

A Three-Level Audit of LLM Alignment for Argument Quality Assessment

Wei-Fan Chen^{*1,2,3} Jinming Yu^{*2} Lucie Flek^{2,3}

¹National Institute of Advanced Industrial Science and Technology (AIST), Japan

²Bonn-Aachen International Center for Information Technology, University of Bonn, Germany

³Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
wf.chen@aist.go.jp, s6jiyuu@uni-bonn.de, flek@bit.uni-bonn.de

Abstract

Large Language Models (LLMs) are increasingly used as automated evaluators of argument quality. However, existing studies typically assess models only through their agreement with human scores, leaving the reasoning process behind these judgments unexplored. In this paper, we propose a three-level audit framework for evaluating the reliability of LLM-based argument quality assessment. The framework distinguishes between (1) **surface alignment**, measuring agreement between LLM-predicted scores and human annotations; (2) **instructional alignment**, assessing whether generated rationales adhere to the intended evaluation criteria; and (3) **faithfulness alignment**, examining whether predicted scores are supported by the generated rationales. To operationalize this audit, we introduce structural rationale prompting,¹ which guides LLMs to generate structured justifications before assigning scores across 11 dimensions of the Dagstuhl-15512 argument quality corpus. We evaluate several LLMs under this framework and find that structural rationale prompting substantially improves agreement with human annotations compared to definition-based prompting. Furthermore, the generated rationales generally follow the evaluation instructions and remain highly consistent with the predicted scores. Overall, our results suggest that auditing LLM evaluators beyond surface score agreement provides deeper insight into the reliability and transparency of LLM-based argument quality assessment.

1 Introduction

The assessment of argument quality remains a foundational challenge in argument mining (Habernal and Gurevych, 2016; Wachsmuth et al., 2017; Lawrence and Reed, 2019). While the theoretical roots of argumentation trace back to Ancient

Greece (Aristotle, 2007), operationalizing these concepts in computational settings requires taxonomies that decompose argument quality into multiple measurable dimensions (Blair, 2011; Wachsmuth et al., 2017). Annotating these dimensions, however, is a labor-intensive and time-consuming process that often requires specially trained annotators and carefully designed guidelines (Toledo et al., 2019; Lauscher et al., 2020). As a result, expert-labeled datasets remain limited, creating a significant bottleneck for training automatic argument quality evaluators and for advancing downstream tasks in argument mining.

Recent advances in Large Language Models (LLMs) have prompted researchers to investigate whether such models can follow natural language instructions to assist with data annotation and evaluation tasks (Gilardi et al., 2023; Chiang and Lee, 2023; Mirzakhmedova et al., 2024). Experimental results across several studies report substantial agreement between LLM judges and human judges, suggesting that LLMs may serve as scalable evaluators in various NLP settings (Zheng et al., 2023; Kocmi and Federmann, 2023). However, these models are typically employed as *black-box evaluators*, whose reliability is assessed primarily through agreement with human scores. Even when reasoning-oriented prompting strategies such as Chain-of-Thought prompting (Wei et al., 2022) are applied, it often remains unclear whether the generated explanations follow the intended evaluation criteria or meaningfully justify the final scores produced by the model. Consequently, the reasoning processes underlying LLM-based evaluations remain largely under-examined.

In this work, we argue that evaluating LLM-based argument quality assessment requires auditing not only predicted scores but also the reasoning processes that lead to them. In practice, an LLM evaluator may fail in several ways: its predicted scores may diverge from human judg-

¹All the codes and prompts used in the paper are available at <https://github.com/aist-cwf/argmining26-argument-quality-assessment>.

*Equal Contribution.

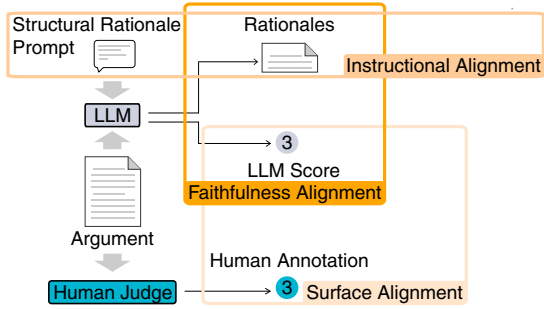


Figure 1: Overview of the proposed three-level audit framework for LLM-based argument quality assessment. Given an argument and an evaluation prompt, the LLM produces rationales and a predicted score. Surface alignment measures agreement between the LLM score and human annotations. Instructional alignment evaluates whether the generated rationales follow the evaluation criteria. Faithfulness alignment examines whether the predicted score is supported by the generated rationales.

