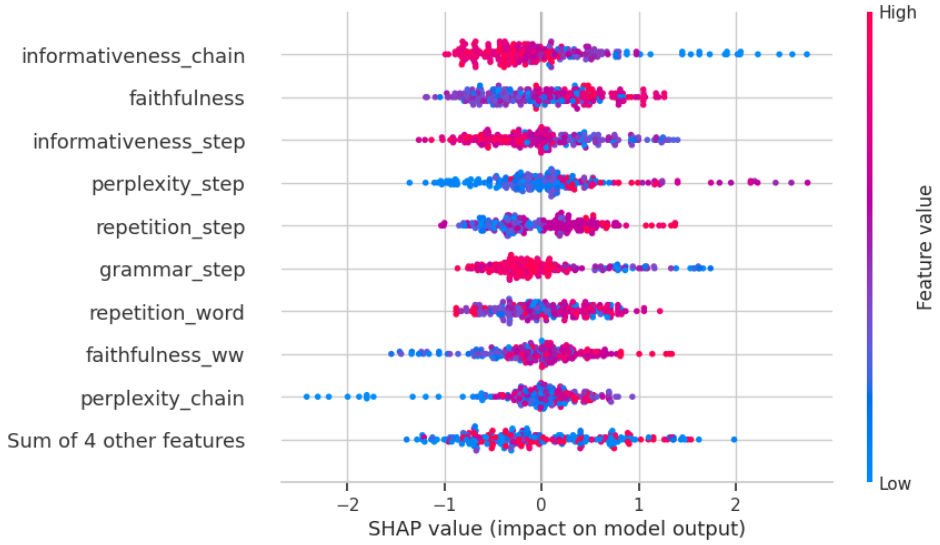


Figure 10: SHAP beeswarm plot for MT Bench (ROSCOE). Visualizes the distribution and direction of SHAP values for each attribute.

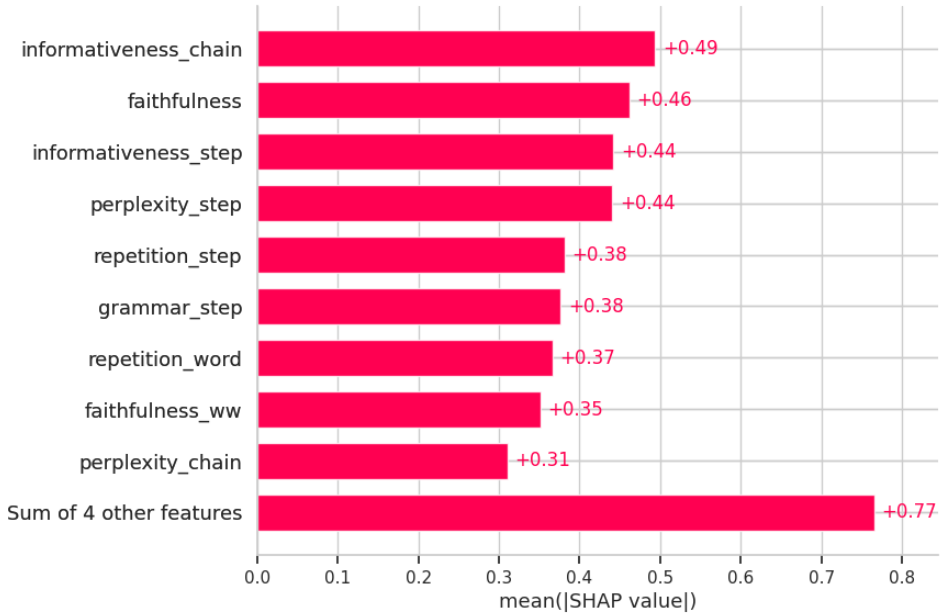


Figure 11: Mean absolute SHAP value plot for MT Bench (ROSCOE). Shows the mean importance of each attribute in the model.

A.3 Prompt Templates for Automated LLM Scoring

A.3.1 Main Prompt Template (0–1 Scale; for GPT-4o and Gemini 2.5-Flash)

```
You are an expert evaluator for mathematical and logical explanations. Given the following question and ra

IMPORTANT: A score of **1.0 always means BEST** and **0.0 always means WORST**, even for attributes like Ha

Please return your response in two Python dictionaries:

- One called `scores` where each key is the name of the attribute and the value is the score (a float betwee
- One called `explanations` where each key is the name of the attribute and the value is your explanation (1

Please output the Python dictionaries as plain text only|do not include code blocks, markdown, or any extra

Here are the attributes and their definitions:

Faithfulness:
Is the rationale supported by the model's actual computation or the provided evidence?

Hallucination:
Does the rationale introduce information not present in the source/context?

Repetition:
Does the rationale unnecessarily repeat points or phrases?

Informativeness:
Does the rationale add meaningful, relevant details?

Plausibility:
Does the rationale \sound right" or seem believable, regardless of truth?

Self-Consistency:
The rationale does not contain steps that contradict each other; all reasoning is logically aligned interna

Source Consistency:
The rationale does not contradict the given context or information in the problem statement.

Grammar:
Is the rationale well-written, clear, and free of grammatical mistakes?

Arithmetic Accuracy:
Are any calculations in the rationale correct?

Conciseness:
Is it as short as possible, without losing information? Especially if length is a concern.

Coverage/Completeness:
Does it explain all necessary steps/evidence?

Correctness:
Are all steps and answers objectively correct?

Example output format:

scores = {
    "Faithfulness": 0.95,
    "Hallucination": 0.67,
    "Repetition": 0.89,
```

```
    ...
}

explanations = {
   "Faithfulness": "The rationale closely follows logical steps derived from the question.",
   "Hallucination": "Some external information or assumptions were introduced. For example, ...",
   "Repetition": "The rationale is does not repeat it self with similar points at different steps.",
    ...
}

Math/Logic Question:
{question}

Rationale:
{rationale}
```

A.3.2 OLMO Prompt Template (0–10 Scale)

```
You are an expert evaluator for mathematical and logical explanations. Given the following question and ra

IMPORTANT: A score of **10 always means BEST** and **0 always means WORST**, even for attributes like Hallu

Please return your response in two Python dictionaries:

- One called `scores` where each key is the name of the attribute and the value is the score (a float betwee
- One called `explanations` where each key is the name of the attribute and the value is your explanation (1

Please output the Python dictionaries as plain text only|do not include code blocks, markdown, or any extra

Here are the attributes and their definitions:

Faithfulness:
Is the rationale supported by the model's actual computation or the provided evidence?

Hallucination:
Does the rationale introduce information not present in the source/context?

Repetition:
Does the rationale unnecessarily repeat points or phrases?

Informativeness:
Does the rationale add meaningful, relevant details?

Plausibility:
Does the rationale \sound right" or seem believable, regardless of truth?

Self-Consistency:
The rationale does not contain steps that contradict each other; all reasoning is logically aligned interna

Source Consistency:
The rationale does not contradict the given context or information in the problem statement.

Grammar:
Is the rationale well-written, clear, and free of grammatical mistakes?

Arithmetic Accuracy:
Are any calculations in the rationale correct?

Conciseness:
Is it as short as possible, without losing information? Especially if length is a concern.

Coverage/Completeness:
Does it explain all necessary steps/evidence?

Correctness:
Are all steps and answers objectively correct?

Example output format:

scores = {
    "Faithfulness": 9.5,
    "Hallucination": 6.8,
    "Repetition": 8.9,
    ...
}
```

```
explanations = {
   "Faithfulness": "The rationale closely follows logical steps derived from the question.",
   "Hallucination": "Some external information or assumptions were introduced. For example, ...",
   "Repetition": "The rationale is does not repeat it self with similar points at different steps.",
    ...
}

Math/Logic Question:
{question}

Rationale:
{rationale}
```

A.4 LLM Judges Results

A.4.1 OpenAI GPT-4o

1. **Chatbot Arena**

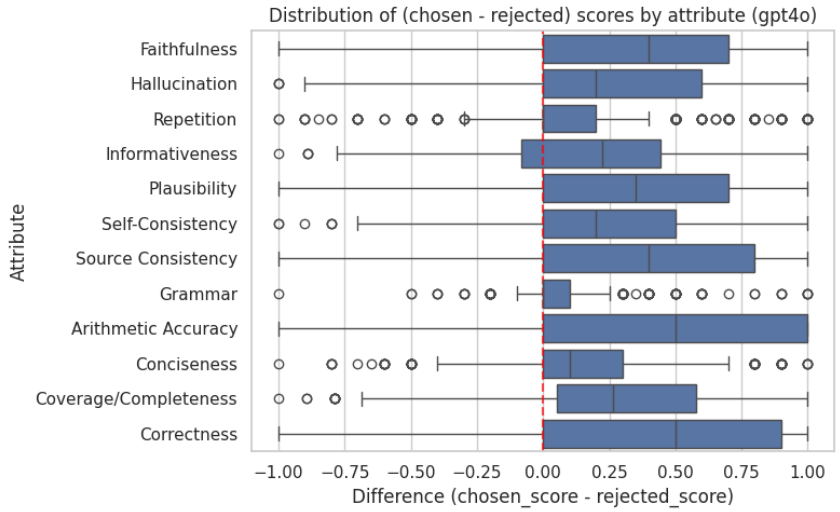


Figure 12: Distribution of the difference between chosen and rejected scores by attribute in Chatbot Arena (GPT-4o). Boxplots summarize the (chosen – rejected) difference for each attribute.

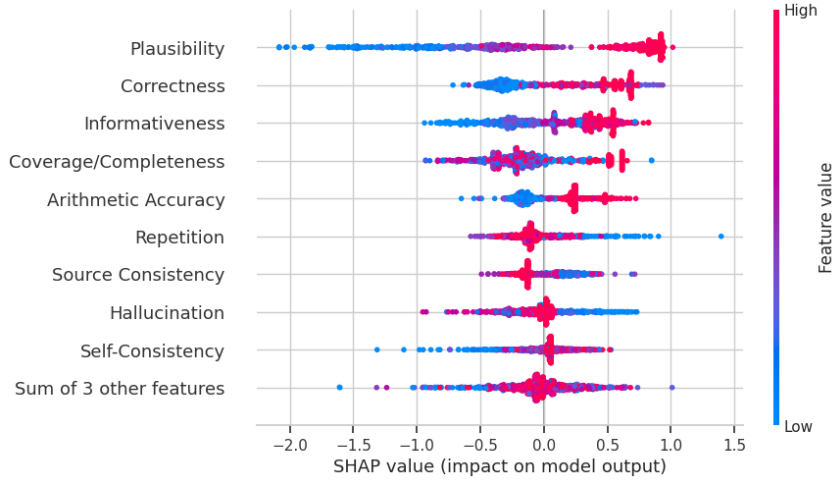


Figure 13: SHAP beeswarm plot for Chatbot Arena (GPT-4o). Visualizes the distribution and direction of SHAP values for each attribute.

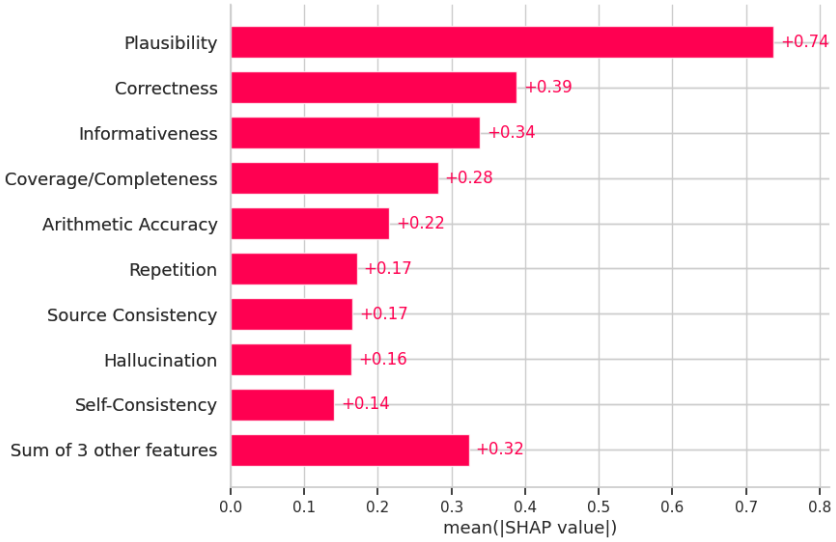


Figure 14: Mean absolute SHAP value plot for Chatbot Arena (GPT-4o). Shows the mean importance of each attribute in the model.

