Persuasiveness of Generated Free-Text Rationales in Subjective Decisions: A Case Study on Pairwise Argument Ranking

Anonymous ACL submission

Abstract

Generating free-text rationales is among the emergent capabilities of Large Language Models (LLMs). These rationales have been found to enhance LLM performance across various NLP tasks. Recently, there has been growing interest in using these rationales to provide insights for various important downstream tasks. In this paper, we analyze generated free-text rationales in tasks with subjective answers, emphasizing the importance of rationalization in such scenarios. We focus on *pairwise argument ranking*, a highly subjective task with significant potential for real-world applications, such as debate assistance. We evaluate the *persuasiveness* of rationales generated by nine LLMs to support their subjective choices. Our findings suggest that open-source LLMs, particularly Llama2-70B-chat, are capable of providing highly persuasive rationalizations, surpassing even GPT models. Additionally, our experiments show that rationale persuasiveness can be improved by controlling its parameters through prompting or through self-refinement.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022; Touvron et al., 2023) have demonstrated a strong ability to generate *free-text rationales* to explain and support their decisions in plain natural language, which adds an essential layer of transparency and interpretability to their outputs. Recently, there has been a growing interest in utilizing these rationales to enhance the usability and reliability of LLM-based applications, thereby reducing the risks posed by LLMs in decision-making processes (Bender et al., 2021).

Existing research on evaluating and analyzing free-text rationales has primarily focused on tasks where there is an expected factual ground truth answer that the model should achieve even without further explanation. Most of this work has

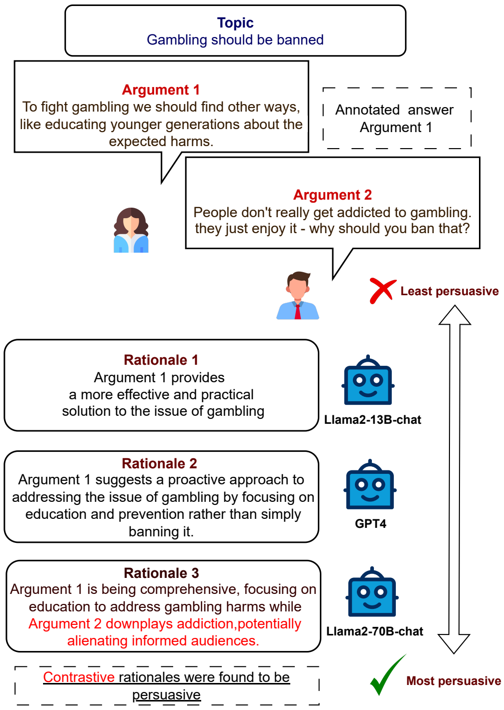


Figure 1: Given two arguments with the same stance on a topic, the model selects the higher quality argument and generates a convincing rationale. We analyze the *persuasiveness* of these rationales.

focused on assessing the plausibility (Wiegreffe and Marasovic, 2021; Marasović et al., 2022) and faithfulness (Wiegreffe et al., 2021) of these rationales to produce accurate answers. Recently, studies have been introduced to also analyze rationales for their utility in learning new concepts (Joshi et al., 2023a) and truth verification (Si et al., 2023).

In this work, we analyze free-text rationales in subjective tasks where annotations, despite agreement, remain subjective. We focus on *rationale persuasiveness* to understand how different LLMs convincingly justify their choices. Specifically, we examine rationales in *pairwise argument ranking* (Gretz et al., 2020; Toledo et al., 2019), a task with inherent subjectivity and significant potential for applications like debate assistance tools (Wachsmuth et al., 2024). In this task, the model recommends one argument from a pair on a controversial topic. We believe that adding persuasive rationales to argument recommendations can enhance their utility in downstream applications. Figure 1 shows examples of rationales generated by various models. While these models agree on the pairwise ranking, their rationales reveal different levels of persuasiveness in supporting Argument 1.

We provide a comprehensive analysis of the persuasive nature of free-text rationales by addressing the following research questions (RQs): **RQ1:** *How do different LLMs compare in generating persuasive rationales?* **RQ2:** *Can we automatically detect the more persuasive rationales?* **RQ3:** *Which characteristics of a rationale contribute to its persuasiveness?* **RQ4:** *Can we control the persuasiveness of generated rationales?* To address these questions, we: (1) Prompt 9 different LLMs to perform zero-shot pairwise ranking and provide rationales for their choices. (2) Use manually annotated rationales to evaluate automatic persuasiveness detection methods, specifically GPT4 (OpenAI, 2023), for ranking rationale persuasiveness, enabling large-scale analysis. (3) Conduct a human evaluation study to rank the persuasiveness of generated rationales and examine the influence of the rationale's content. (4) Experiment with enhancing rationale persuasion by prompting the model with key aspects for persuasion learned from prior steps and explore automatic self-improvement techniques to assess if the model can improve its persuasiveness.

Our findings can be summarized in four key points: (1) Open-source LLMs, particularly Llama2-70B-chat, excelled in generating persuasive rationales, even outperforming GPT4. (2) GPT4 closely matched human rankings of the persuasiveness of the rationales, although a perfect agreement was unattainable due to the inherent subjectivity of the task. (3) Contrastive rationales, which justify why the alternative argument was not chosen, emerged as the most influential factor in persuasiveness. (4) Prompting the model with persuasiveness factors can enhance the persuasiveness of the generated rationales.

2 Related Work

Argument Quality Ranking Argument quality ranking is a key task in argument quality estimation, which can be approached in two main settings: *(1) pointwise ranking,* where arguments are individually assessed based on a quality score like interpretability (Swanson et al., 2015), human quality annotations (Toledo et al., 2019; Gretz et al., 2020); and *(2) pairwise ranking*, where the quality of the arguments is estimated in comparison to each other, using factors such as persuasiveness (Habernal and Gurevych, 2016; Simpson and Gurevych, 2018) or aggregated preferences (Toledo et al., 2019). *Our work adopts the pairwise ranking framework in a zero-shot setting.*

LLMs for Argument Quality Ranking Despite their strong performance in various tasks, Wang et al. (2023a) demonstrated that LLMs, particularly the GPT-3.5-turbo, struggle to match supervised models in point-wise and pair-wise ranking tasks, even in few-shot settings. Instead of relying solely on existing benchmarks, Mirzakhmedova et al. (2024) showed that LLMs, especially PALM2 and GPT-3, are effective in annotating argument quality, particularly when combined with human annotations. Recently, Wachsmuth et al. (2024) suggested that LLMs could open new directions in argument quality research, such as fact-checking and argument optimization. *In this work, we analyze the persuasiveness of rationales generated by different LLMs, proposing that LLMs can enhance argument quality-based applications by providing users with persuasive explanations to support their decisions.*

Evaluating Free-Text Rationalization Evaluating free-text rationales has primarily focused on their ability to aid models in reaching correct answers. Metrics such as accuracy differences between predictions with and without rationales (Hase et al., 2020; Wiegreffe et al., 2021) and information-theoretic measures (Chen et al., 2023) assess how rationale content supports model performance. Wiegreffe and Marasovic (2021) established criteria for evaluating rationales, including surface form for validity and grammatical correctness, support for association between the rationale and the label, and *contrast* with alternative labels. Building on this, Joshi et al. (2023a) introduced *novelty*, measuring the extent of new information provided by the rationale, enhancing its utility in

Dataset	# Argument Pairs (Unfiltered)	# Argument Pairs (Filtered)	# Rationales	# Rationale Pairs for Persuasion Annotated	Full
IBM-9k	400	30	270	204	1080
IBM-30k	1534	144	1296	-	5184

Table 1: Summary of datasets for evaluating free-text rationales. Unfiltered is the total argument pairs sampled, Filtered is the subset with unanimous LLM agreement, and Annotated is the subset used for human evaluation.

human-ai collaboration tasks. In the context of persuasiveness, Ajwani et al. (2024) found that LLMs can convincingly support incorrect predictions in the NLI task. *Given our study's focus is close to rationale utility, we adopt the dimensions introduced by Joshi et al. (2023a) to evaluate our rationale content. We focus on persuasiveness for subjective tasks like pairwise argument ranking. We also included a large number of models and evaluation measures.*

Persuasiveness in LLMs Prior research on the persuasiveness of LLMs has compared generated arguments with those written by humans. Bai et al. (2023) conducted a randomized control trial showing that GPT-3 can write persuasive political arguments comparable to human ones. Similarly, Palmer and Spirling found that GPT-3's texts on controversial topics were as persuasive as those written by crowdsource workers. Salvi et al. (2024) demonstrated that personalization enhances GPT4's persuasiveness in conversations. Rescala et al. (2024) also showed that GPT4 can detect persuasiveness in debates as effectively as crowdsource workers. *However, most of this research has focused on large commercial LLMs and analyzing the arguments themselves. We shift the focus to the persuasiveness of rationales. Additionally, we include a broader range of LLMs for a more comprehensive analysis.*

3 Experimental Settings

3.1 Datasets

To assemble the free-text rationales evaluation set, we used argument pairs from two datasets: *IBM-ArgQ-9.1kPairs* (IBM-9k) (Toledo et al., 2019) and *IBM-30k-rank* (IBM-30k) (Gretz et al., 2020). The IBM-9k dataset contains pairs of arguments either supporting or opposing a topic, with annotations for the higher-quality argument. The IBM-30k dataset includes individual arguments annotated with quality scores ranging from 0 to 1.