ments, its explanations may not follow the intended evaluation criteria, or its reasoning may not actually support the predicted scores. To systematically analyze these potential failure modes, we propose a three-level audit framework for LLM-based argument quality assessment. As illustrated in Figure 1, the framework distinguishes three complementary alignment dimensions: surface alignment, which measures agreement between LLM-predicted scores and expert human annotations; instructional alignment, which evaluates whether generated rationales adhere to the intended evaluation instructions; and faithfulness alignment, which examines whether predicted scores are supported by the generated rationales. These dimensions guide our investigation of the reliability and transparency of LLM-based argument evaluation through the following research questions.

- (1) Surface Alignment: To what extent do LLM scores correlate with human annotations across diverse argument quality dimensions?
- (2) Instructional Alignment: To what extent do the rationales generated by LLMs adhere to the prompted instructions and criteria defined for each argument quality dimension?
- (3) Faithfulness Alignment: To what extent are the predicted scores supported by the generated rationales?

To operationalize this audit, we introduce structural rationale prompting, a prompting strategy that guides LLMs to generate step-wise rationales

based on predefined evaluation criteria before assigning quality scores. We apply this approach to assess argument quality across all 11 dimensions of the Dagstuhl-15512 argument quality corpus (Wachsmuth et al., 2017). Using the generated rationales and predicted scores, we analyze the alignment between LLM outputs and human judgments across the three alignment dimensions defined above.

The first research question examines whether LLMs can reliably annotate argument quality in a manner comparable to human experts. To address this question, we evaluate the statistical alignment between LLM-predicted scores and expert human annotations, measuring agreement using Krippendorff’s α and prediction error using RMSE. We further analyze whether generating structured rationales improves agreement between LLMs and human experts. Among the evaluated models, *mistral-small-latest* achieves the strongest performance (average $\alpha = 0.80$ and average RMSE = 0.63). Moreover, structural rationale prompting yields substantially higher agreement ($\alpha = 0.67$ vs. 0.28, both using GPT-3.5-Turbo) than the “expert prompt” used in Mirzakhmedova et al. (2024), which provides only definitions of the argument quality dimensions.

While the first research question evaluates alignment between LLM-predicted scores and human annotations, the second examines the quality of the generated rationales. Because the Dagstuhl-15512 dataset does not provide gold-standard justifications, we assess whether LLM-generated rationales adhere to the structured evaluation instructions. We validate an independent judge LLM (GPT-5.2) against human expert annotations on a small subset, yielding a Krippendorff’s α of 0.56, and then use it to evaluate the remaining rationales. On a 3-point scale, the results show consistently high levels of instructional alignment across the four evaluated models, with scores ranging from 2.00 (Mistral-7B-Instruct) to 2.79 (Mistral-small-latest).

The third research question investigates whether the models’ predicted scores are supported by their stated rationales. To assess this relationship, annotators are asked to infer argument quality scores based solely on the generated rationales, without access to the original model predictions. These rationale-based scores are then compared with the model-predicted scores to derive a measure of faithfulness. As in the previous evaluation, we validate

an independent judge LLM (GPT-5.2) against human expert annotations on a small subset (Krippendorff’s $\alpha = 0.56$) before applying it to the remaining instances. On a 3-point scale, the results indicate consistently high faithfulness alignment across all evaluated models, with scores ranging from 2.80 (Mistral-7B-Instruct and GPT-3.5-Turbo) to 2.96 (Llama-3.3-70B-Instruct).

Our contributions are as follows: (1) We introduce a reproducible auditing framework for analyzing the reliability of LLM-based argument quality evaluation through three alignment dimensions: surface alignment, instructional alignment, and faithfulness alignment. (2) We propose structural rationale prompting, a prompting strategy that guides LLMs to generate structured justifications based on predefined argument quality criteria before assigning scores. (3) Through experiments on the Dagstuhl-15512 argument quality corpus, we show that structural rationale prompting improves alignment with human annotations and that modern LLMs generate rationales that largely follow evaluation instructions while remaining highly consistent with their predicted scores.

2 Related Work

This section first reviews the evolution of argument quality assessment from theory-grounded taxonomies to neural scoring models. We then examine the emerging literature on LLM Interpretability, focusing on generating faithful rationales within structured reasoning frameworks.