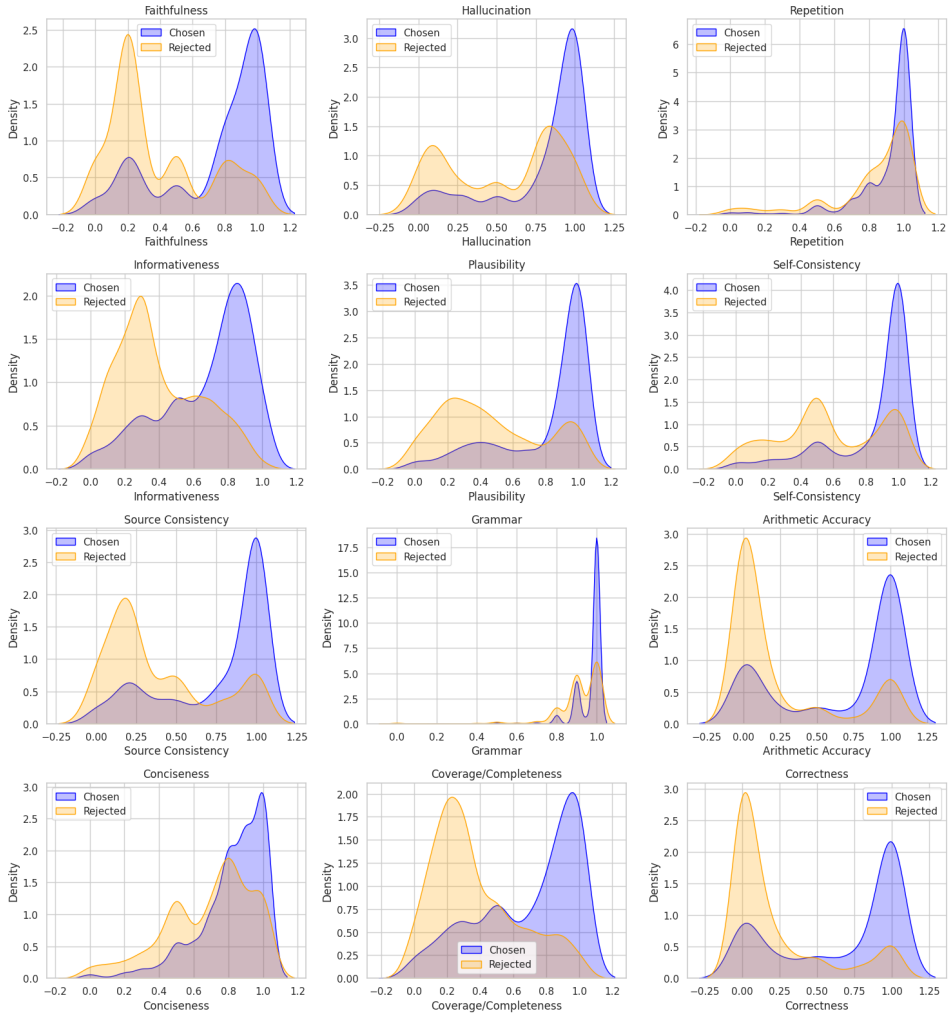


Figure 15: Distribution of attribute values for chosen vs. rejected rationales in Chatbot Arena (GPT-4o). Each subplot shows the density of scores for each attribute.

2. **Mt Bench**

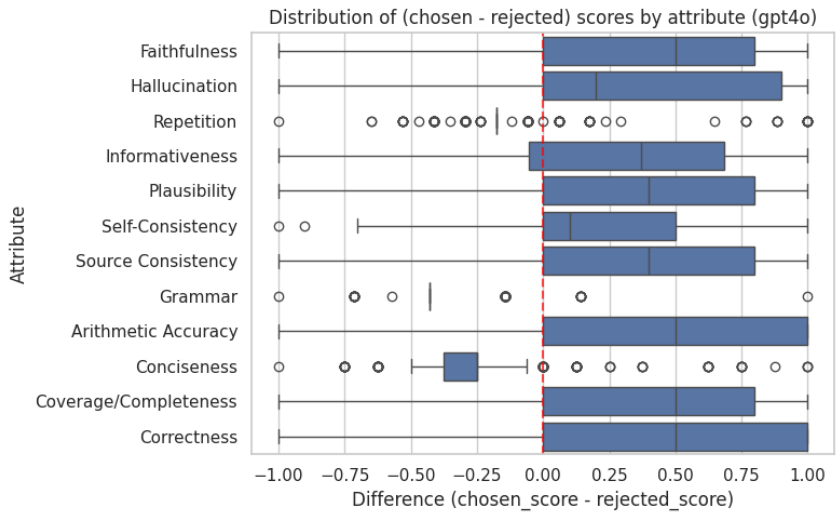


Figure 16: Distribution of the difference between chosen and rejected scores by attribute in MT Bench (GPT-4o). Boxplots summarize the (chosen – rejected) difference for each attribute.

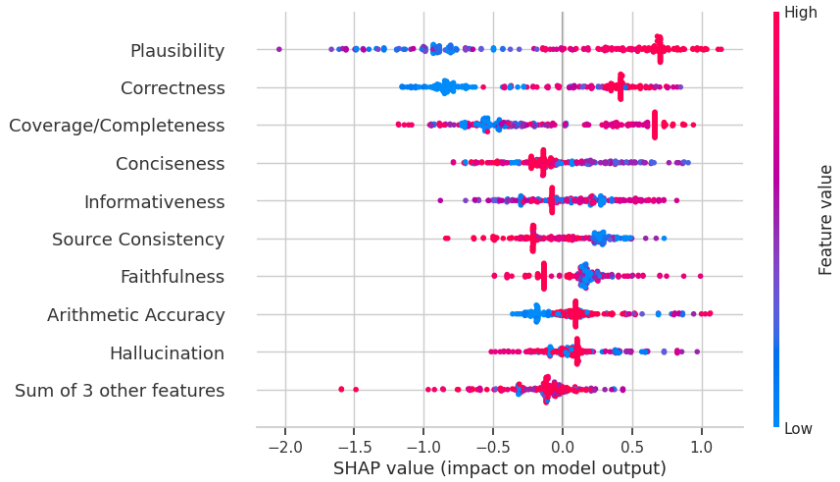


Figure 17: SHAP beeswarm plot for MT Bench (GPT-4o). Visualizes the distribution and direction of SHAP values for each attribute.

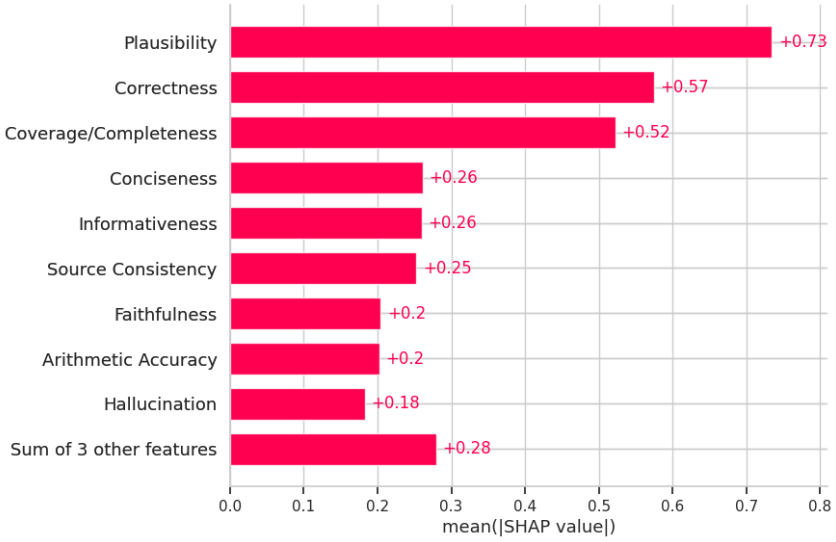


Figure 18: Mean absolute SHAP value plot for MT Bench (GPT-4o). Shows the mean importance of each attribute in the model.

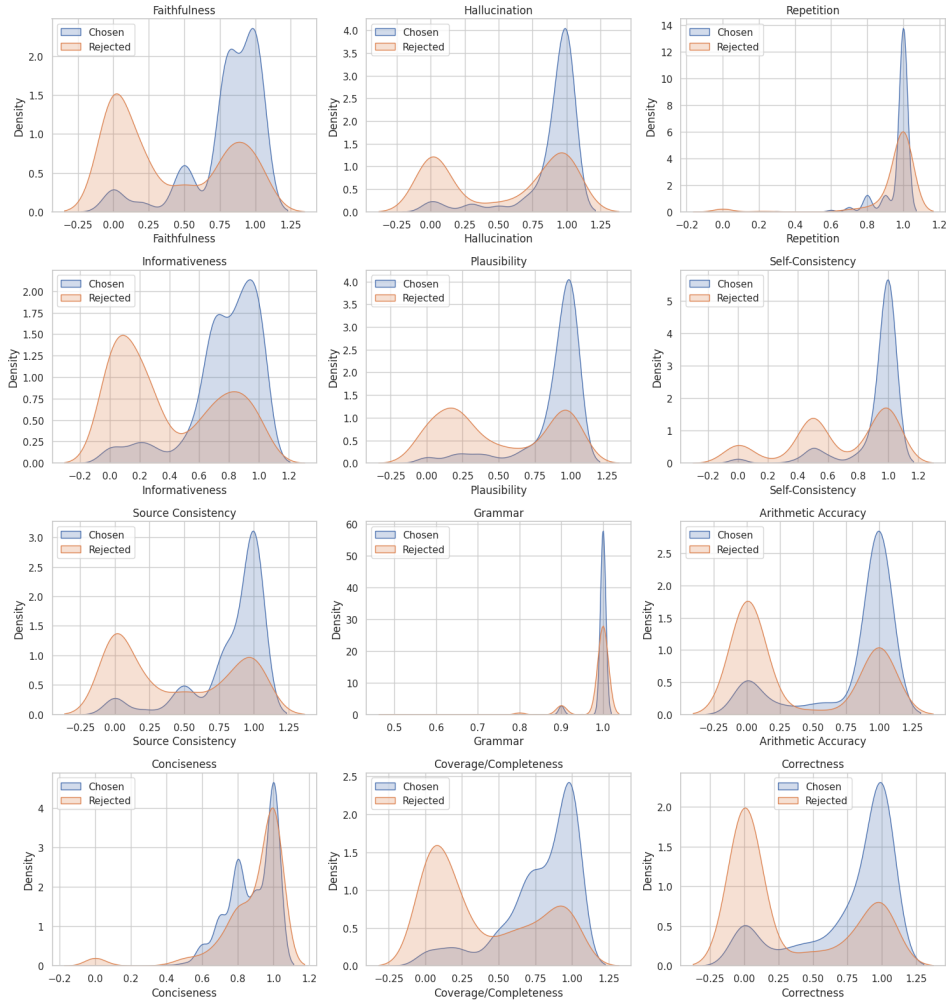


Figure 19: Distribution of attribute values for chosen vs. rejected rationales in MT Bench (GPT-4o). Each subplot shows the density of scores for each attribute.

A.4.2 Google Gemini-2.5-Flash

1. **Chatbot Arena**

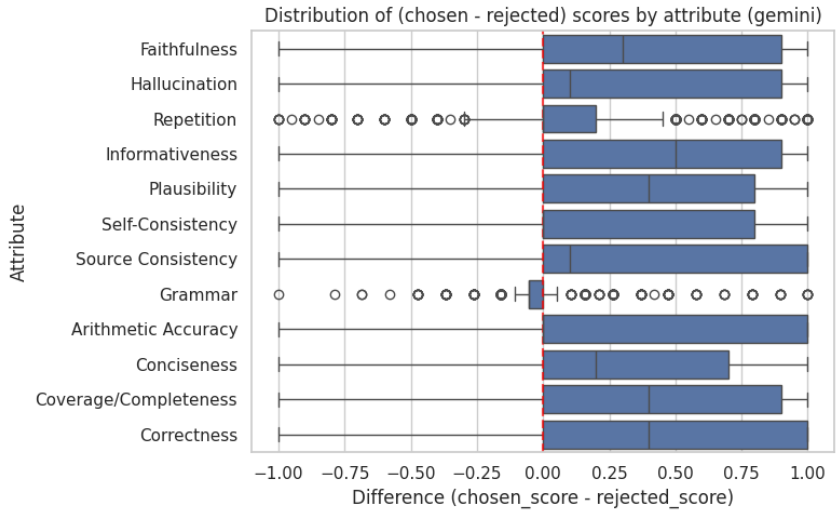


Figure 20: Distribution of the difference between chosen and rejected scores by attribute in Chatbot Arena (Gemini 2.5-Flash). Boxplots summarize the (chosen – rejected) difference for each attribute.

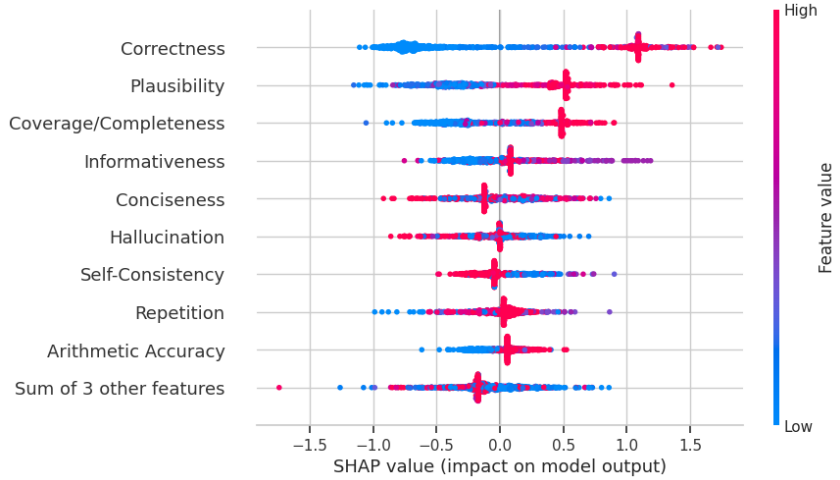


Figure 21: SHAP beeswarm plot for Chatbot Arena (Gemini 2.5-Flash). Visualizes the distribution and direction of SHAP values for each attribute.