From the IBM-9k dataset, we randomly selected 400 argument pairs from the test set, evenly distributed across 20 topics. This set was used for manual analysis and evaluation due to its quality control measures, which ensure that argument pairs advocate the same stance, are of high quality, and have comparable lengths to avoid length bias (Potash et al., 2017). These pairs were used to prompt the LLMs for argument predictions and supporting rationales. We filtered out pairs where any LLM failed to predict the annotated winning argument, focusing on pairs with unanimous agreement to ensure a fair comparison between the generated rationales. This left us with 30 argument pairs [1], each with rationales generated by 9 models, totaling 270 rationales. Comparing these rationales for persuasiveness resulted in 1080 rationale pairs for evaluation.

For the IBM-30k dataset, we created a pairwise ranking set by sampling arguments that (1) have a similar stance, (2) vary in length by a maximum of 20% to avoid bias, (3) each appear at most once to diversify the comparison set while reducing computation cost of prompting, and (4) have different quality scores, allowing us to assess the influence of the quality differences on the persuasiveness. This resulted in 1534 pairs. We followed a similar prompting and filtering technique used for the IBM-9k dataset, which left us with 144 unanimously agreed upon pairs, totaling $144 * 9 = 1296$ rationales. Comparing these rationales for persuasiveness resulted in 5184 persuasion pairs. This dataset acts as an extended test set to assess whether our findings on the IBM-9k dataset will generalize to other topics and arguments. Table 1 [2] shows the statistics of the datasets included in our work.

3.2 Models

Considered LLMs Our study employs a set of LLMs to investigate the influence of various features on the generated rationales. (1) **Open-source** models include *Llama2* (Touvron et al., 2023) and *Vicuna* (Zheng, 2023), while **closed-source**

[1]Appendix A shows that considering agreement among all models leads to a significant reduction in the number of argument pairs

[2]Data and code to be released upon acceptance.

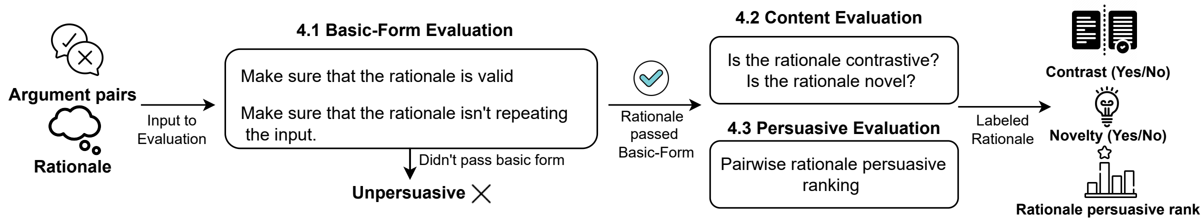


Figure 2: For the input argument pair and rationale, we filter out invalid or repetitive rationales (Section 4.1). The qualified rationales are then analyzed based on their content (Section 4.2) and persuasiveness (Section 4.3).

models include *GPT models* (GPT-3.5-turbo and GPT4) recognized for superior performance on many downstream NLP tasks (Wang et al., 2023b). (2) **Instruction tuning** is represented by the *chat* versions of Llama2 and Vicuna, where the latter is primarily fine-tuned based on human preferences between pairwise model generations. (3) For each LLM family, we test various **model sizes**, namely 7-B and 13-B versions of *Vicuna* and *Llama2* (both chat and non-chat versions) and *Llama2-70B-chat*.[3]

For all open-source models, we utilized the Hugging Face library implementations (Wolf et al., 2019). As for the OpenAI models, we employed the OpenAI API [4] to prompt both GPT-3.5-turbo and GPT4. To reduce randomness in generation, we set the temperature during decoding to 0.

Prompting LLMs for Ranking Arguments and Generating Rationales Our prompt is structured to contain three components. (1) **System Message:** This includes a designated system setting assumed by the model during the task. (2) **Task Description:** We describe the ranking task, assigning numerical identifiers to arguments as recommended for LLM-based ranking tasks (Sun et al., 2023; Pradeep et al., 2023). To improve clarity, we include ranking criteria from prompts used by human annotators for assessing argument quality (Toledo et al., 2019; Gretz et al., 2020). Additionally, we instruct the model to generate reasoning to support its chosen argument. (3) **Formatting Examples:** We present the model with input format and the expected output. To prevent bias, we provide two formatting examples, one where argument 1 is the winner (the selected argument in pairwise ranking) and another where argument 2 is the winner. This ensures the model includes all expected components in its output. [5]

4 Rationale Evaluation

Figure 2 outlines our evaluation process, which consists of three key stages: (1) **Basic-Form Evaluation:** This initial stage filters out meaningless rationales, ensuring only valid ones proceed for further analysis, similar to the concept of surface-form evaluation (Joshi et al., 2023a). (2) **Content Evaluation:** We assess the rationale's content by analyzing its support through *contrast* and its informativeness through *novelty*, aiming to understand how rationale content influences its persuasiveness. (3) **Persuasiveness Evaluation:** We assess the rationale's persuasiveness relative to other generated rationales supporting the chosen argument.

We rely on human annotators to evaluate each stage, using the 270 rationale subset from the IBM-9k described in Table 1. This annotated set is used to: (1) Analyze the influence of rationale content on the rationale persuasiveness, and (2) Explore automatic persuasiveness evaluation methods to reduce the cost of human evaluation, especially in utility-driven tasks (Joshi et al., 2023a).

We use Mechanical Turk workers for annotations at each evaluation stage. Each dimension in basic form and content evaluation is assessed with yes/no questions, with the majority vote of three workers determining the final decision on contrast and novelty. For persuasiveness, rationales are evaluated through pairwise comparisons, where three workers vote on whether a rationale is more, less, or equally persuasive compared to another. Workers also write $1-2$ sentences explaining their choices to ensure critical thinking. Rigorous quality assurance methods ensure annotation reliability. [6]

[3]GPT4 is reported to have massive parameters, though the exact details are undisclosed.

[4]https://openai.com/blog/openai-api

[5]The exact prompt is included in Appendix B.

[6]Appendix C details our annotation guidelines, quality assurance methods, and annotator reliability.

Analysis Type	Examples
Input Argument Example	**Topic: "Flu vaccination should not be mandatory"** **Argument 1:** "the state ought not intervene in individuals' right of bodily autonomy regardless of societal harms. forcing vaccinations is a crude invasion to someone's body and should therefore not be done." **Argument 2:** "forcing people to participate in a system they do not trust, often with passion and for irrational reasons, might lead them to distrust conventional medicine as a whole, increasing the risk to their health." Decision: "Argument 1".
Invalid reasoning	**Reasoning:** "argument 1 is more convincing than argument 2
Repetitive reasoning	**Reasoning:** "It is a crude invasion to someone's body and should therefore not be done."
Valid reasoning	**Reasoning:** "The first argument emphasizes the importance of bodily autonomy and the limitation of state intervention in personal choices. It presents a clear ethical principle that can be applied consistently across different situations. The second argument highlights potential negative consequences of mandatory vaccination. , which, while valid, may not be as persuasive as the ethical principle of bodily autonomy. "

Table 2: Examples of *Invalid Reasoning* (second row), *Repetitive Reasoning* (third row) from non-compliant sets, and *Valid Reasoning* (fourth row) with highlighted text: Repetitive part , Contrastive part , Novel part .

4.1 Basic-Form Evaluation

We examine two aspects of the quality of a rationale to assess the rationale form: *Validity*: Is the rationale grammatically correct and coherent? *Repetition*: Does the rationale merely reiterate the input argument, either fully or in summary, without adding any new insight or justification?

Examples of rationales that fail to meet these criteria are provided in Table 2, with invalid rationales shown in the second row and repetitive rationales in the third row. If a rationale doesn't meet these basic-form requirements, it is disregarded from further evaluation and deemed unpersuasive by default.

4.2 Content Evaluation

For *contrast*, we assess the LLM's ability to refute the argument it did not choose. Our goal is to determine if refuting the alternative argument enhances the rationale's persuasiveness. For *novelty*, we evaluate whether the rationale introduces new information or a new perspective not explicitly mentioned in the arguments, thereby increasing its persuasiveness. An example of a valid rationale with highlighted contrastive and novel (new perspective) parts can be found in Table 2, row 4[7].

[7]We also analyzed rationale content for support by evaluating *association* (Wiegreffe et al., 2021; Wiegreffe and Marasović, 2021), determining if the rationale highlights key points in the chosen argument. Most LLMs supported their choices through association, offering no unique information for persuasiveness ranking.