Argument Quality Assessment The computational argument quality assessment has evolved across three distinct paradigms: theory-driven feature engineering, neural representation learning, and the current era of generative modeling.

Early work in argument quality assessment was heavily grounded in classical rhetoric and informal logic. Following the Aristotelian tradition, researchers initially focused on identifying the components of an argument using the Toulmin model (Toulmin, 2003) or Walton’s argumentation schemes (Walton et al., 2008). Habernal and Gurevych (2016) pioneered the *convincingnes task*, using crowdsourcing to determine which of two arguments is more persuasive. This was further refined by Wachsmuth et al. (2017), who established the foundational taxonomy of 11 fine-grained quality dimensions, providing the Dagstuhl-15512 corpus which remains the gold standard for multi-

dimensional assessment.

With the rise of deep learning, the field shifted toward neural models that could capture latent linguistic features. Potash et al. (2017) and Stab and Gurevych (2017) explored the use of Recurrent Neural Networks for scoring argument persuasiveness and structure. The introduction of Transformer-based models like BERT (Devlin et al., 2019) significantly improved performance on quality scoring tasks, as seen in the work of Gretz et al. (2020), who released a large-scale dataset of 30k arguments scored for quality. Subsequent research explored cross-domain generalizability (Toledo et al., 2019) and the use of graph neural networks to model the relational structure between conflicting arguments (Ye and Teufel, 2021).

Most recently, the paradigm has shifted toward using LLMs as judges. For example, Stahl et al. (2025) fine-tuned an LLM via the proposed specialized instruction in various computational argument tasks including quality assessment. In fact, while models like GPT-4 have demonstrated high correlation with human scores in general text evaluation (Chiang and Lee, 2023), their application to the specific nuances of argumentation is an ongoing area of research. Liu et al. (2023) introduced G-Eval, which uses Chain-of-Thought to improve evaluation consistency. However, as noted by Wang et al. (2023), LLM evaluators are susceptible to self-preference bias and hallucinated justifications. Our work builds on this line of inquiry by moving beyond simple scoring to audit the faithfulness of the rationales generated.

Rationale Generation and LLM Faithfulness

The emergence of Large Language Models has shifted focus from purely discriminative tasks to generative reasoning. The introduction of Chain-of-Thought prompting by Wei et al. (2022) demonstrated that forcing a model to generate intermediary reasoning steps significantly enhances performance on complex tasks. In the domain of evaluation, Liu et al. (2023) and Chan et al. (2023) found that eliciting these rationales leads to higher alignment with human-assigned scores. However, the nature of these rationales remains a subject of intense scrutiny. Researchers have debated whether these outputs represent a process-oriented logic or merely a result-oriented justification (Zelikman et al., 2022; Lightman et al., 2023).

A critical concern in LLM interpretability is faithfulness: the degree to which a generated

Quality Dimension	Definition
Cogency	
Local acceptability	Premises worthy of being believed.
Local relevance	Premises accept/reject conclusion.
Local sufficiency	Premises enough to draw conclusion.
Effectiveness	
Credibility	Makes author worthy of credence.
Emotion appeal	Makes audience open to arguments.
Clarity	Uses correct, unambiguous language and avoids deviation from the issue.
Appropriateness	Language supports credibility and emotions, in proportion to the issue.
Arrangement	Argues in the right order.
Reasonableness	
Global acceptability	Audience accepts use of argument.
Global relevance	Argument helps arrive at conclusion.
Global sufficiency	Enough rebuts counterarguments.

Table 1: Definition and categorization of each quality dimension, adapted from Wachsmuth et al. (2017) and Mirzakhmedova et al. (2024).

rationale accurately reflects the model’s internal decision-making process. Jacovi and Goldberg (2020) established the formal criteria for faithfulness, distinguishing it from mere plausibility (how human-like an explanation sounds). Chen et al. (2025) examined faithfulness hallucination by creating a heatmap-like “Halumap,” and Chen et al. (2024) tried to generate more faithful outputs based on a multi-task learning approach. Recent audits have revealed a disturbing trend of post-hoc rationalization, where models generate convincing but fundamentally disconnected explanations for their outputs (Turpin et al., 2023).

To move beyond black-box behavior, several studies have proposed methods to audit the causal link between rationales and outputs. Lyu et al. (2023) explored the use of faithful chain-of-thoughts by enforcing symbolic intermediate steps. Meanwhile, the use of Judge LLMs to evaluate the quality and logic of another model’s reasoning has become a standard diagnostic tool (Zheng et al., 2023). Our study contributes to this line of work by applying these “faithfulness audits” to the specific, highly subjective domain of 11-dimension argumentation quality.