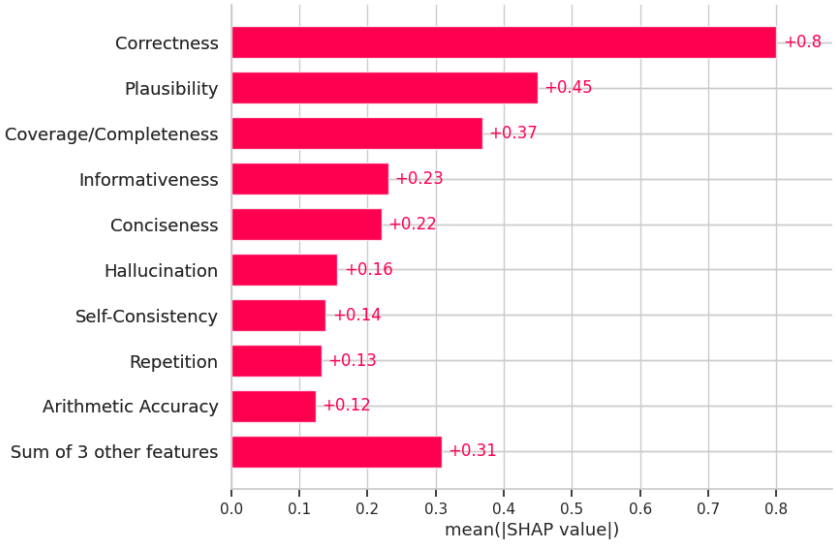


Figure 22: Mean absolute SHAP value plot for Chatbot Arena (Gemini 2.5-Flash). Shows the mean importance of each attribute in the model.

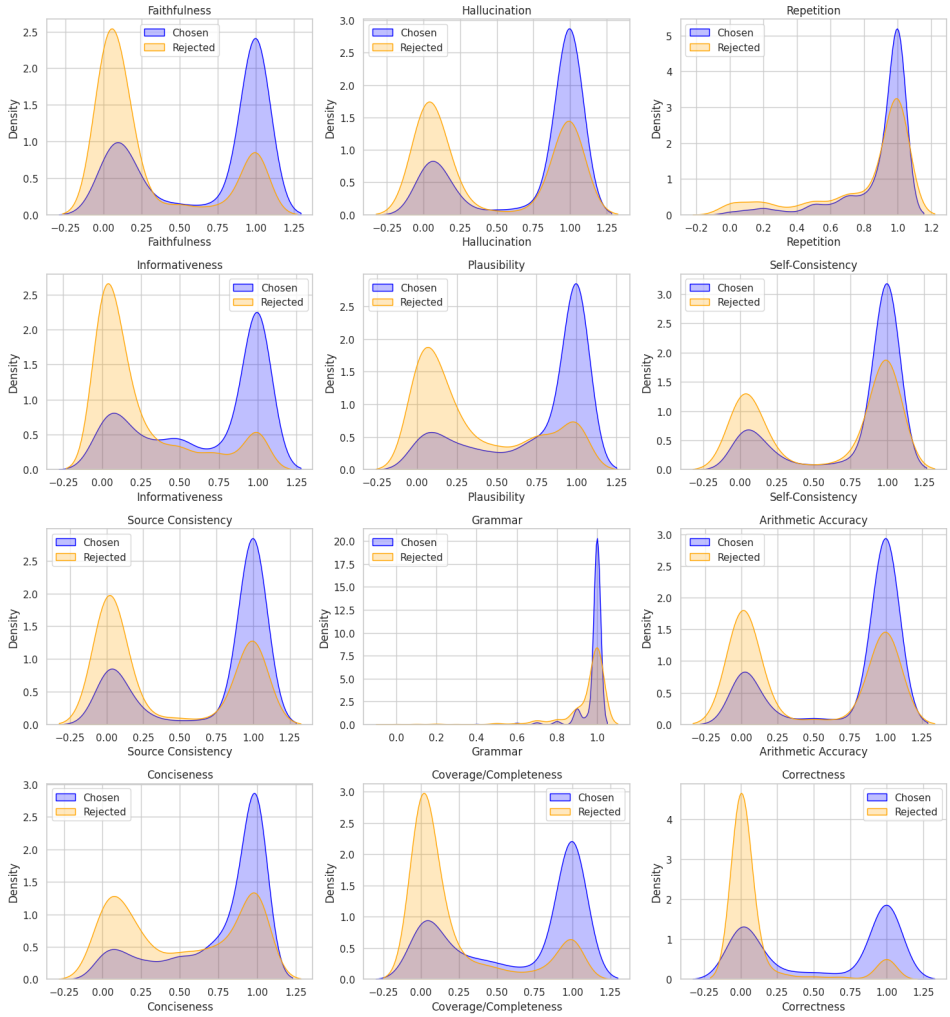


Figure 23: Distribution of attribute values for chosen vs. rejected rationales in Chatbot Arena (Gemini 2.5-Flash). Each subplot shows the density of scores for each attribute.

2. **Mt Bench**

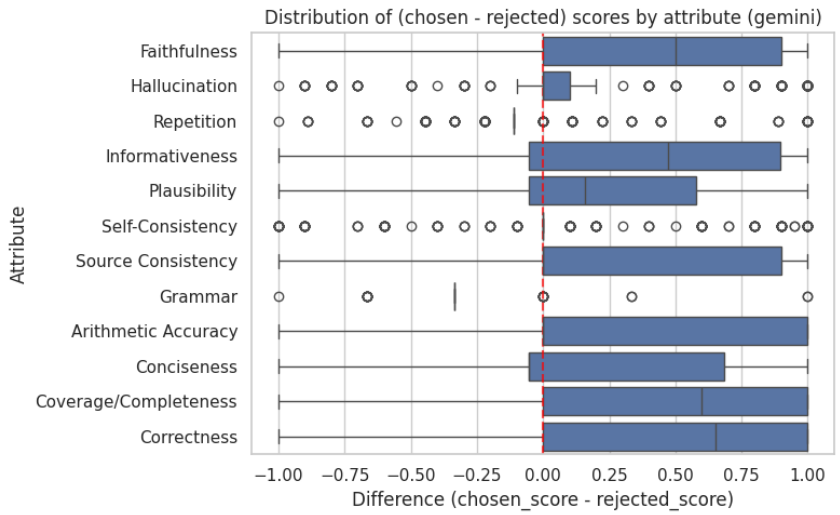


Figure 24: Distribution of the difference between chosen and rejected scores by attribute in MT Bench (Gemini 2.5-Flash). Boxplots summarize the (chosen – rejected) difference for each attribute.

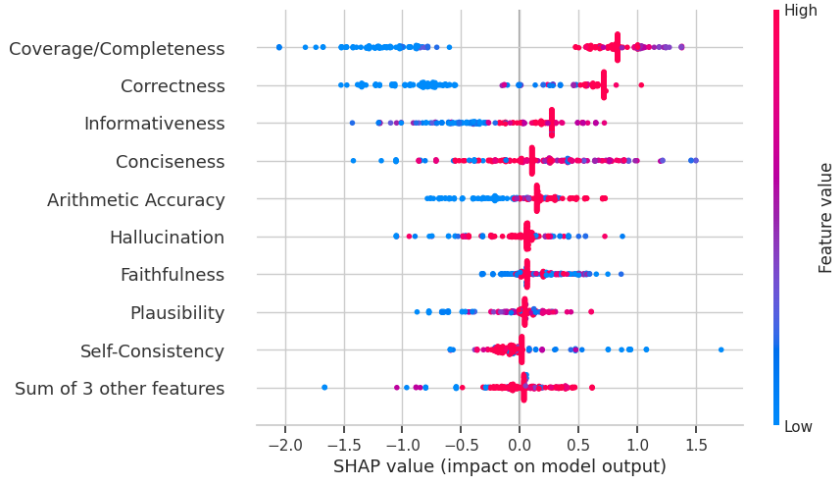


Figure 25: SHAP beeswarm plot for MT Bench (Gemini 2.5-Flash). Visualizes the distribution and direction of SHAP values for each attribute.

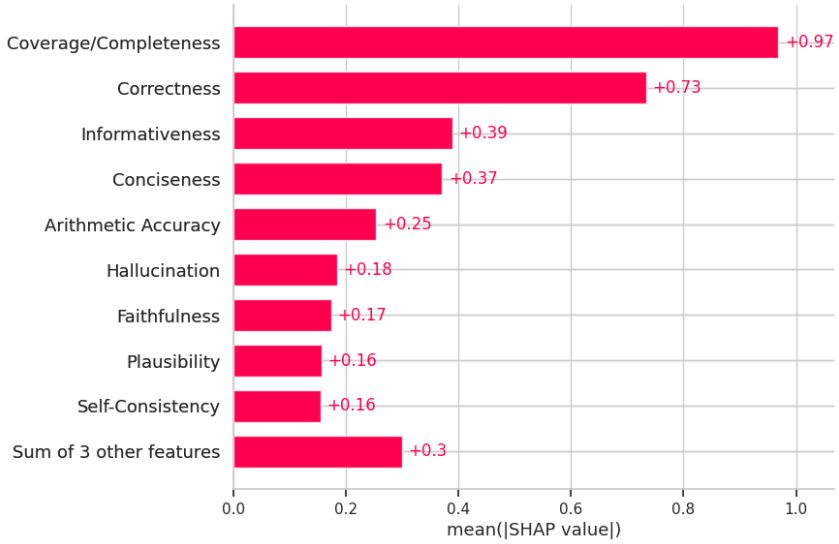


Figure 26: Mean absolute SHAP value plot for MT Bench (Gemini 2.5-Flash). Shows the mean importance of each attribute in the model.

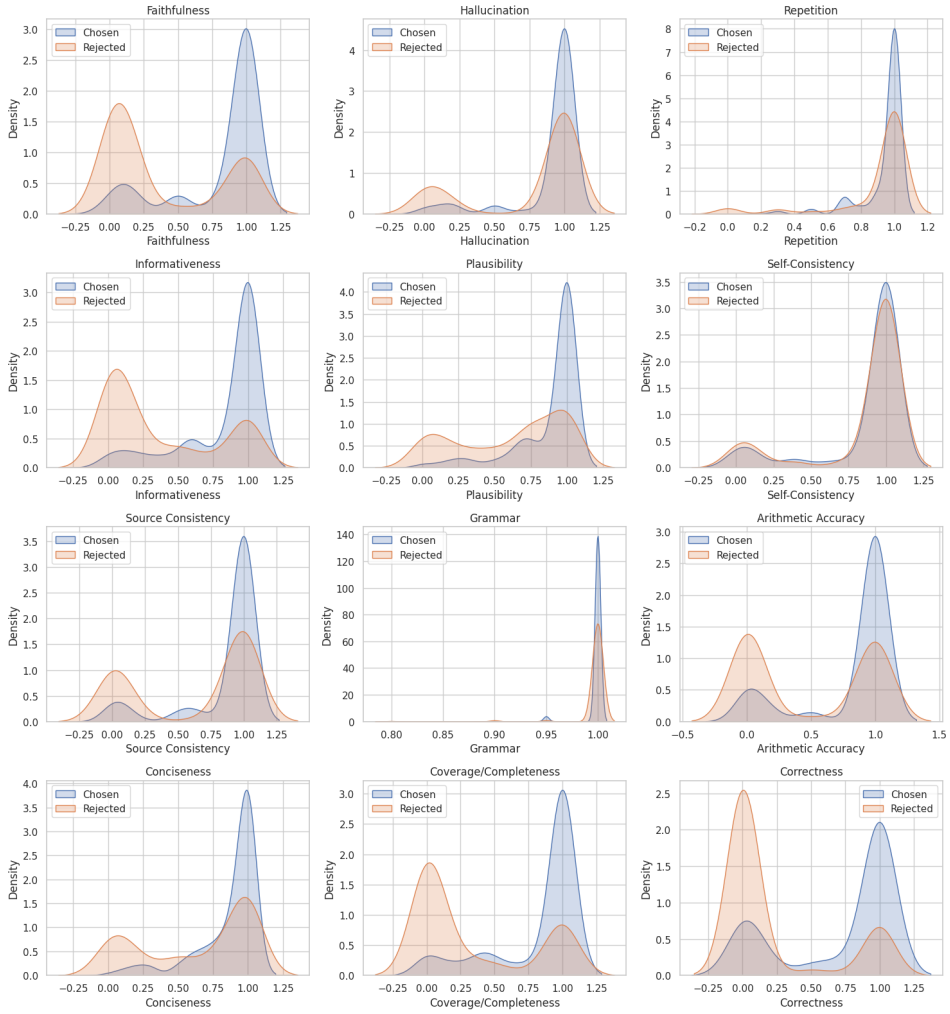


Figure 27: Distribution of attribute values for chosen vs. rejected rationales in MT Bench (Gemini 2.5-Flash). Each subplot shows the density of scores for each attribute.