4.3 Persuasiveness Evaluation

In recommendation tasks, persuasive explanations help users understand why a certain item or choice is recommended, convincing them to accept it (Wang et al., 2014; Tran et al., 2023). Similarly, in argument ranking, persuasiveness of the rationales can be defined as the ability to convincingly justify the model's recommendation of one argument over another. Due to the subjective nature of this task, we opted against assigning a single persuasiveness score. Instead, we evaluate persuasiveness through pairwise comparisons, allowing us to assess the persuasiveness abilities of different models supporting the same choice.

Human Evaluation of Persuasiveness Due to the quadratic nature of pairwise comparisons, we randomly select one third of the rationale pairs for persuasion described in Table 1, resulting in 360 pairs. After excluding rationales that do not meet basic quality standards, we are left with 204 pairs for human annotations. We refer to this subset as IBM-9k (annotate set).

Automatic Evaluation of Persuasiveness To assess persuasiveness rankings on a larger scale across all pairs in our study (both the full IBM-9k and IBM-30k pairs), we utilize GPT4 for automatic persuasiveness ranking. GPT4 is selected for its proven effectiveness in evaluating various downstream tasks (Liu et al., 2023; Chiang and Lee, 2023). We benchmark GPT4's rankings against human persuasiveness rankings on the annotated set

Model	IBM9k (Annotated Set)		IBM9k (Full Pairs)	IBM-30k-rank			
					Quality Differences		
				Full Pairs	0-0.25	0.25-0.5	0.5-1.
	APR (δ) with Human-Eval ↑	APR (δ) GPT4 Eval ↑	APR (δ) GPT4 Eval ↑	APR (δ) GPT4 Eval ↑			
Llama2-13B-Chat	2.28(0.48)	2.14(0.37)	3.42(1.95)	3.75(1.74)	3.69(1.73)	3.85(1.69)	3.60(2.02)
Llama2-7B-Chat	3.14(1.46)	3.42(0.78)	3.85(2.20)	4.15(2.09)	4.11(2.07)	4.43(2.13)	3.66(2.12)
Vicuna-7B	3.63(0.80)	4.18(1.72)	4.39(1.49)	3.75(1.36)	3.61(1.38)	3.75(1.28)	4.60(1.24)
Vicuna-13B	4.36(1.56)	3.72(1.10)	4.67(1.46)	4.45(1.41)	4.60(1.37)	4.17(1.41)	4.45(1.55)
GPT-3.5-Turbo	5.18(1.16)	6.00(1.48)	6.14(1.53)	5.08(1.37)	5.11(1.35)	4.95(1.53)	5.00(1.00)
GPT4	5.72(1.55)	5.72(1.19)	5.92(1.27)	5.82(1.06)	5.86(0.93)	5.82(1.18)	5.66(1.34)
Llama2-70B-Chat	**7.00 (1.09)**	**6.18 (1.77)**	**6.57 (1.66)**	**6.29 (0.98)**	**6.14 (1.07)**	**6.09 (1.71)**	**5.91 (0.91)**

Table 3: Average Persuasive Rank (APR) (δ) for 7 instruction-tuned LLMs and datasets. δ denotes the standard deviation. ↑ indicates higher persuasiveness. Rows are sorted by Human-Eval APR in ascending order.

and then report its persuasiveness ranking scores across all IBM-9k rationale pairs (IBM-9k Full Pairs) and the IBM-30k dataset.

Persuasive Ranking Metric For both human and automatic evaluations, we use the scoring formula proposed by Qin et al. (2023) in ranking passages for retrieval tasks, to rank persuasiveness of the rationales. The score s_i for a rationale r_i is given by:

$$s_i = 1 \cdot \sum_{\substack{j=1 \\ j \neq i}}^{M} \mathbb{I}_{r_i > r_j} + 0.5 \cdot \sum_{\substack{j=1 \\ j \neq i}}^{M} \mathbb{I}_{r_i = r_j} \tag{1}$$

where M is the total number of considered models and r_i and r_j are the rationales from model i and model j, respectively. This formula adds 1 to the score s_i if a rationale r_i is considered more persuasive than r_j, and 0.5 if it is considered equally persuasive. To determine the overall persuasiveness of each model, we use the s_i scores to rank the models' generated rationales for each argument pair and report the **Average Persuasiveness Rank (APR)** of each model as the final persuasiveness score, ranging from 1 ranked the least persuasive and M, which is the total number of models included in the comparison, as the most persuasive.

To compute Equation 1 using GPT4, we instruct the model to compare the persuasiveness of rationale 1 and rationale 2 in supporting the argument. Same as human evaluation, we include a third option for GPT4 to select if it finds both rationales equally persuasive. Furthermore, following the method described by Qin et al. (2023), we present the rationale pairs to GPT4 twice, each time with the order of rationales switched. If GPT4's decision differs between the two prompts, we consider the rationales to be equally persuasive and increase the s score of each rationale by 0.5 [8].

5 Results and Analysis

5.1 Persuasiveness Rankings of Rationales

Human and Automatic Persuasive Rankings (RQ1, RQ2) Table 3 presents the APR in all data sets. Llama2-7B and Llama2-13B were excluded from the rankings because their basic-form annotations indicated a consistent failure in quality check, making them the least persuasive by default[9]. Therefore the APR is reported across 7 LLMs instead of 9. Llama2-70B-chat consistently generated the most persuasive rationales. This was evident in both human and automatic rankings with GPT4, surpassing even closed-source GPT models. This result highlights the potential of open-source models like Llama2-70B-chat in tasks such as pairwise argument ranking.

For the IBM-9k annotated set, GPT4 did not perfectly match the APR with human evaluation. However, GPT4 agreed with human evaluation in the persuasiveness ranking order of the included LLMs, except for the rankings of GPT4 vs. GPT-3.5-Turbo and Vicuna-7B vs. Vicuna-13B. This suggests that GPT4 can differentiate between the persuasiveness of rationales when differences are significant, but may disagree with human judgment when the persuasiveness scores are close.

For the IBM-30k data set, the variation in the difference in the quality of arguments had a limited effect on the persuasiveness of the rationale. The rationale generated by Llama2-70B chat remained the most persuasive, followed by those of GPT4 and GPT-3.5-turbo. This indicates that different LLMs tend to follow a similar rationalization strategy regardless of the quality difference. For

[8] Prompt in Appendix D.

[9] Appendix E details the basic-form distribution across all models.

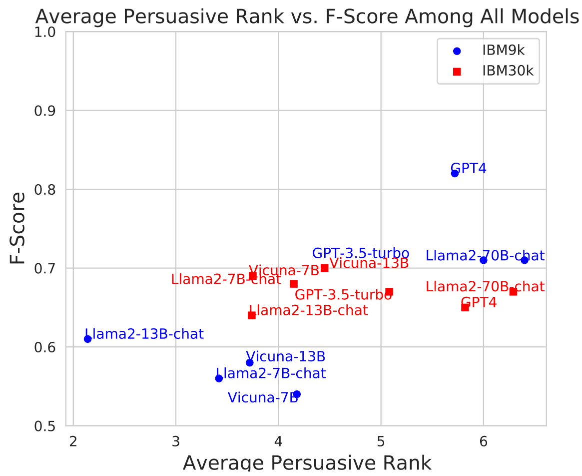


Figure 3: Persuasion Ranking vs F-score

all datasets, we found that instruction tuning and model size improves persuasiveness [10].

5.2 What contributes to the rationale persuasiveness? (RQ3)

Model Accuracy ≠ Rationale Persuasion Figure 3 shows that the LLM's ability to accurately predict the annotated higher-ranked argument, measured by the F1 score between the LLM's predicted argument and the annotated argument on the full unfiltered argument pairs of the IBM-9k and IBM-30k datasets, does not necessarily correlate with higher persuasiveness scores measured by GPT4 across the IBM-9k annotated set and the IBM-30k full set. This is further supported by the insignificant Pearson correlation results, with $p > 0.05$ for both datasets.

For example, despite having the highest persuasiveness rank, Llama2-70B-chat falls behind GPT4 in the F1 score for the IBM-9k dataset. This trend is more apparent with the IBM-30k pairs, where both GPT4 and GPT-3.5-turbo have lower F1 scores compared to the Vicuna models, yet achieve higher persuasive rankings. The drop in F1 scores can be attributed to the quality variation in the IBM-30k test set, affecting the LLM's ability to agree with the annotated higher argument, but having limited impact on how the model supports its prediction. These observations indicate that a model's ability to convincingly support an argument extends beyond mere accuracy in predicting the labeled argument, suggesting a complex interplay of factors that influence a model's persuasive capabilities.

[10]Details are in Appendix F.

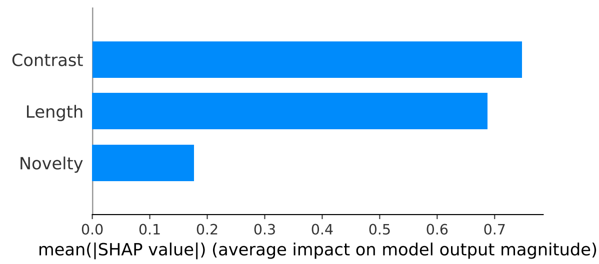


Figure 4: *SHAPLey* values of each feature. The higher the value , the higher the impact on average persuasiveness rank.