3 Experimental Setup

This section presents the experimental setup that underlies the three alignment research questions investigated in this study. We begin by introducing the argument quality assessment task, followed by descriptions of the dataset, the prompting strategies used to elicit model judgments and rationales, and

Quality Dimension	#Instances
Cogency	
Local acceptability	82
Local relevance	99
Local sufficiency	113
Effectiveness	
Credibility	115
Emotion appeal	130
Clarity	89
Appropriateness	53
Arrangement	81
Reasonableness	
Global acceptability	96
Global relevance	66
Global sufficiency	136

Table 2: The number of instances in each dimension where there are perfect agreement among the annotators.

the LLMs evaluated in our experiments.

3.1 Task

Following the argument quality taxonomy proposed by Wachsmuth et al. (2017), our task is to assess arguments along 11 fine-grained dimensions of argument quality. These dimensions capture different aspects of argumentative strength and effectiveness. Table 1 summarizes the dimensions together with their definitions.

3.2 Dataset

Based on this taxonomy, Wachsmuth et al. (2017) constructed the Dagstuhl-15512 argument quality corpus by annotating 320 arguments drawn from the UKPConvArgRank dataset (Habernal and Gurevych, 2016). Each argument was evaluated by three expert annotators with respect to the defined quality dimensions. The reported inter-annotator agreement, measured using Krippendorff’s α , ranges from 0.23 to 0.60 across dimensions. Annotators assigned scores on a 3-point Likert scale (1 = low, 2 = medium, 3 = high) or selected *cannot judge*, which occurred only once among approximately 14,000 individual assessments.

To obtain a single reference label for comparison with LLM predictions, we aggregate the scores from the three annotators using MACE (Hovy et al., 2013).² MACE estimates the reliability of each annotator and infers a latent true label based on these reliability estimates, thereby providing a principled aggregation of multiple annotations.

We also follow Mirzakhmedova et al. (2024) in

²We also experimented with simple score averaging and observed very similar results.

Task: Evaluate the quality of the Local Sufficiency of the given argument in the context of a debate forum. The input includes an issue, a stance, and a corresponding argument. Your goal is to think step-by-step about the structural flow of the argumentation and assign a quality score for **Local Sufficiency** (from 1 to 3).

Input

- **Issue**: ISSUE
- **Stance**: STANCE
- **Argument**: ARGUMENT

Definition of Local Sufficiency: The premises of an argument should be seen as sufficient if, together, they provide enough support to make it rational to draw the argument’s conclusion. If you identify more than one conclusion in the comment, try to adequately weight the sufficiency of the premises for each conclusion when judging about their “aggregate” sufficiency—unless there are particular premises or conclusions that dominate your view of the author’s argumentation. Notice that you may see premises as sufficient even though you do not personally accept all of them, i.e., sufficiency does not presuppose acceptability.

Step-by-Step Instructions:

- (1) Identify the conclusion — Determine the main claim that the author is trying to prove or support.
- (2) List the premises — Extract the key reasons or pieces of evidence offered to support the conclusion. Ensure they are clearly separable and relevant.
- (3) Evaluate logical connections — Analyze whether the premises logically and directly support the conclusion without major gaps or irrelevant claims.
- (4) Assess premise quality and evidence support — Check if the premises are factually sound, supported by evidence, and free from bias or unsupported assumptions.
- (5) Make a global judgment
 - Score 3 (High) — Premises are relevant, coherent, and collectively sufficient to justify the conclusion.
 - Score 2 (Medium) — Premises are partially sufficient but lack depth or contain minor gaps in reasoning.
 - Score 1 (Low) — Premises are weak, biased, or fail to support the conclusion logically.

Examples:

- Argument: *example_argument*
 - Issue: *example_issue*
 - Stance: *example_stance*
 - Reasoning: *example_reasoning*
 - Final score: *example_score*
-

Table 3: The structural rationale prompt for assessing the *local sufficiency* quality dimension. The teletype tokens are the placeholders for the argument to be evaluated. Two examples for few shots learning are provided in the end but are omitted here due to space constraints. The full example and the model outputs can be found in Sec A.1.

constructing a *perfect-agreement subset*, consisting of instances for which all annotators assigned the same score. This subset represents cases with unambiguous human judgments and therefore provides a useful setting for analyzing model behavior under high annotation certainty. The number of instances in the perfect-agreement subset is reported in Table 2.