A.4.3 OLMo 2-32b

1. **Chatbot Arena**

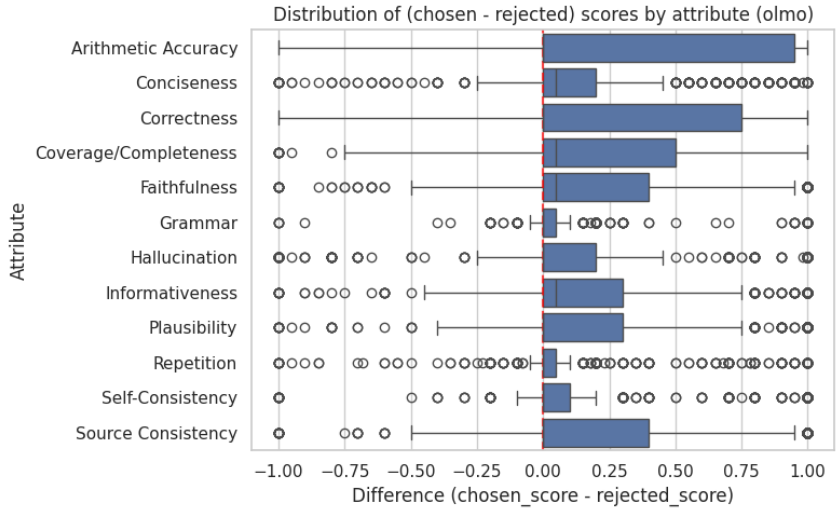


Figure 28: Distribution of the difference between chosen and rejected scores by attribute in Chatbot Arena (OLMo 2-32b). Boxplots summarize the (chosen – rejected) difference for each attribute.

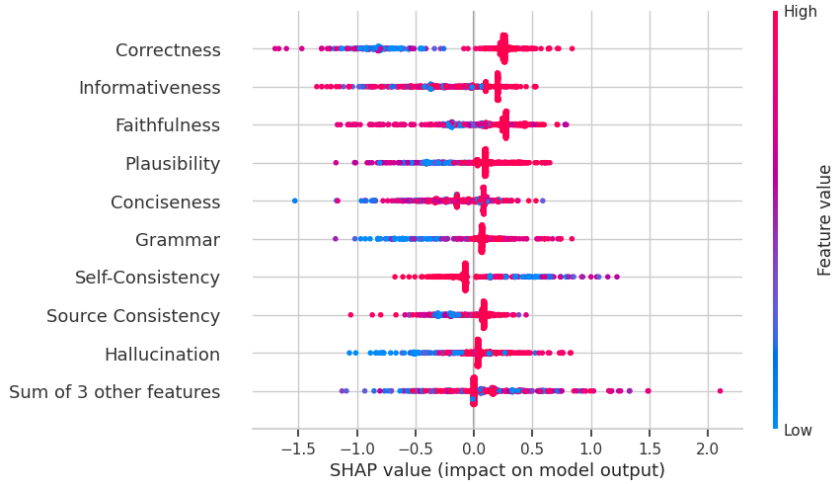


Figure 29: SHAP beeswarm plot for Chatbot Arena (OLMo 2-32b). Visualizes the distribution and direction of SHAP values for each attribute.

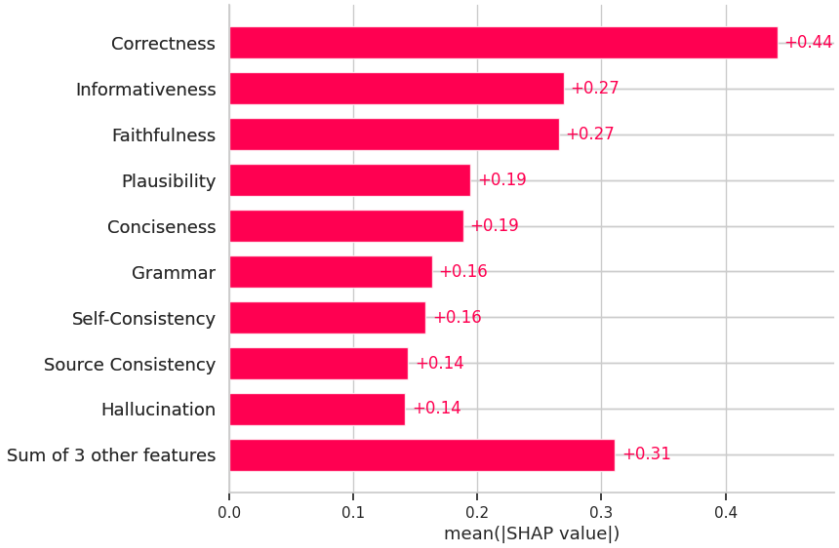


Figure 30: Mean absolute SHAP value plot for Chatbot Arena (OLMo 2-32b). Shows the mean importance of each attribute in the model.

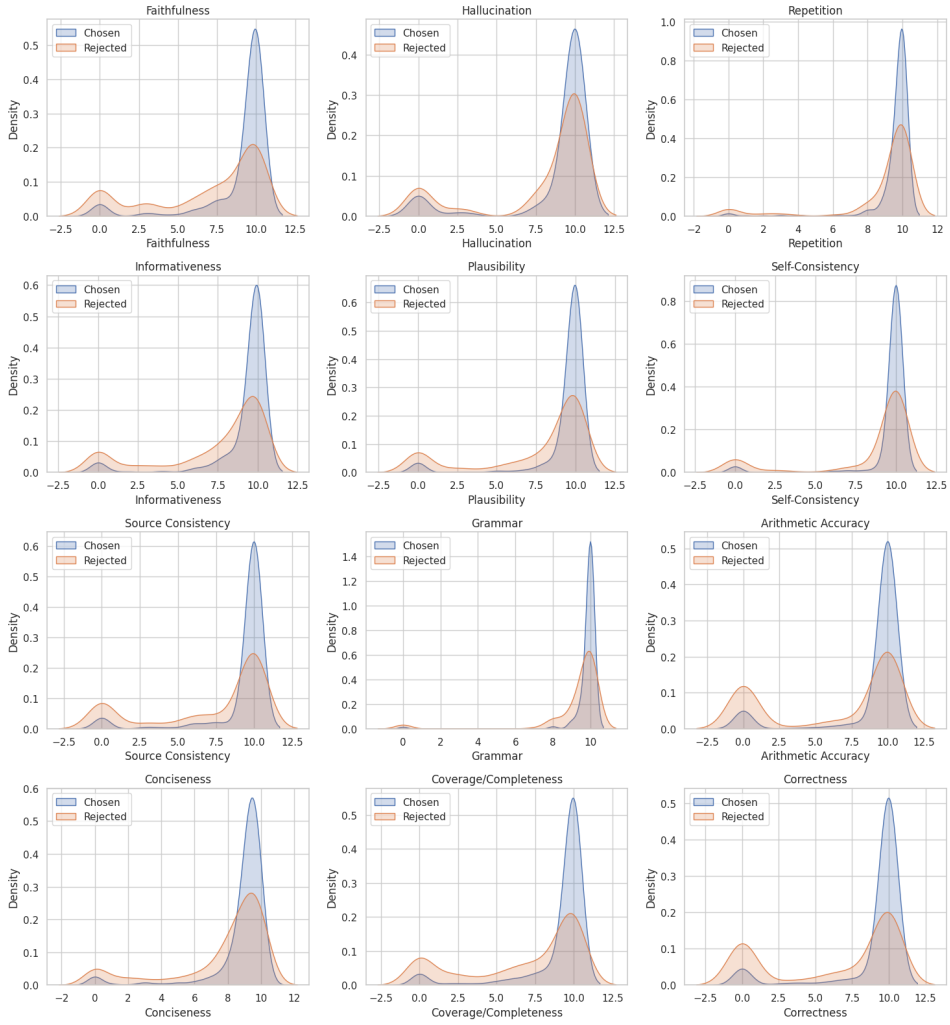


Figure 31: Distribution of attribute values for chosen vs. rejected rationales in Chatbot Arena (OLMo 2-32b). Each subplot shows the density of scores for each attribute.

2. **Mt Bench**

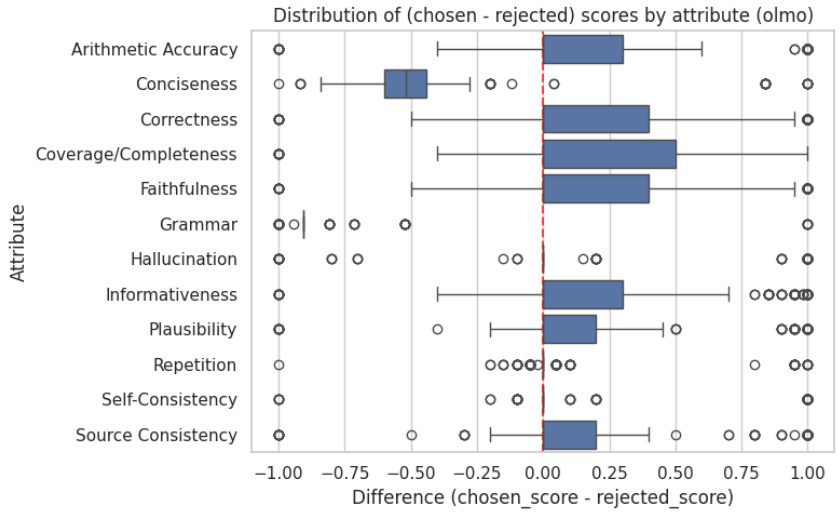


Figure 32: Distribution of the difference between chosen and rejected scores by attribute in MT Bench (OLMo 2-32b). Boxplots summarize the (chosen – rejected) difference for each attribute.

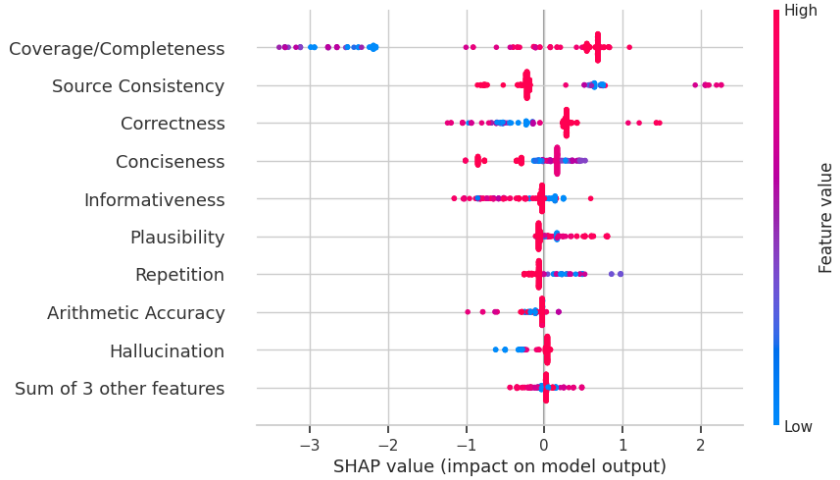


Figure 33: SHAP beeswarm plot for MT Bench (OLMo 2-32b). Visualizes the distribution and direction of SHAP values for each attribute.

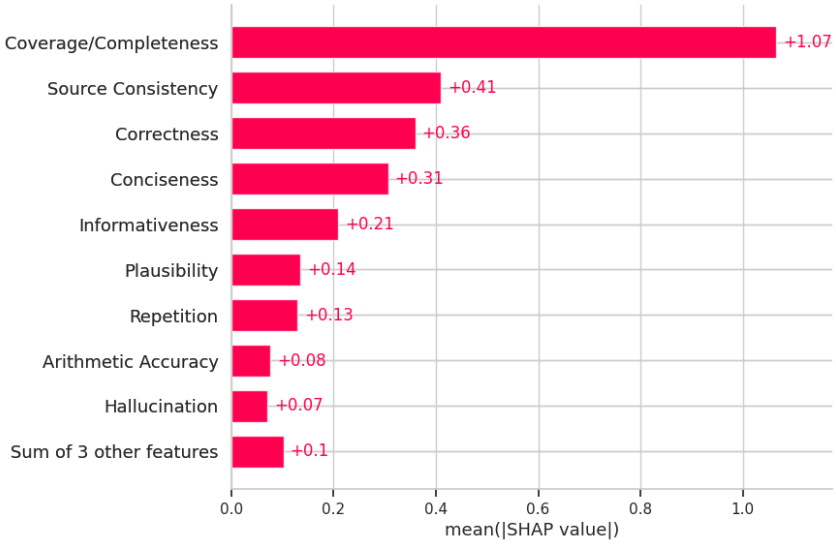


Figure 34: Mean absolute SHAP value plot for MT Bench (OLMo 2-32b). Shows the mean importance of each attribute in the model.