Rationale Content Analysis In addition to *Contrast* and *Novelty*, we also explore the observable characteristic of rationale *Word length* on the persuasiveness ranking of the rationales and investigate the role of these attributes. We formulate this as a regression task, employing a *random forest regressor* (f) to predict persuasiveness ranking based on the features: length (X_{length}), contrast ($X_{contrast}$), and novelty ($X_{novelty}$). $Ranking = f(X_{length}, X_{contrast}, X_{novelty})$. We convert the contrast and novelty majority votes for each rationale into binary values. Upon estimating f, we use the **SHAP explainer** (Lundberg and Lee, 2017) to determine the impact of each feature on the persuasiveness ranking. We particularly used SHAP as it takes into consideration the feature interaction when estimating the individual feature impact on the predictions.

Figure 4 shows that *contrast* is the most influential factor in persuasiveness. This aligns with studies advocating for contrastive explanations in truth verification (Si et al., 2023) but deviates from Joshi et al. (2023b), where contrast had minimal influence on rationale utility. We hypothesize that this is intuitive, given the nature of our task. By weakening the alternative arguments, we can make the argument choice more acceptable and enhance the rationale's persuasiveness. *Length* is also significant, indicating that more detailed explanations may improve persuasiveness. Lastly, *novelty* has a less pronounced impact, suggesting that while new information is valuable, its role is secondary to contrast and length in this context.

To understand the content contribution independent of content length, we cluster rationales into two groups: *High-Persuasive (HP)* and *Low-Persuasive (LP)* clusters, using k-means clustering. We then control for length variations by focusing only on rationales with word lengths within 20%

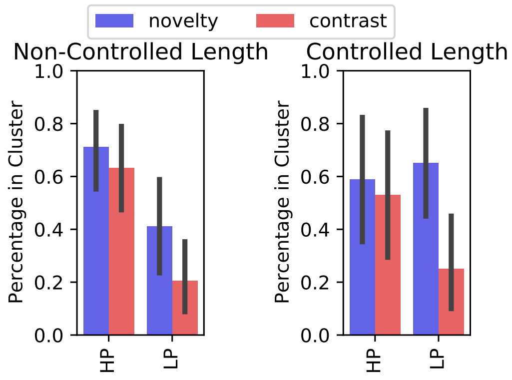


Figure 5: *Contrast* and *Novelty* % in different categories of rationale rating.

Model	APR GPT4 Eval ↑
Llama2-7B-Chat	4.31(2.86)
Llama2-7B-chat-persuasion-prompted	6.65(2.97)
Llama2-7B-chat-persuasion-refined	5.15(2.88)
Llama2-13B-Chat	3.68(2.00)
Llama2-70B-Chat	**7.89 (2.05)**
Vicuna-7B	5.57(1.74)
Vicuna-13B	5.52(2.06)
GPT-3.5-Turbo	7.63(1.53)
GPT4	7.21(1.39)

Table 4: (APR) LLMs on the IBM9k (Full Pairs) dataset using GPT4. *Italicized* rows indicate the Llama2-7B-chat models experimented for enhanced persuasiveness.

of each other. This ensures that any observed differences in persuasiveness are primarily due to content, not length. Figure 5 illustrates that in the IBM-9k annotated set of rationales , both novelty and contrast percentages are significantly higher (ANOVA-test, $p < 0.05$) in the High-Persuasive group. However, in the controlled length rationale set, only contrast exhibits a significant increase in the High-Persuasive group (ANOVA-test, $p < 0.05$). These results verify the SHAP analysis, emphasizing the importance of contrast in persuasion. Conversely, the presence of novelty in lengthy rationales may act as a confounding factor, potentially inflating its significance.

5.3 Controlling Persuasiveness (RQ4)

We aim to use the insights from the previous session to improve the model's ability to generate persuasive rationales. We experimented with **Re-prompting the LLM:** This involved asking the model to provide two sentences supporting its chosen argument and two sentences refuting the alternative argument. The goal was to encourage the model to include contrastive rationales with sufficient length, proven influential for persuasiveness. we compare this method against **Evaluate and Refine:** which is a form of self-refinement (Huang et al., 2022). The model first assesses whether the generated rationale was persuasive. If the model determines that the rationale is not persuasive, it then generates a more persuasive one. Both methods were applied to the Llama2-7B-chat model, which, as shown in Appendix E, had a low rate of generating contrastive rationales. We refer to the new rationales generated by the model as *Llama2-7B-chat-persuasion-prompted* and *Llama2-7B-chat-persuasion-refined*, respectively [11].

Table 4 shows that *Llama2-7B-chat-persuasion-prompted* ranks higher in persuasiveness with GPT4-based ranking compared to both Llama2-7B-chat and self-refined rationales (*Llama2-7B-chat-persuasion-refined*), which emphasizes the importance of contrast and detail in enhancing rationale persuasiveness. However, the new rationales still lag behind Llama2-70B-chat and GPT models, indicating that larger models may rely on persuasive factors unexplored in our work. *Evaluate and Refine* method did not improve persuasiveness compared to prompting with persuasive parameters, suggesting that LLMs benefit more from alignment on persuasive factors.

6 Conclusion and Future Work

This paper presents a comprehensive analysis of the persuasiveness of free-text rationales generated by various LLMs. Our results show that open-source models, particularly Llama2-70B-chat, generate highly persuasive rationales, surpassing strong closed-source GPT models. While GPT4's rankings generally align with human judgments, discrepancies arise due to the task's inherent subjectivity. We proposed a detailed human evaluation studying key factors contributing to persuasiveness. We found that *contrastive rationales*,where the model justifies its choice and refutes the alternative, the most significant. We also demonstrated that prompting models with specific persuasiveness parameters enhances rationale persuasiveness. Future work will explore the user acceptance of model-chosen arguments and investigate other subjective tasks beyond pairwise argument ranking.

[11] Prompts are in Appendix G.

7 Limitations

This study primarily utilized rationale evaluation taxonomies to assess persuasiveness. Future work could incorporate additional factors from persuasive theory to gain a deeper understanding of what different LLMs rely on to support their choices. Our annotated sample size is relatively small, as we prioritized quality control over a larger quantity of annotations. Although we hypothesize that our results would be consistent with a larger sample, it would strengthen our findings to re-evaluate our methods on a broader dataset. Additionally, expanding the study to other domains where the task is inherently subjective, beyond pairwise argument ranking, would provide a more comprehensive evaluation.

8 Ethical Statement

Persuasive rationales can enhance transparency, particularly in subjective tasks, by making recommendations more acceptable to users. However, there is a potential ethical concern that persuasive rationales could be used adversarially to promote biased or nonfactual arguments. Therefore, it is crucial to consider the ethical implications of deploying persuasive rationales and to develop safeguards to prevent misuse.

References

Rohan Ajwani, Shashidhar Reddy Javaji, Frank Rudzicz, and Zining Zhu. 2024. Llm-generated blackbox explanations can be adversarially helpful. *arXiv preprint arXiv:2405.06800*.

Hui Bai, Jan Voelkel, Johannes Eichstaedt, and Robb Willer. 2023. Artificial intelligence can persuade humans on political issues.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. Rev: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023a. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023b. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128, Toronto, Canada. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval:

Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? *arXiv preprint arXiv:2404.09696*.

OpenAI. 2023. Gpt-4 technical report. Technical report.

Alexis Palmer and Arthur Spirling. Large language models can argue in convincing and novel ways about politics: Evidence from experiments and human judgement.

Peter Potash, Robin Bhattacharya, and Anna Rumshisky. 2017. Length, interchangeability, and external knowledge: Observations from predicting argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 342–351, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? *arXiv preprint arXiv:2404.00750*.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint arXiv:2403.14380*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large language models help humans verify truthfulness–except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment-new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Thi Ngoc Trang Tran, Alexander Felfernig, Viet Man Le, Thi Minh Ngoc Chau, and Thu Giang Mai. 2023. User needs for explanations of recommendations: In-depth analyses of the role of item domain and personal characteristics. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 54–65.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. *arXiv preprint arXiv:2403.16084*.

Beidou Wang, Martin Ester, Jiajun Bu, and Deng Cai. 2014. Who also likes it? generating the most persuasive social explanations in recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Yiran Wang, Xuanang Chen, Ben He, and Le Sun. 2023a. Contextual interaction for argument post quality assessment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10420–10432.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Qinyuan Zheng. 2023. WKU_NLP at SemEval-2023 task 9: Translation augmented multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1525–1530, Toronto, Canada. Association for Computational Linguistics.

A Argument Pairs Agreement Distribution

Figure 6 illustrates that the number of agreed-upon argument pairs decreases as more models are included in the analysis. The "High Accuracy" category includes GPT-4, GPT-3.5-turbo, and Llama2-70B-chat. The "Instruction Tuned" category adds the remaining instruction-tuned models to the high-accuracy models: Llama2-7B-chat, Llama2-13B-chat, Vicuna-13B, and Vicuna-7B. Finally, the "All Models" category includes the non-instruction-tuned models Llama2-7B and Llama2-13B in addition to those in the previous categories. For a more comprehensive analysis, we included all models in our analysis.