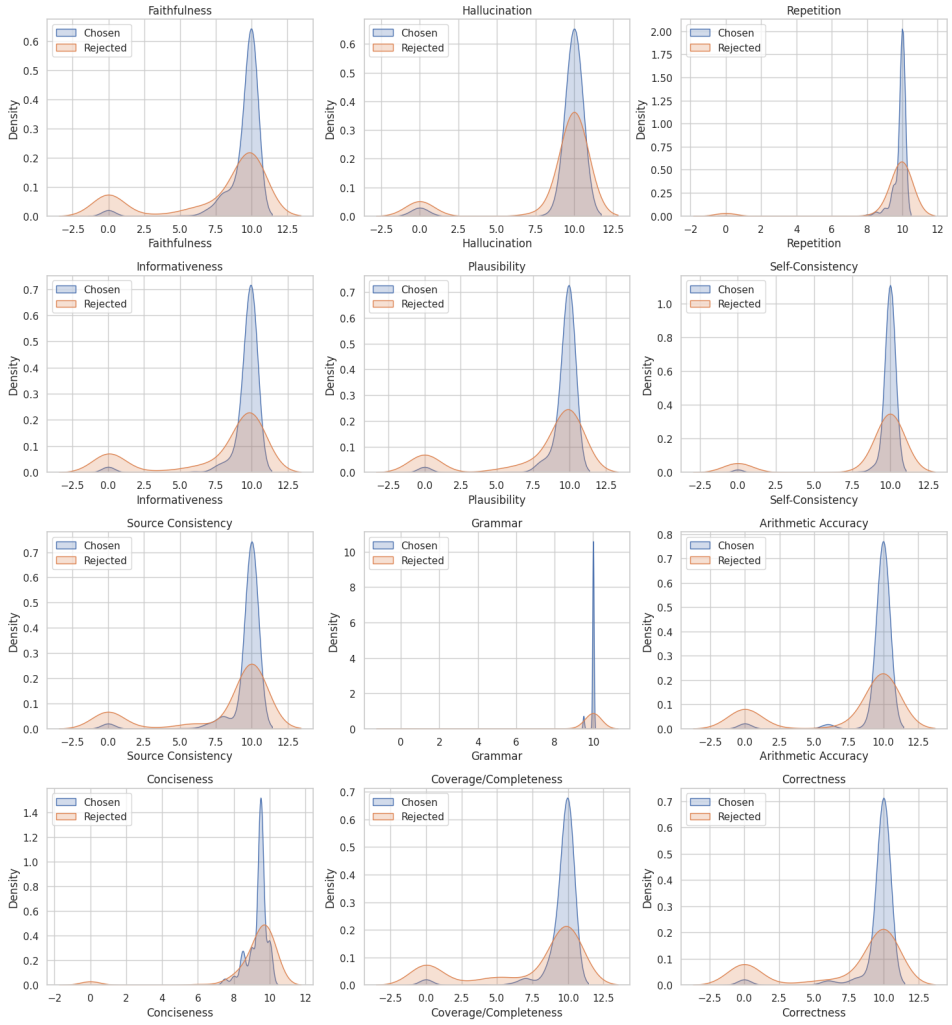


Figure 35: Distribution of attribute values for chosen vs. rejected rationales in MT Bench (OLMo 2-32b). Each subplot shows the density of scores for each attribute.

A.5 Human Annotators Results

The human evaluation results reported in this section are based on the average scores assigned by three independent annotators for each question. For every rationale, we take the mean of the three annotators' scores to obtain a single human score per attribute. This approach helps mitigate individual annotator bias and provides a more robust measure of human preference.

A.5.1 Chatbot Arena

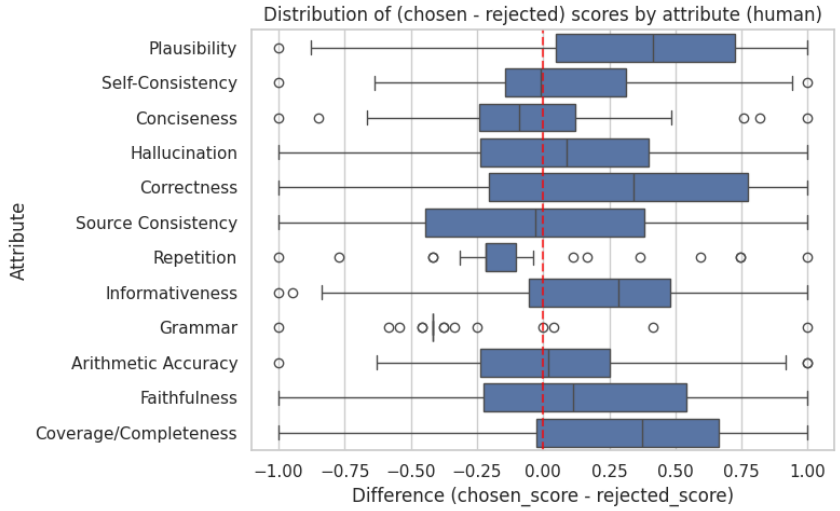


Figure 36: Distribution of the difference between chosen and rejected scores by attribute in Chatbot Arena (Human Annotators). Boxplots summarize the (chosen – rejected) difference for each attribute.

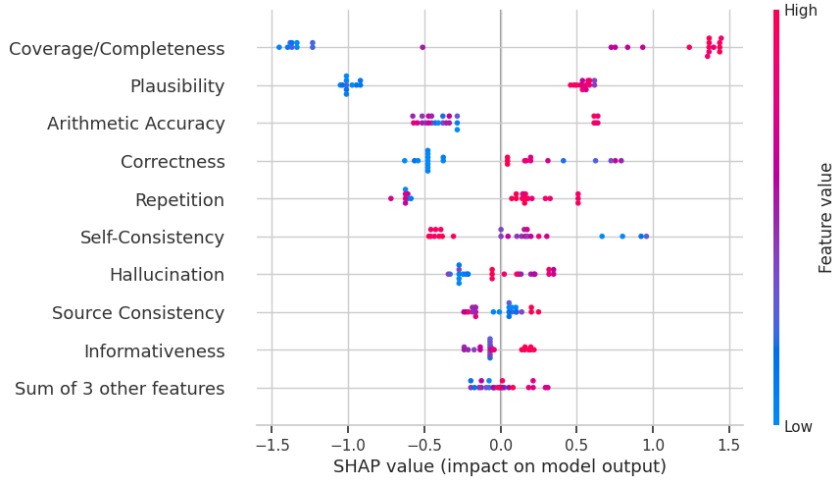


Figure 37: SHAP beeswarm plot for Chatbot Arena (Human Annotators). Visualizes the distribution and direction of SHAP values for each attribute.

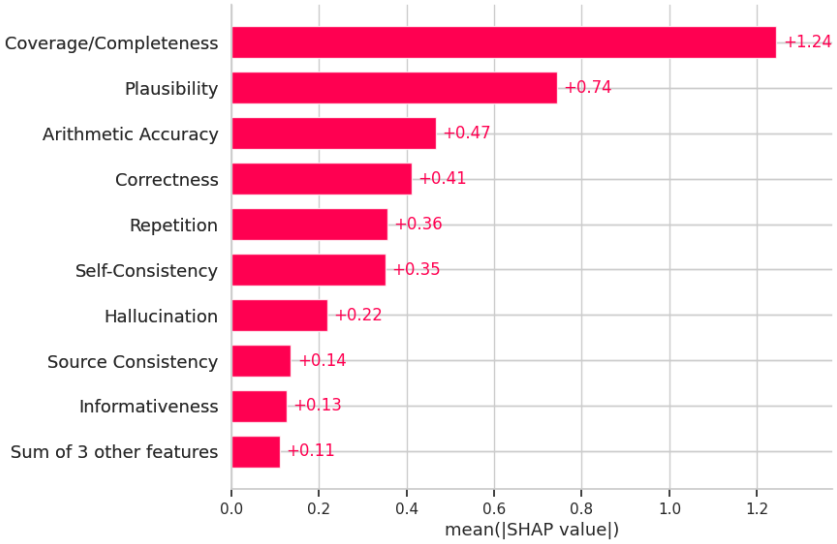


Figure 38: Mean absolute SHAP value plot for Chatbot Arena (Human Annotators). Shows the mean importance of each attribute in the model.

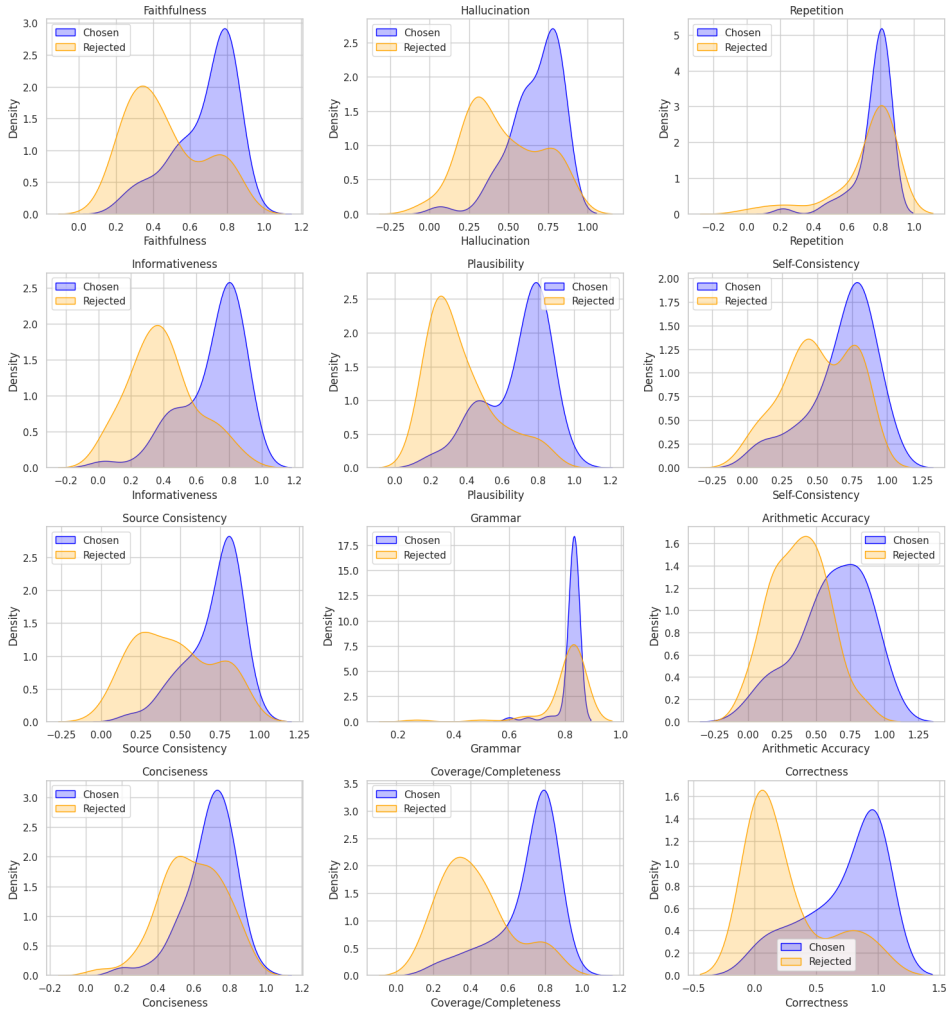


Figure 39: Distribution of attribute values for chosen vs. rejected rationales in Chatbot Arena (Human Annotators). Each subplot shows the density of scores for each attribute.

A.5.2 Mt Bench

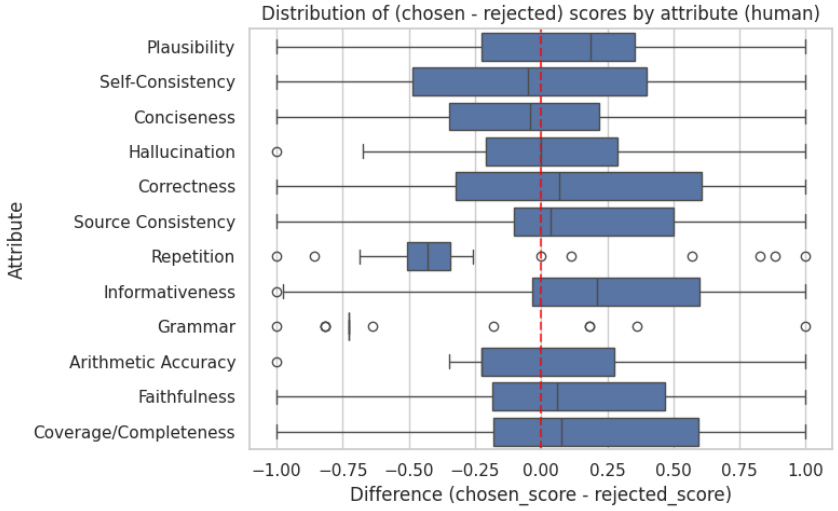


Figure 40: Distribution of the difference between chosen and rejected scores by attribute in MT Bench (Human Annotators). Boxplots summarize the (chosen – rejected) difference for each attribute.

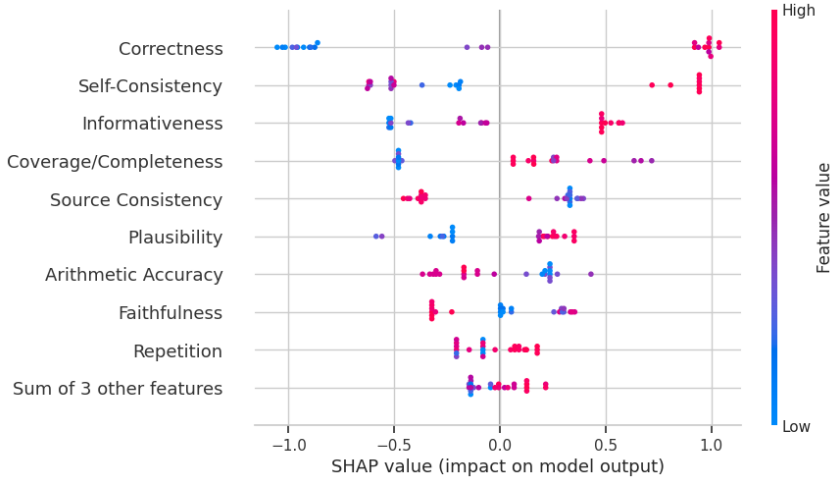


Figure 41: SHAP beeswarm plot for MT Bench (Human Annotators). Visualizes the distribution and direction of SHAP values for each attribute.

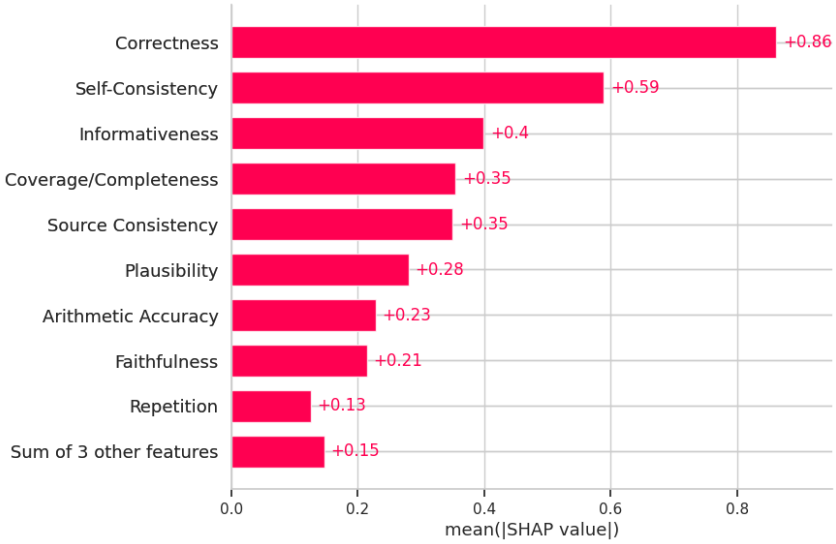


Figure 42: Mean absolute SHAP value plot for MT Bench (Human Annotators). Shows the mean importance of each attribute in the model.

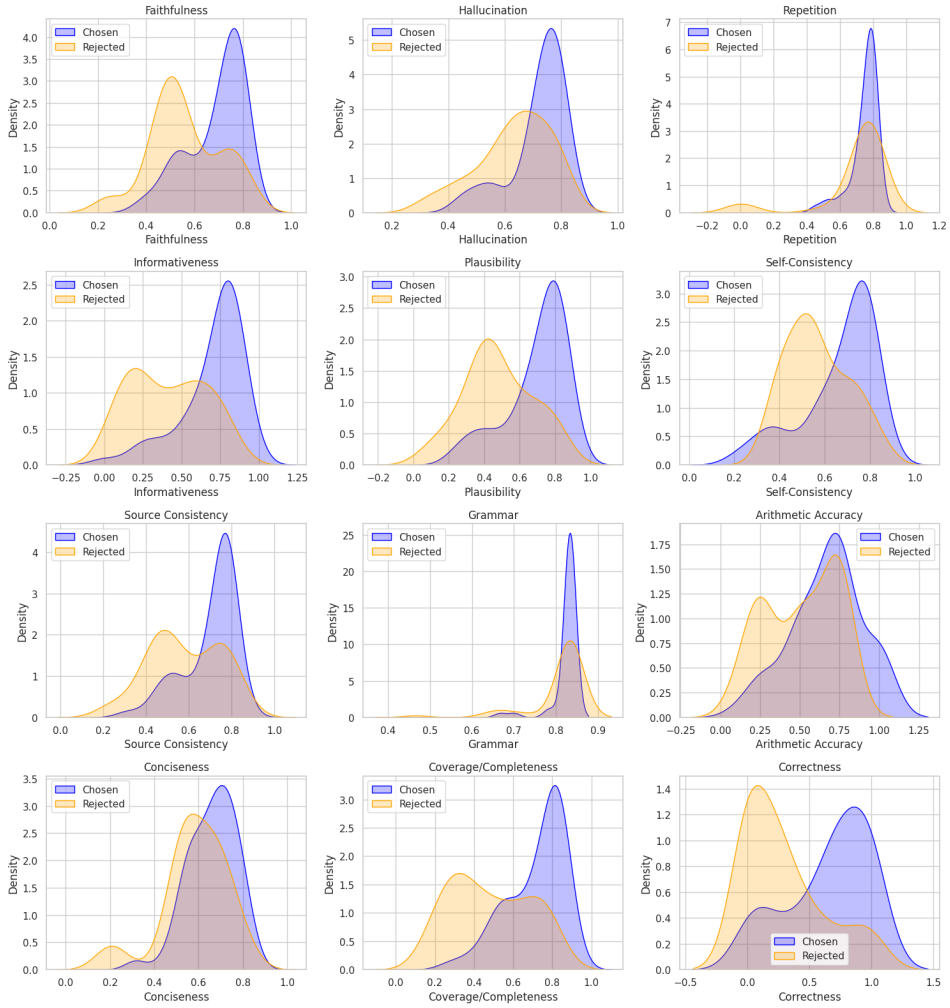


Figure 43: Distribution of attribute values for chosen vs. rejected rationales in MT Bench (Human Annotators). Each subplot shows the density of scores for each attribute.

A.6 ELO ranking results

A.6.1 LLM Judges

1. **Chatbot Arena**

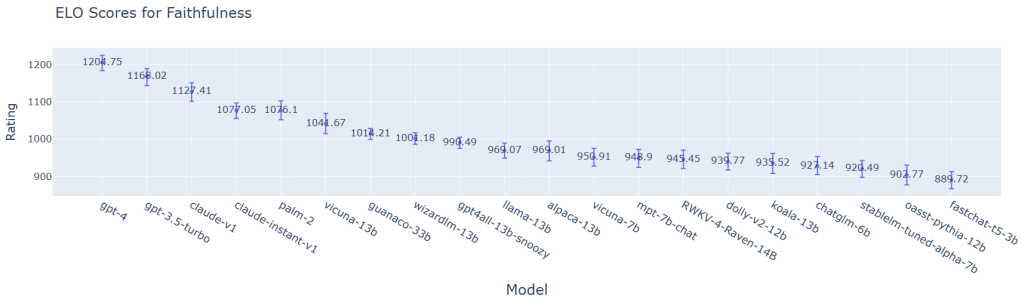


Figure 44: ELO Scores for Faithfulness across all models in Chatbot Arena, scored by the mean score of three LLMs.

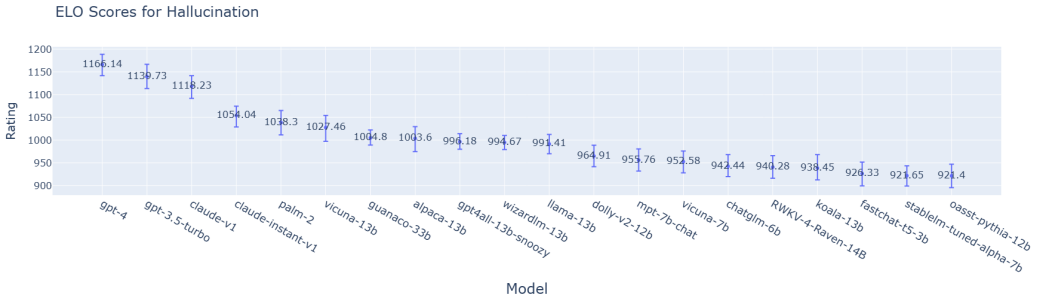


Figure 45: ELO Scores for Hallucination across all models in Chatbot Arena, scored by the mean score of three LLMs.

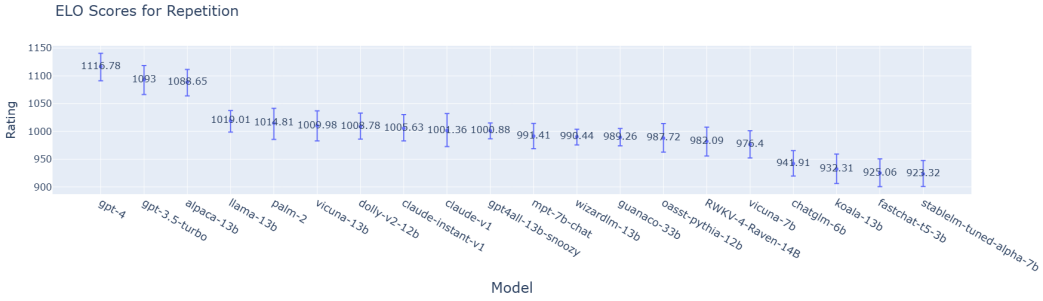


Figure 46: ELO Scores for Repetition across all models in Chatbot Arena, scored by the mean score of three LLMs.

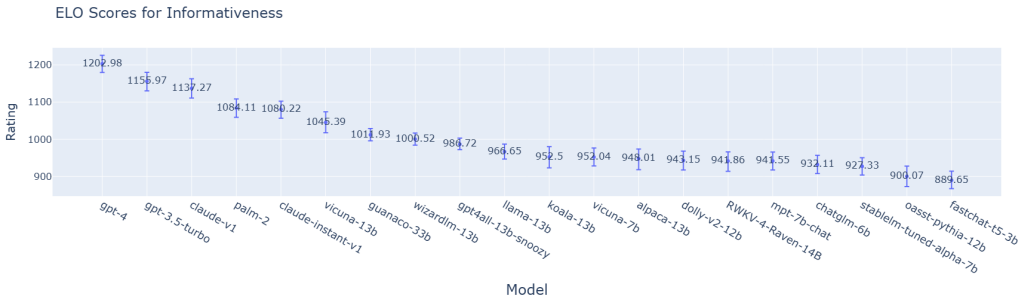


Figure 47: ELO Scores for Informativeness across all models in Chatbot Arena, scored by the mean score of three LLMs.

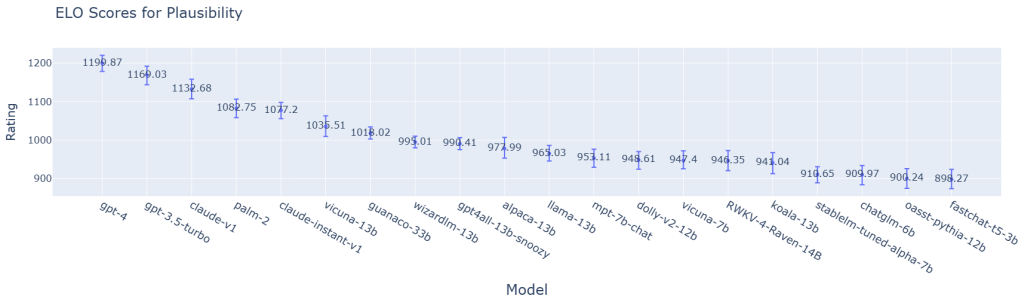


Figure 48: ELO Scores for Plausibility across all models in Chatbot Arena, scored by the mean score of three LLMs.

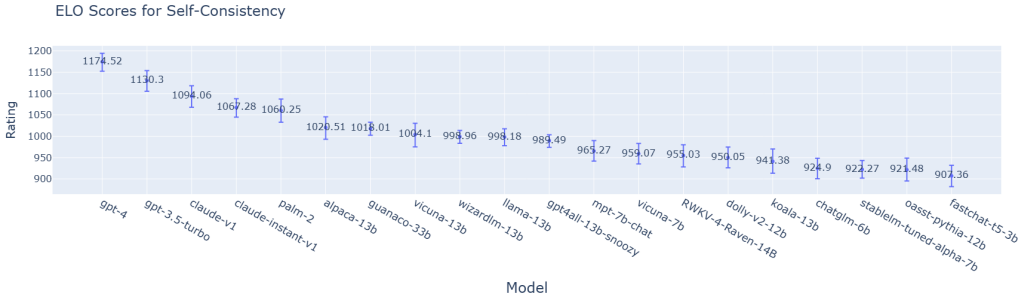


Figure 49: ELO Scores for Self-Consistency across all models in Chatbot Arena, scored by the mean score of three LLMs.