Obtaining Rationalization Pairs For each argument pair, we generate 9 different rationales from the included LLMs. Using pairwise comparisons to rank these rationales results in 36 combinations per argument pair. Consequently, for the total filtered argument pairs, we have 1080 rationale pairs for the IBM-9k dataset and 5184 for the IBM-30k dataset.

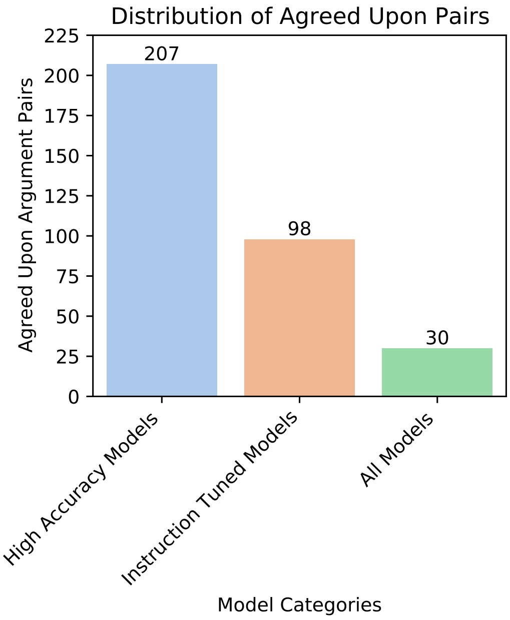


Figure 6: Distribution of argument pairs across different categories of models in the IBM-9k sampled set.

B Pairwise Ranking prompt

Table 5 shows the exact prompt used for our first stage pairwise ranking. The "Expected Output" section of the prompt indicates the format in which the model generates responses and not an actual output.

C Mechanical Turk HITS

C.1 Basic-form Evaluation in Detail

Evaluation Process and Worker Reliability We employ Mechanical Turk workers with more than 95% approval rate and more than 5000 approved HITS. Workers are instructed to select **yes** if the rationale is both valid and devoid of any partial or full repetition of the chosen argument. If the rationale violates either of these conditions, Turkers are directed to choose **no**. Additionally, they are required to specify the reason for rejecting the rationale, selecting between "invalid rationale" or "repetitive rationale." At first, workers were given 20 examples to help them understand the task requirements and estimate its difficulty. Along with the task description, clear instructions and examples were provided to avoid any possible confusion. Three workers evaluated each sample. The reliability of the workers was measured using Krippendorff's alpha (α) (Krippendorff, 2011). The initial score of 0.53 was achieved for basic labeling criteria and

Pairwise Ranking Prompt

System message You possess the art of argumentation.

Task definition You will receive two arguments, each identified by a numerical identifier [] and a Topic. *Disregarding your own opinion on the topic, given the arguments, the human decision, and the human reasoning, decide which argument you would recommend.* Choose argument [1] if you recommend argument [1] over argument [2]. Choose argument [2] if you recommend argument [2] over argument [1]. Format your output in a JSON format with "decision" and "reasoning" keys.

Reminder: Make sure to choose only one argument and provide a convincing reasoning why you choose this argument over the other one. Generate only the JSON output with decision and reasoning, do not generate any additional thought process or discussion.

Formatting Examples

Example 1:

```
{
  "topic": "topic 1",
  "1": "argument 1",
  "2": "argument 2",
  "model_decision": "argument 1",
  "model_reasoning": "reason model chose argument 1"
}
```

Output:

```
{
  "decision": 1,
  "reasoning": "reason for choosing argument 1"
}
```

Example 2

```
{
  "topic": "topic 2",
  "1": "argument 1",
  "2": "argument 2",
  "model_decision": "argument 1",
  "model_reasoning": "reason model chose argument 1"
}
```

Output

```
{
  "decision": 2,
  "reasoning": "reason for choosing argument 2"
}
```

Annotation Example

```
{
  "topic": "{}",
  "1": "{}",
  "2": "{}",
  "model_decision": "argument {}",
  "model_reasoning": "{}"
}
```

Expected Output (generated by the model in json format)

```
{
  "decision": "...",
  "reasoning": "..."
}
```

Table 5: Pairwise argument ranking prompt. *italicized* part in **Task definition** is the prompt given to human annotators described in (Gretz et al., 2020; Toledo et al., 2019) .

0.27 for identifying reasons for non-compliance. To improve the evaluation quality, we disqualified workers who failed to answer hidden test questions and introduced a set of 20 examples with revised guidelines. This led to an improved score of 0.80 for basic form labeling and 0.66 for identifying reasons for non-compliance on the additional set of 20 examples. Using these revised guidelines, we evaluated the final set of 270 rationales. The reliability score for this phase was 0.76 for basic form labeling and 0.71 for identifying reasons for failure, whether due to validity or repetition (non-compliance). The majority votes from workers' assessments were used to evaluate each sample. Samples that failed to meet basic form criteria, as determined by the majority vote, were excluded from further evaluation phases.

Basic-form HIT Figure 7 shows the actual MTurk HIT given to Turkers to evaluate the basic form. First, workers are asked to select YES/NO based on the validity and repetition criteria. If they select NO, they are asked to choose a reason between **Invalid** and **Repetitive** for selecting NO.

C.2 Content Evaluation HITs

Annotator Qualification Process Similar to the basic-form evaluation, we conduct this step using YES/NO questions to determine whether the rationale is contrastive or novel. These questions are answered by proficient English-speaking Mechanical Turk workers who have passed our qualification test. Content evaluation began with a qualification task for our annotators, all of whom are proficient in English. This initial task consisted of annotating 10 sample rationales. The samples were selected based on their known, expected annotations in novelty and contrast to ensure the accuracy of the qualification process. Each sample was reviewed by 5 workers. Only those workers who accurately completed at least 8 out of the 10 questions and achieved more than 90% agreement with the expected annotations were retained for the subsequent evaluation.

Final Content Evaluation For each sample, we employ three qualified workers to assess both *contrast* and *novelty* aspects, using a binary YES/NO selection. The final label for each rationale is determined by the majority vote among these workers. For the complete final evaluation set, we computed Krippendorff's alpha coefficient, resulting in a value of 0.82 for contrast, indicating a high level of annotator agreement, while it stood at 0.31 for novelty, suggesting a relatively lower agreement. We attribute this discrepancy to the complexity of determining whether certain information constitutes a novel viewpoint or not [12].

Figure 8 shows the Mechanical Turk HIT given to Mechanical Turk workers to evaluate contrast while figure 9 shows the Mechanical Turk HIT given to Mechanical Turk workers to evaluate novelty.

C.3 Persuasiveness Evaluation Details and HIT Guidelines

To verify the clarity and efficacy of our instructions, we present workers with a set of 10 pairs selected from distinct topics. Five of these pairs exhibit significant differences in rationale form, including variations in length and level of detail, while the remaining five pairs are comparable in lengths. We intentionally provide easier examples to ensure that workers follow the guidelines. Annotators had perfect agreement for the set where rationales varied significantly. For the comparable rationale pairs, the interannotator reliability, as measured by Krippendorff's Alpha, reached 0.55. The interannotator reliability for the full set, reached 0.64. We use these annotation guidelines to obtain the final persuasion set, achieving a Krippendorff's Alpha score of 0.56.

Figure 10 shows the Mechanical Turk HIT for evaluating pairwise persuasiveness. Workers are prompted to choose between rationale 1, and rationale 2, or indicate that both are equally persuasive. Additionally, they are requested to provide 1-2 sentences as explanations for their decisions.

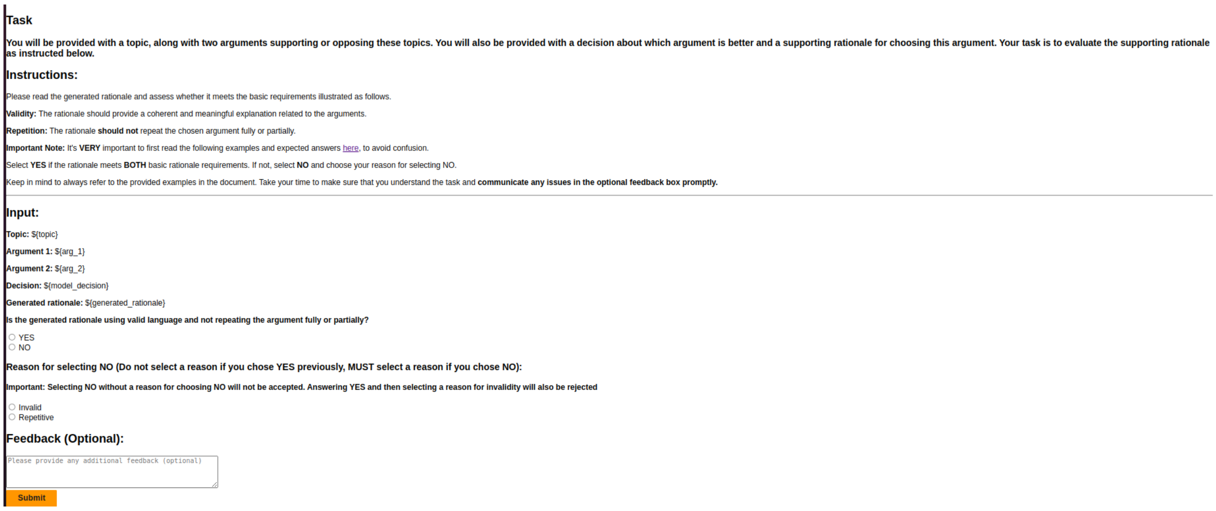
D Persuasion Evaluation with GPT4

Table 6 shows the components of the prompt we have used in pairwise persuasion ranking of the rationale.