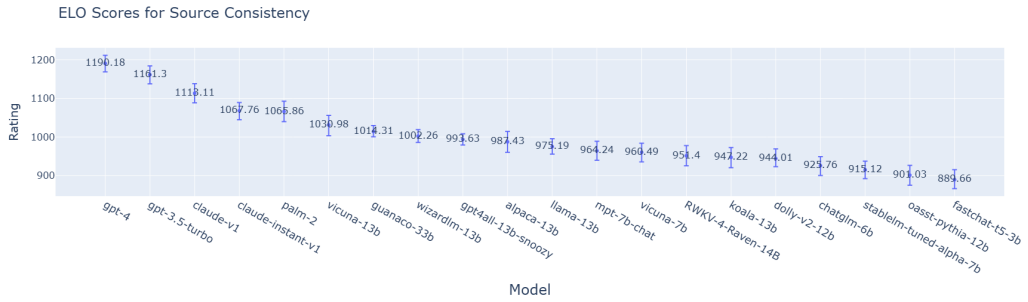


Figure 50: ELO Scores for Source Consistency across all models in Chatbot Arena, scored by the mean score of three LLMs.

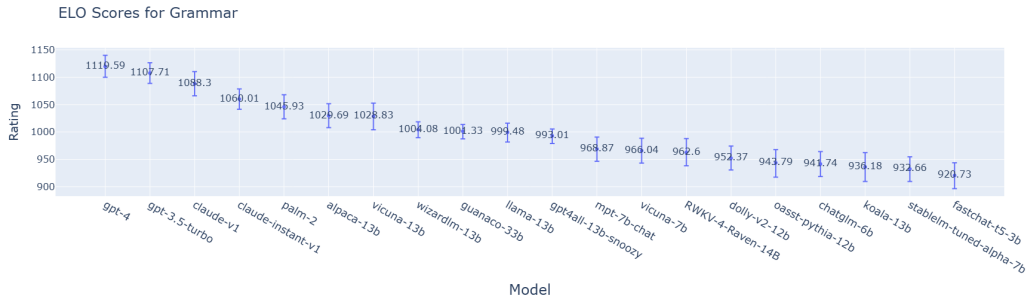


Figure 51: ELO Scores for Grammar across all models in Chatbot Arena, scored by the mean score of three LLMs.

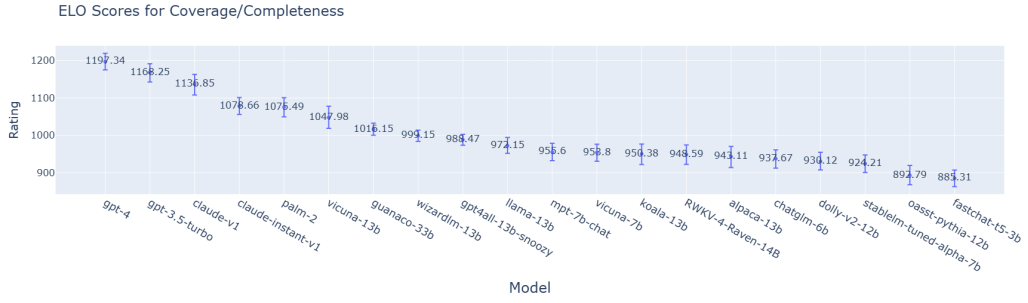


Figure 52: ELO Scores for Coverage/Completeness across all models in Chatbot Arena, scored by the mean score of three LLMs.

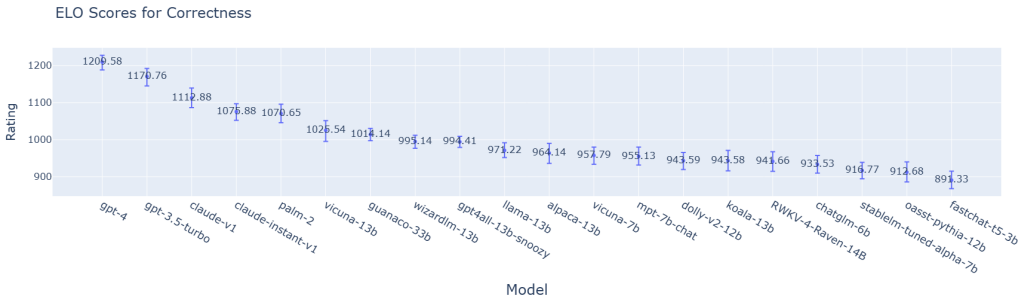


Figure 53: ELO Scores for Correctness across all models in Chatbot Arena, scored by the mean score of three LLMs.

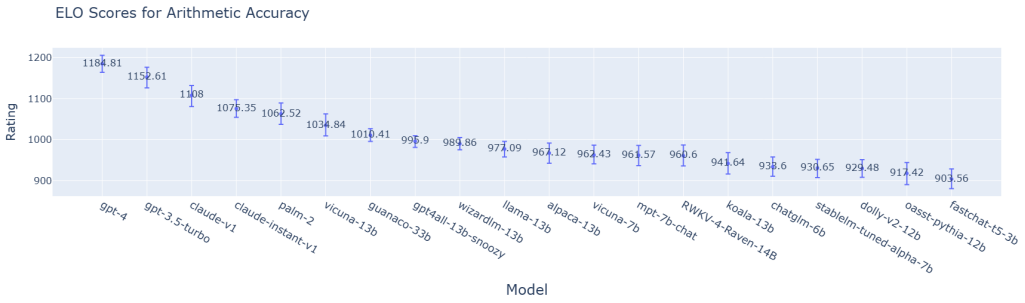


Figure 54: ELO Scores for Arithmetic Accuracy across all models in Chatbot Arena, scored by the mean score of three LLMs.

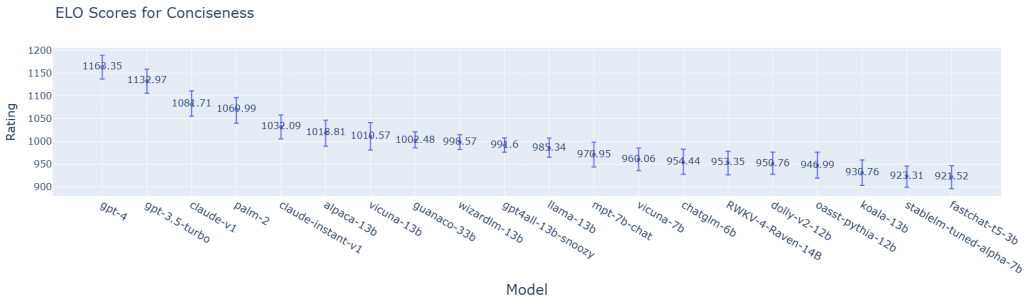


Figure 55: ELO Scores for Conciseness across all models in Chatbot Arena, scored by the mean score of three LLMs.

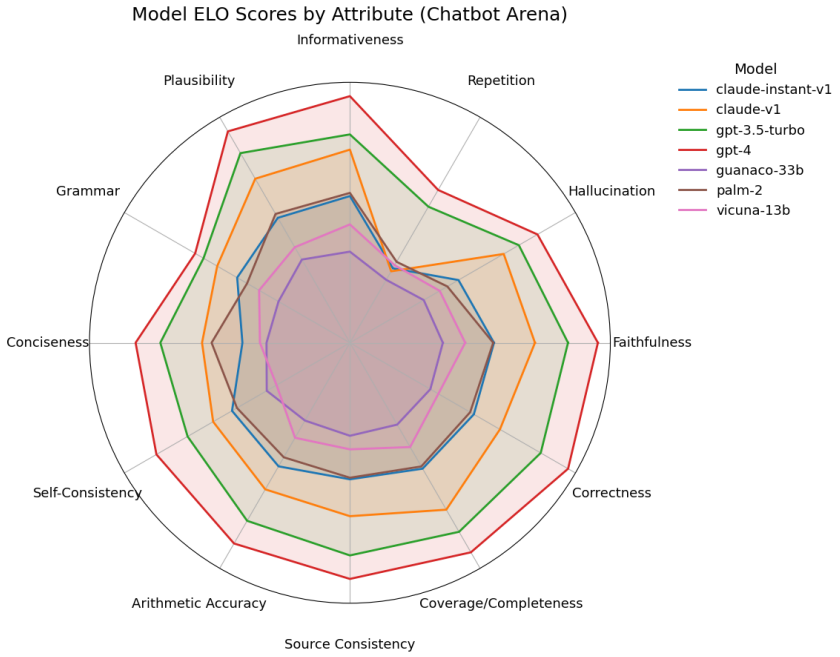


Figure 56: Radar chart of model ELO scores by attribute in Chatbot Arena, computed as the mean of the three LLM judges.

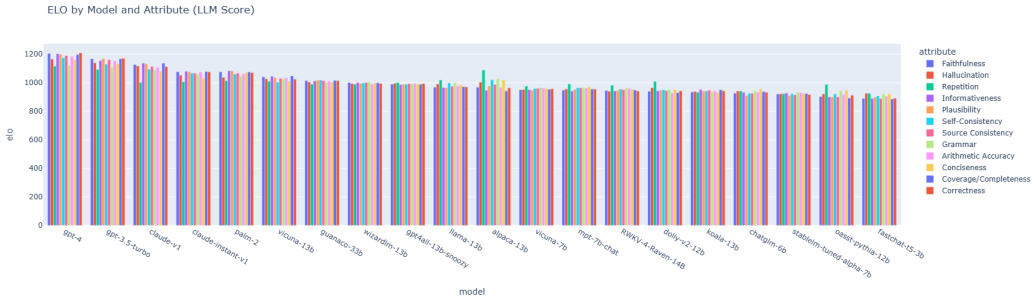


Figure 57: Bar plot of model ELO scores by model and attribute in Chatbot Arena, averaged across all LLM judges.

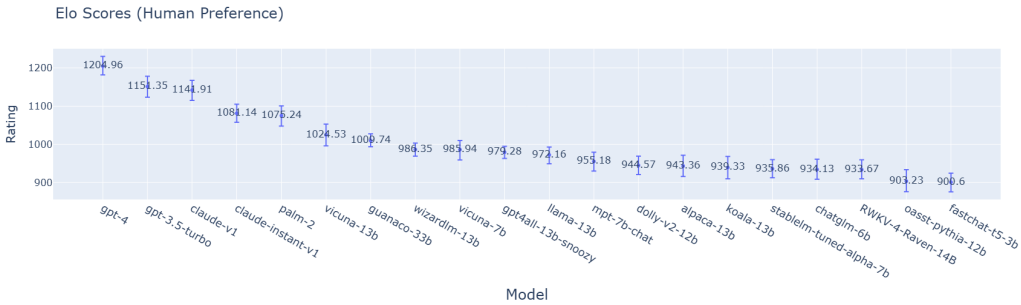


Figure 58: ELO scores for all models based on human preference labels in Chatbot Arena. Here, the score is computed directly from the human-chosen vs. rejected outcomes. Error bars indicate the confidence interval.

2. Mt Bench

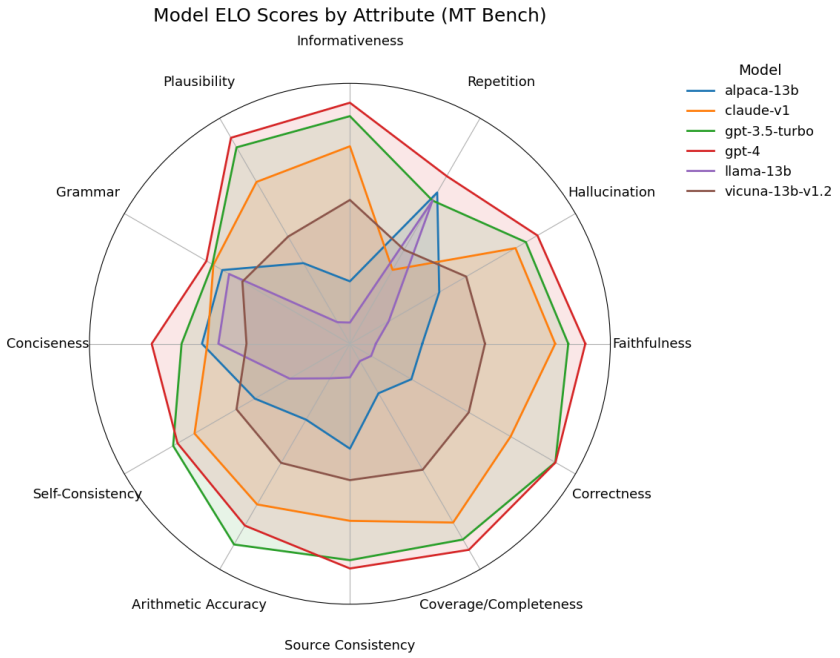


Figure 59: Radar chart comparing model ELO scores by attribute on MT Bench (LLM scores). Each axis represents one evaluation attribute, and each polygon represents a different model.