E Characteristics of the generated rationale per model

Basic Form Figure 11 illustrates the percentage of rationales that failed to meet the basic form criteria across all models, along with the breakdown

[12] Experiments with random workers (with over 95% approval rate and over 5,000 approved HiTs) on the same subset yielded Krippendorff's alpha values of 0.17 and 0.18 for contrast and novelty, respectively. These findings emphasize the importance of our qualification process in obtaining reliable annotations.



Task

You will be provided with a topic, along with two arguments supporting or opposing these topics. You will also be provided with a decision about which argument is better and a supporting rationale for choosing this argument. Your task is to evaluate the supporting rationale as instructed below.

Instructions:

Please read the generated rationale and assess whether it meets the basic requirements illustrated as follows.

Validity: The rationale should provide a coherent and meaningful explanation related to the arguments.

Repetition: The rationale **should not** repeat the chosen argument fully or partially.

Important Note: It's **VERY** important to first read the following examples and expected answers here, to avoid confusion.

Select **YES** if the rationale meets **BOTH** basic rationale requirements. If not, select **NO** and choose your reason for selecting NO.

Keep in mind to always refer to the provided examples in the document. Take your time to make sure that you understand the task and **communicate any issues in the optional feedback box promptly.**

Input:

Topic: ${topic}

Argument 1: ${arg_1}

Argument 2: ${arg_2}

Decision: ${model_decision}

Generated rationale: ${generated_rationale}

Is the generated rationale using valid language and not repeating the argument fully or partially?

YES
NO

Reason for selecting NO (Do not select a reason if you chose YES previously, MUST select a reason if you chose NO):

Important: Selecting NO without a reason for choosing NO will not be accepted. Answering YES and then selecting a reason for invalidity will also be rejected

Invalid
Repetitive

Feedback (Optional):

Please provide any additional feedback (optional)

Submit

Figure 7: A screenshot form basic-form MTurk HIT for basic-form evaluation.

of reasons for failure between invalidity and repetition. The figure shows that Llama-2-7B and Llama-2-13B Chat predominantly generated invalid rationales, suggesting flaws in their reasoning capabilities regarding their choices. Conversely, models of similar sizes that underwent instruction tuning, namely Llama2-7B Chat, Llama2-13B Chat, Vicuna-7B, and Vicuna-13B, demonstrated proficiency in generating meaningful rationales. This emphasizes the significance of instruction tuning in rationalization. Notably, the common observation among samples failing to meet basic requirements was repetition, indicating a tendency among models to reiterate their chosen arguments partially or fully.

Content Evaluation Figure 12 reveals that, among all models, Llama2-70B Chat consistently provided rationales that justified not choosing the alternative argument (contrast). Similarly, GPT4 predominantly generated rationales characterized by contrast. However, the majority of rationales generated by other models did not offer justifications for not selecting the alternative argument.

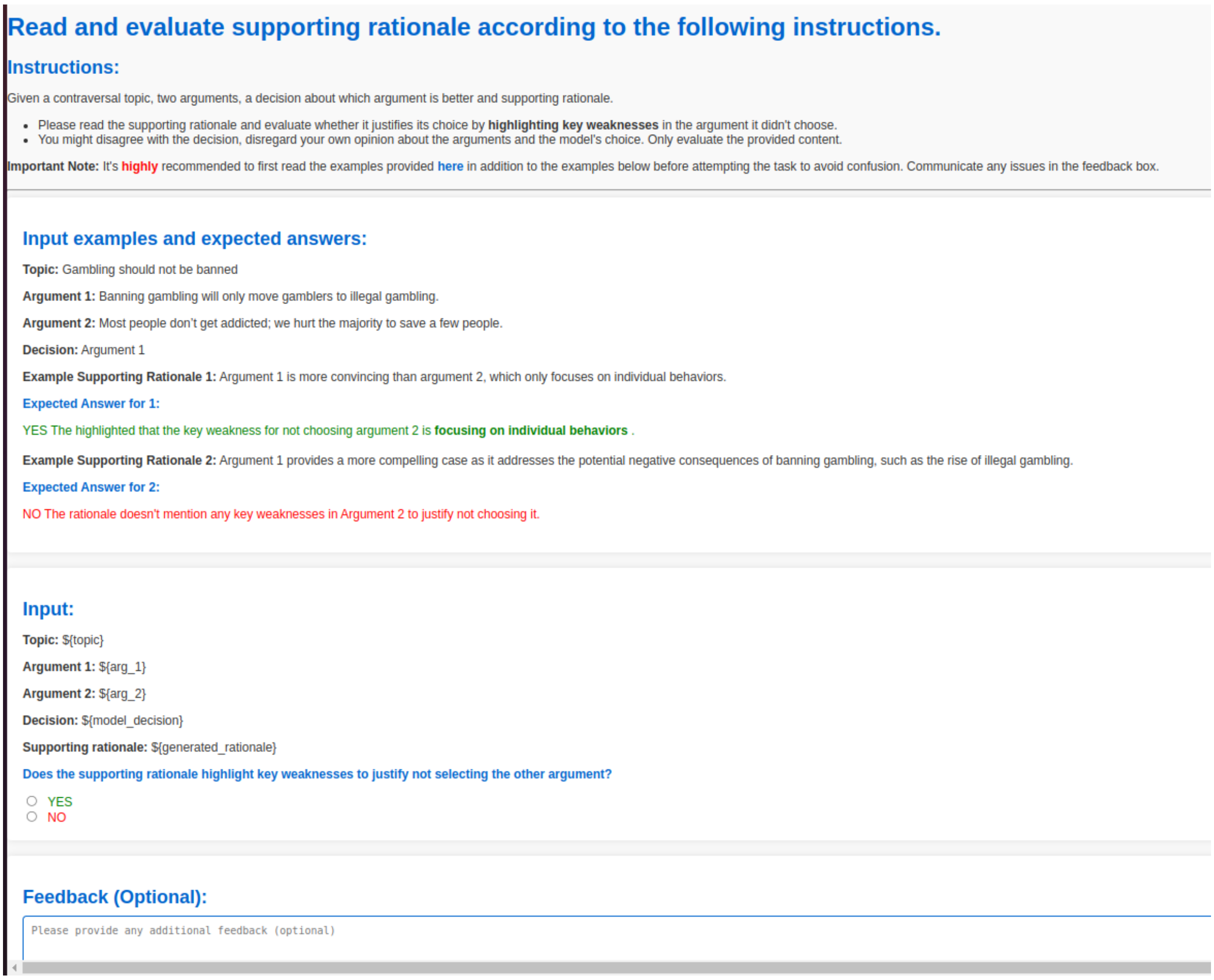
In analyzing novelty, it appears that the model scale, demonstrated by Llama2-70B, GPT4, and GPT-3.5-turbo, plays a role in enhancing the models' capacity to offer novel information in their generated rationales, beyond what is explicitly stated in the arguments.

F Characteristics of Models Capable of Generating Highly Persuasive Rationales

(1) **Instruction Tuning:** Among the models we analyzed, those that had not undergone instruction tuning (Llama2-7B and Llama2-13B) failed to provide valid rationales justifying the models' choices. This indicates that mere auto-regressive training is insufficient and that instruction tuning is essential for creating effective rationales. (2) **Scale:** The results also highlight that scaling up the parameters within the same model framework enhances persuasiveness. For example, Llama2-70B-chat was found to be more persuasive than its lower parameter counterparts, Llama2-13B-chat and Llama2-7B-chat. (3) **Further Tuning with Instructions Obtained from a Stronger LLM:** Vicuna models ranked higher compared to their Llama2 counterparts in the case of the IBM-ArgQ-9.1kPairs dataset, while Vicuna-13B consistently ranked higher on average compared to Llama2-7B-chat and Llama2-13B-chat in terms of the IBM-30k-rank dataset. This suggests that further instruction tuning, based on more advanced models, can improve a model's capability to generate more compelling rationales.

G Rationale Persuasiveness Improvement

Re-prompt the LLM Table 7 displays the prompt used to instruct LLMs to generate a more persuasive rationale. The model was prompted to compose 2 sentences supporting the chosen argument and 2 sentences indicating reasons for not choosing the alternative argument. This approach



Read and evaluate supporting rationale according to the following instructions.