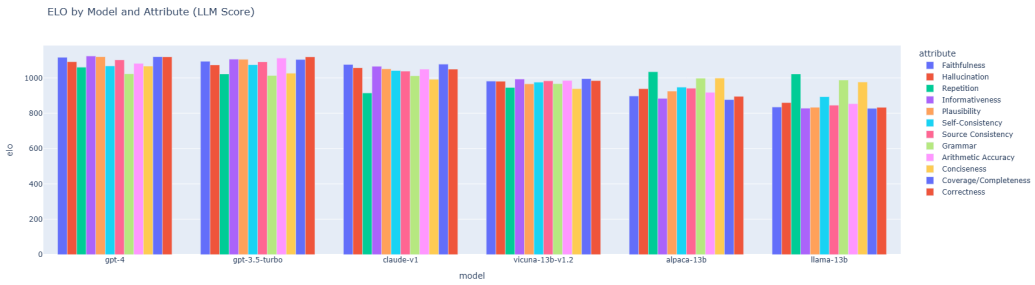


Figure 60: Bar chart showing ELO scores for each model and attribute on MT Bench (LLM scores). Each color indicates a different evaluation attribute.

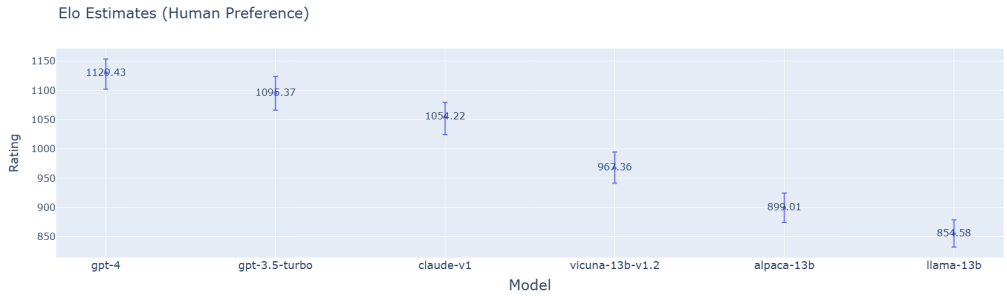


Figure 61: Bar plot of ELO estimates based on human preferences for each model on MT Bench. Error bars indicate uncertainty in ELO estimation.

A.6.2 Human Annotator

1. **Chatbot Arena**

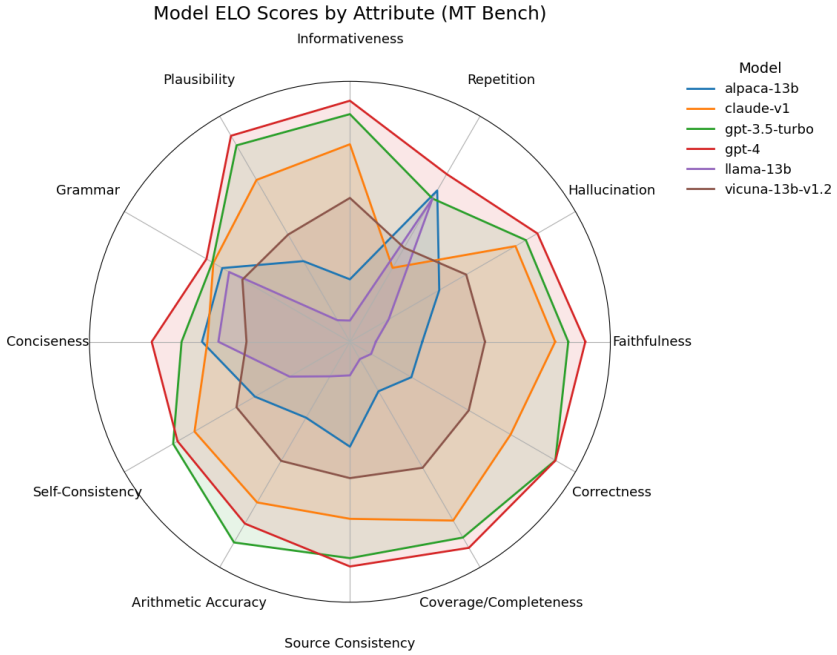


Figure 62: Radar chart comparing model ELO scores by attribute on MT Bench (LLM Judges). Each axis corresponds to a specific evaluation attribute, and each line represents a different model.

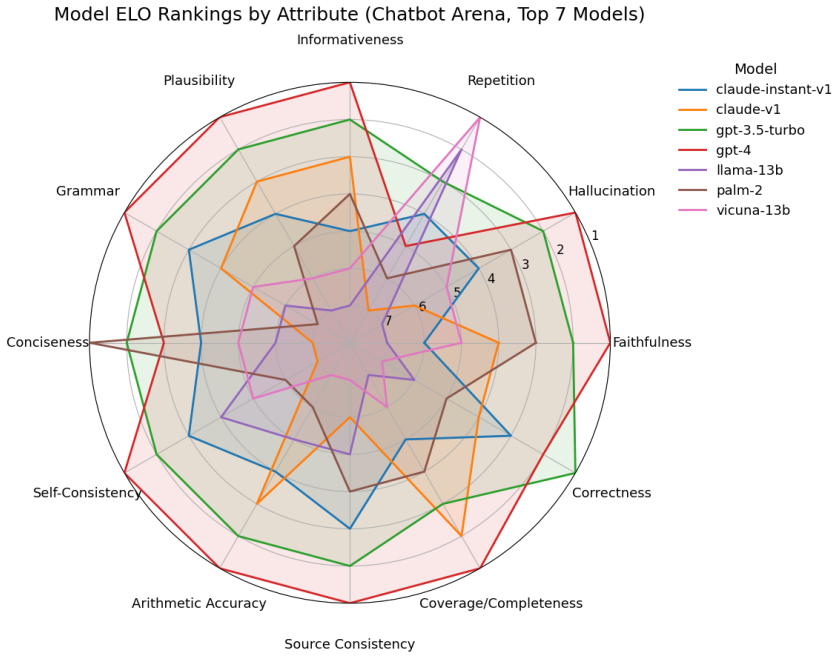


Figure 63: Radar chart comparing model ELO rankings by attribute for the top 7 models in Chatbot Arena (LLM Judges). Each axis is an evaluation attribute, and each polygon represents a model.

2. Mt Bench

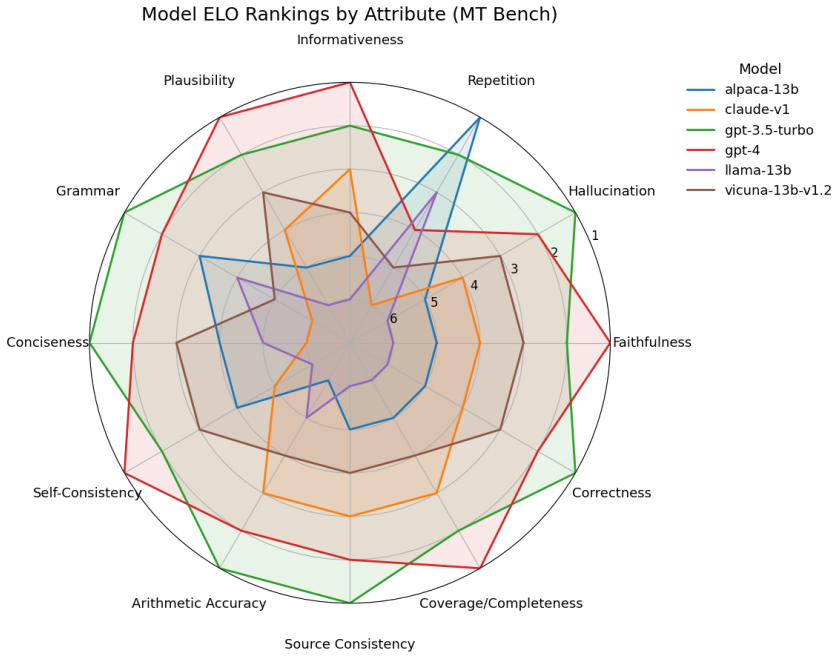


Figure 64: Radar chart showing model ELO rankings by attribute on MT Bench. Each axis represents an evaluation attribute, and each polygon represents a model. (Lower is better)

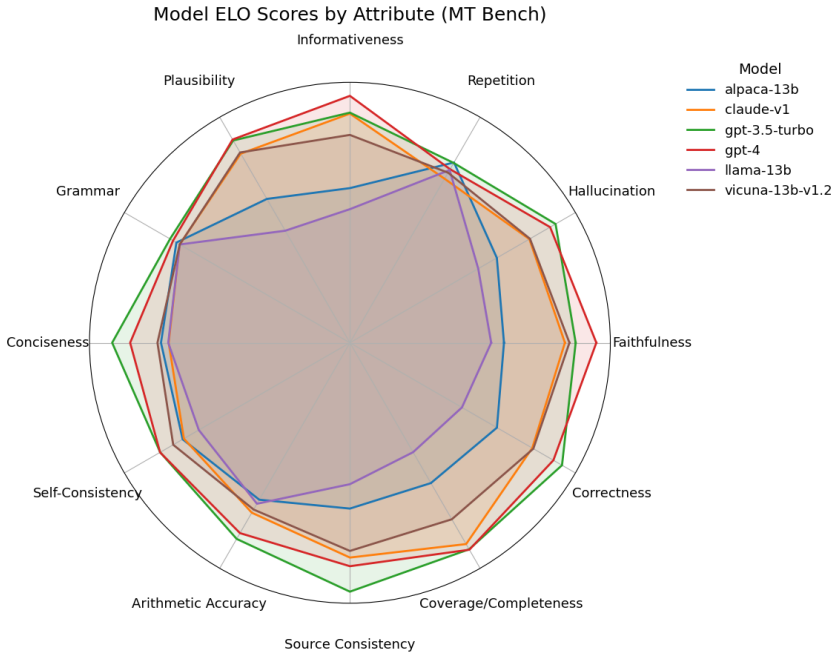


Figure 65: Radar chart showing model ELO scores by attribute on MT Bench. Each axis represents an evaluation attribute, and each polygon represents a model. (Higher is better)

A.7 Examples

An example illustrating a mistake made by the GPT-4o model on the Correctness attribute: In this case, both the chosen and rejected rationales provide incorrect final answers. However, GPT-4o assigns a correctness score of 1.0 to the chosen rationale, failing to recognize the error in the solution.

Question:
x+y = 4z, x*y = 4z^2, express x-y in z
Chosen rationale:
We can start by using the formula for (x+y)^2: (x+y)^2 = x^2 + 2xy + y^2 Substituting the given expressions for x+y and xy, we get: (4z)^2 = x^2 + 2xy + y^2 16z^2 = x^2 + 2(4z^2) + y^2 Rearranging, we get: x^2 - y^2 = 16z^2 - 8z^2 x^2 - y^2 = 8z^2 Now we can factor the left side using the difference of squares formula: (x-y)(x+y) = 8z^2 Substituting the given expression for x+y, we get: (x-y)(4z) = 8z^2 Dividing both sides by 4z, we get: x-y = 2z Therefore, x-y can be expressed as 2z.
Rejected rationale:
x+y = 4z, x*y = 4z^2, express x-y in z

Model: gpt

Attribute	Chosen Score	Rejected Score	Chosen Explanation	Rejected Explanation
Faithfulness	1.0	0.1	The rationale directly uses the provided equations and follows logical algebraic steps to reach the answer.	The rationale simply restates the problem and does not provide any computation or explanation.
Hallucination	1.0	1.0	No extraneous information or assumptions are introduced; all steps are based on the given equations.	No new or extraneous information is introduced; it only repeats the question.
Repetition	1.0	1.0	There is no unnecessary repetition; each step is unique and advances the solution.	There is no unnecessary repetition; the rationale is a single restatement.
Informativeness	1.0	0.1	The rationale clearly explains each algebraic manipulation and substitution, making the reasoning transparent.	The rationale adds almost no meaningful or relevant details beyond the question itself.
Plausibility	1.0	0.2	The steps and final answer are mathematically sound and believable.	It does not provide an answer or reasoning, so it is not plausible as a solution.
Self-Consistency	1.0	1.0	All steps logically follow from one another without contradiction.	There are no internal contradictions, as nothing is actually explained.
Source Consistency	1.0	1.0	The rationale is fully consistent with the information given in the problem statement.	It does not contradict the problem statement; it merely repeats it.
Grammar	1.0	1.0	The explanation is clear, well-structured, and free of grammatical errors.	The rationale is grammatically correct and clear.
Arithmetic Accuracy	1.0	0.0	All calculations, such as squaring and rearranging terms, are performed correctly.	No calculations are performed, so arithmetic accuracy is absent.
Conciseness	1.0	1.0	The rationale is as brief as possible while still providing all necessary details.	It is as short as possible, since it only repeats the question.
Coverage/Completeness	1.0	0.1	Every necessary step is included, from substitution to the final answer.	It does not explain any steps or provide an answer, so coverage is minimal.
Correctness	1.0	0.0	All steps and the final answer (x-y = 2z) are objectively correct.	No steps or answers are given, so correctness is absent.

Figure 66: Enter Caption