Instructions:

Given a contraversal topic, two arguments, a decision about which argument is better and supporting rationale.

- Please read the supporting rationale and evaluate whether it justifies its choice by **highlighting key weaknesses** in the argument it didn't choose.
- You might disagree with the decision, disregard your own opinion about the arguments and the model's choice. Only evaluate the provided content.

Important Note: It's **highly** recommended to first read the examples provided **here** in addition to the examples below before attempting the task to avoid confusion. Communicate any issues in the feedback box.

Input examples and expected answers:

Topic: Gambling should not be banned

Argument 1: Banning gambling will only move gamblers to illegal gambling.

Argument 2: Most people don't get addicted; we hurt the majority to save a few people.

Decision: Argument 1

Example Supporting Rationale 1: Argument 1 is more convincing than argument 2, which only focuses on individual behaviors.

Expected Answer for 1:

YES The highlighted that the key weakness for not choosing argument 2 is **focusing on individual behaviors** .

Example Supporting Rationale 2: Argument 1 provides a more compelling case as it addresses the potential negative consequences of banning gambling, such as the rise of illegal gambling.

Expected Answer for 2:

NO The rationale doesn't mention any key weaknesses in Argument 2 to justify not choosing it.

Input:

Topic: ${topic}

Argument 1: ${arg_1}

Argument 2: ${arg_2}

Decision: ${model_decision}

Supporting rationale: ${generated_rationale}

Does the supporting rationale highlight key weaknesses to justify not selecting the other argument?

○ YES
○ NO

Feedback (Optional):

Please provide any additional feedback (optional)

Figure 8: A screenshot form MTurk HIT for contrast evaluation.

ensures that the model includes *contrast* and sufficient detail in its rationalization, which has been shown to enhance persuasiveness.

Evaluate and Refine Table 8 shows the prompt used in the *evaluate and refine* method to let the LLM decide if it needs to improve its rationale persuasiveness or not.

Examples of the improved rationales Table 9 presents examples from various improvement methods. Notably, the evaluate and refine method indicated that the original rationale was sufficiently persuasive, suggesting a possible tendency of models to concur with the persuasiveness of their prior outputs. Conversely, when the LLM is explicitly instructed on how to construct a persuasive rationale, the content changes accordingly.

GPT4 pairwise persuasion ranking Prompt

Task definition You will be presented with a topic and two arguments, labeled as "ARG1" and "ARG2." One of these arguments, either "ARG1" or "ARG2," is identified as the winner argument ("WINNER_ARG"). Additionally, two different rationales supporting the winner argument are provided, each indicated by a numerical identifier [1] or [2]. Your task is to *determine which rationale is more persuasive or if they are equally persuasive in supporting the "WINNER_ARG".*

Formatting Examples

```
{
//Three formatting examples for each type of output.
// Actual formatting examples are truncated to save pace.
}
```

Annotation Example

```
    {
    //Actual input
    }
    Think step by step then decide.
```

Table 6: GPT4 based persuasion ranking prompt.

GPT4 pairwise persuasion ranking Prompt

Task definition You will receive two arguments, each identified by a numerical identifier [] and a Topic. Disregarding your own opinion on the topic, given the two arguments, decide which argument you would recommend. **Provide a compelling reasoning consists of 2 sentences justifying the argument you choose and 2 sentences stating your reasoning for not choosing the other argument.** Choose [1] if you if you recommend argument [1] over argument [2]. Choose [2] if you recommend argument [2] over argument [1]. Format your output in a JSON format with "decision" and "reasoning" keys.:

Formatting Examples

```
{
//Similar Examples to Table 4
// Actual formatting examples are truncated to save pace.
}
```

Annotation Example

```
    {
    //Actual input
    }
```

Table 7: Pairwise argument ranking with prompting LLM for a more persuasive rationalization. **Bolded** part are intended to prompt LLMs to generate contrastive rationales with enough details.

Evaluate and Refine Prompt

Task definition You will receive two arguments, each identified by a numerical identifier [] and a Topic. You will also receive a decission about which argument is better in the format of ARG1 or ARG2 in the format of "WINNER ARG". You will also receive a "supporting rationale" supporting the "WINNER ARG". Your task is to decide whether the supporting rationale is "convincing" in terms of supporting the "WINNER ARG". Choose "YES" if you think the rationale is convincing in supporting the "WINNER ARG". Choose "NO" if you think it's not convincing in supporting the "WINNER ARG". If you choose "YES", generate "no further improvement needed" as your "improved rationale" . If you choose "NO" generate a more persuasive version of the rationale.

Formatting Examples

```
{
//2 formatting examples to ensure the model follows them
// Actual formatting examples are truncated to save pace.
}
```

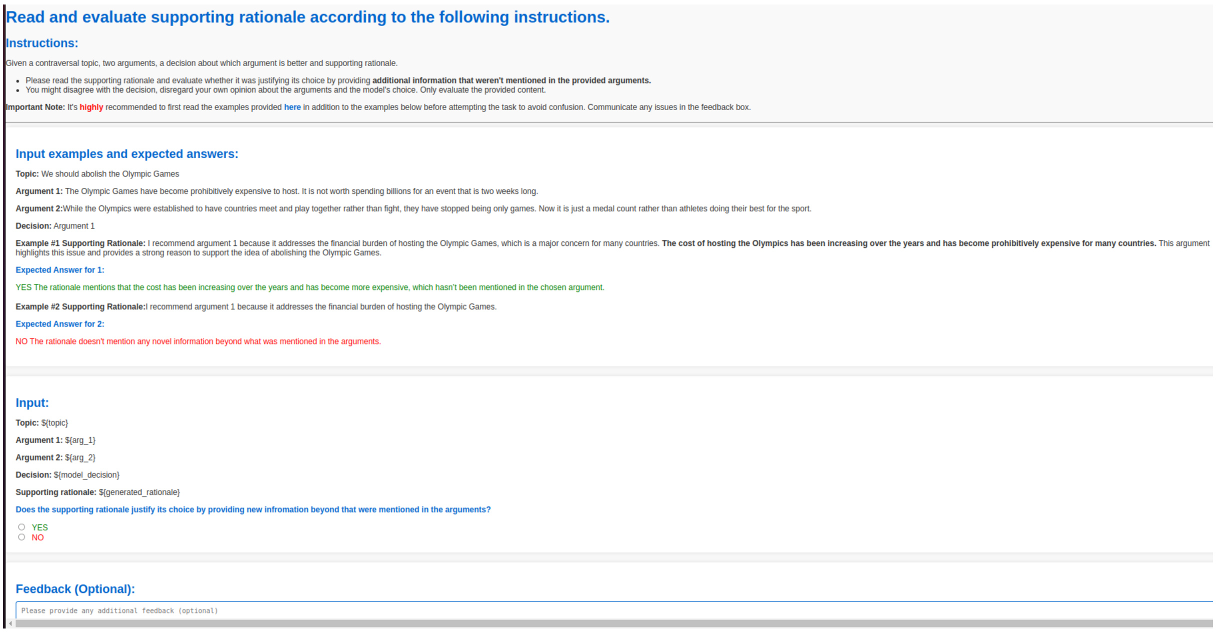
Annotation Example

```
    {
    //Actual input
    }
```

Table 8: Prompt used to instruct the LLM to decide first if it needs to improve the rationale persuasiveness , then improve its persuasiveness if necessary.

Inputs
Topic: We should promote autonomous cars **Argument 1:** Research into the development of fully autonomous vehicles will invariably produce breakthroughs in sensor technology, machine vision, and other related domains that will find application in other fields benefitting society. **Argument 2:** Autonomous vehicles will be safer especially trucks. People who say trucking is the most popular job in America are wrong and only a relatively few people will lose jobs. See `https://www.marketwatch.com/story/no-truck-driver-isnt-the-most-common-job-in-your-state-2015-02-12` **Decision:** Argument 1
Generated Rationale
Llama2-7B-chat: The argument that research into the development of fully autonomous vehicles will produce breakthroughs in related domains is more convincing because it highlights the potential for positive impact across multiple fields, rather than simply focusing on the safety benefits of autonomous vehicles.
Llama2-7B-chat-persuasion-prompted: While argument 2 mentions potential job losses in the trucking industry, the benefits of promoting autonomous cars far outweigh these potential drawbacks. The development of fully autonomous vehicles will lead to breakthroughs in sensor technology, machine vision, and other related domains, which will have a broader impact on society. Additionally, autonomous vehicles will be safer, including trucks, which will save lives and reduce accidents. Overall, argument 1 presents a more compelling case for promoting autonomous cars.
Llama2-7B-chat-second-stage The argument that research into the development of fully autonomous vehicles will produce breakthroughs in related domains is more convincing because it highlights the potential for positive impact across multiple fields, rather than simply focusing on the safety benefits of autonomous vehicles.

Table 9: Different rationales outputs by different methods.



Read and evaluate supporting rationale according to the following instructions.

Instructions:

Given a contraversal topic, two arguments, a decision about which argument is better and supporting rationale.

- Please read the supporting rationale and evaluate whether it was justifying its choice by providing **additional information that weren't mentioned in the provided arguments.**
- You might disagree with the decision, disregard your own opinion about the arguments and the model's choice. Only evaluate the provided content.

Important Note: It's **highly** recommended to first read the examples provided **here** in addition to the examples below before attempting the task to avoid confusion. Communicate any issues in the feedback box.

Input examples and expected answers:

Topic: We should abolish the Olympic Games

Argument 1: The Olympic Games have become prohibitively expensive to host. It is not worth spending billions for an event that is two weeks long.

Argument 2: While the Olympics were established to have countries meet and play together rather than fight, they have stopped being only games. Now it is just a medal count rather than athletes doing their best for the sport.

Decision: Argument 1

Example #1 Supporting Rationale: I recommend argument 1 because it addresses the financial burden of hosting the Olympic Games, which is a major concern for many countries. **The cost of hosting the Olympics has been increasing over the years and has become prohibitively expensive for many countries.** This argument highlights this issue and provides a strong reason to support the idea of abolishing the Olympic Games.

Expected Answer for 1:

YES The rationale mentions that the cost has been increasing over the years and has become more expensive, which hasn't been mentioned in the chosen argument.

Example #2 Supporting Rationale:I recommend argument 1 because it addresses the financial burden of hosting the Olympic Games.

Expected Answer for 2:

NO The rationale doesn't mention any novel information beyond what was mentioned in the arguments.

Input:

Topic: ${topic}

Argument 1: ${arg_1}

Argument 2: ${arg_2}

Decision: ${model_decision}

Supporting rationale: ${generated_rationale}

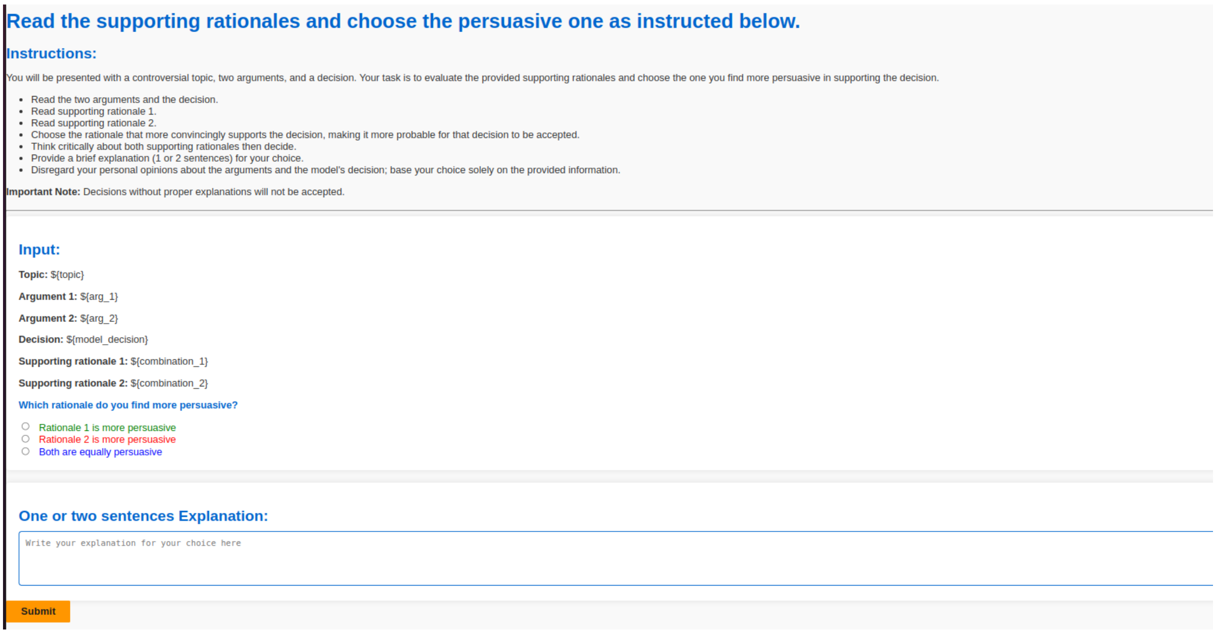
Does the supporting rationale justify its choice by providing new infromation beyond that were mentioned in the arguments?

- YES
- NO

Feedback (Optional):

Please provide any additional feedback (optional)

Figure 9: A screenshot form MTurk HIT for novelty evaluation.



Read the supporting rationales and choose the persuasive one as instructed below.

Instructions:

You will be presented with a controversial topic, two arguments, and a decision. Your task is to evaluate the provided supporting rationales and choose the one you find more persuasive in supporting the decision.

- Read the two arguments and the decision.
- Read supporting rationale 1.
- Read supporting rationale 2.
- Choose the rationale that more convincingly supports the decision, making it more probable for that decision to be accepted.
- Think critically about both supporting rationales then decide.
- Provide a brief explanation (1 or 2 sentences) for your choice.
- Disregard your personal opinions about the arguments and the model's decision; base your choice solely on the provided information.

Important Note: Decisions without proper explanations will not be accepted.

Input:

Topic: ${topic}

Argument 1: ${arg_1}

Argument 2: ${arg_2}

Decision: ${model_decision}

Supporting rationale 1: ${combination_1}

Supporting rationale 2: ${combination_2}

Which rationale do you find more persuasive?

- Rationale 1 is more persuasive
- Rationale 2 is more persuasive
- Both are equally persuasive

One or two sentences Explanation:

Write your explanation for your choice here

Submit

Figure 10: A screenshot form MTurk HIT for persuasion evaluation.

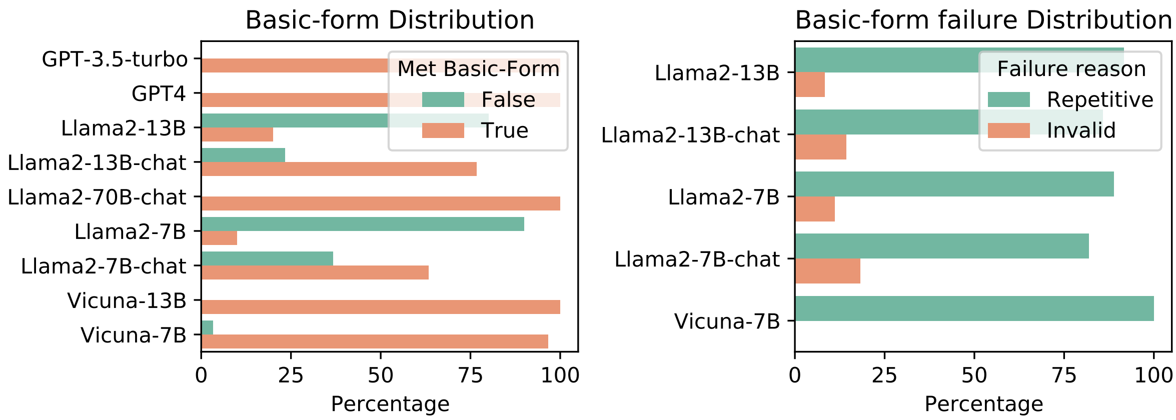


Figure 11: Basic-form distribution across models

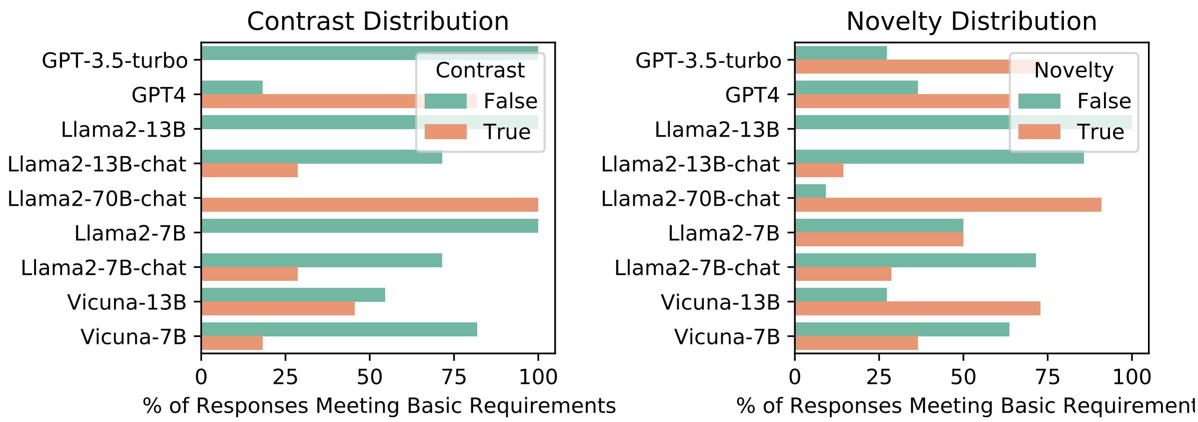


Figure 12: Contrast and Novelty distribution among models for samples met basic-form requirements.