3.3 Prompt Design

Prompt design for LLMs is known to be highly sensitive, as even small variations in wording can lead to substantially different outputs. Rather than focusing on prompt engineering for optimal performance, this work introduces a structural rationale prompting framework designed to guide language models through a structured reasoning process before assigning argument quality scores. The goal of this framework is to encourage models to approximate the reasoning procedure of expert annotators when evaluating arguments.

Structural Rationale Prompt Table 3 illustrates the structural rationale prompt using the *local suffi-*

ciency dimension as an example. The prompt first provides the issue, stance, and argument, followed by the definition of local sufficiency taken from the annotation guidelines in Wachsmuth et al. (2017). Based on this definition and prior computational approaches to modeling the dimension, we formulate step-by-step instructions that guide the model in producing a structured rationale prior to assigning a score.

For the local sufficiency dimension, both the definition and prior computational work (Rahimi et al., 2014) suggest that evaluation requires several reasoning steps: (1) identifying the premises and the conclusion of the argument, (2) assessing the logical support provided by each premise for the conclusion, and (3) aggregating these assessments to determine whether the premises are collectively sufficient. Based on this reasoning procedure, we formulate the structured rational prompt shown in the example prompt. Importantly, our goal is not to claim that these prompts represent an optimal design for argument quality annotation. Instead, they serve as an analytical tool for examining whether

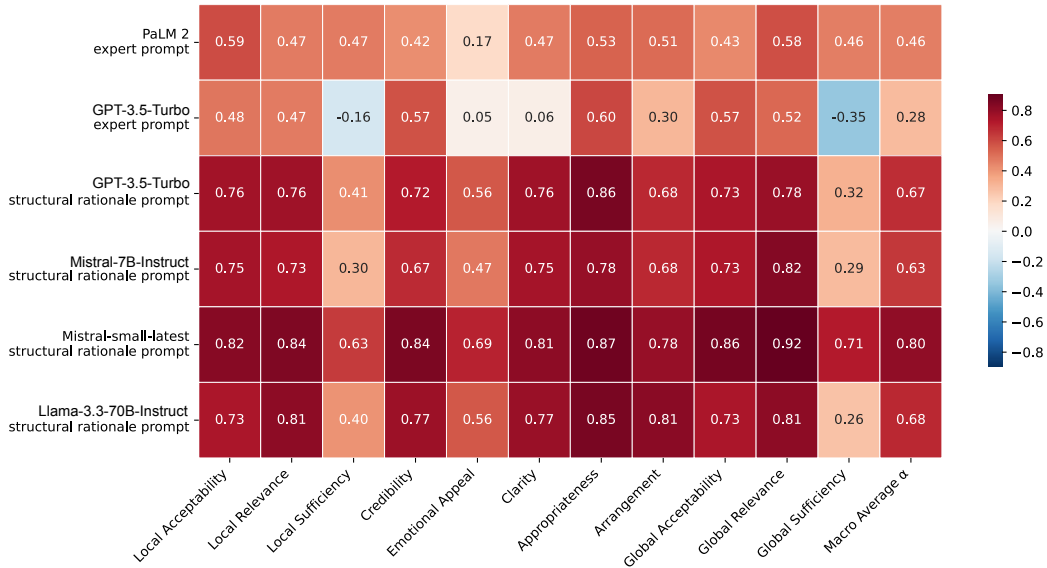


Figure 2: Krippendorff’s α measuring agreement between human annotations and LLM predictions across models and prompting strategies. Results for the expert prompt are taken from [Mirzakhmedova et al. \(2024\)](#). Structural rationale prompting consistently improves agreement across the evaluated dimensions.

LLMs follow the intended evaluation procedure and whether their predicted scores are supported by the generated rationales.

All experiments are conducted three times, and we report the average performance across runs. The few-shot examples used in the prompts are sampled from the Dagstuhl-15512 corpus and are excluded from the corresponding evaluation sets to avoid overlap between demonstration and test instances.

Expert Prompt To enable comparison with the results reported by [Mirzakhmedova et al. \(2024\)](#), we also evaluate the so-called “expert prompt.” This prompting strategy provides the LLM with only the definition of each argument quality dimension taken from the annotation guidelines, without additional step-by-step instructions.

3.4 Models

As our evaluation setup, we prompt several LLMs with varying architectures and model scales: (1) Llama-3.3-70B-Instruct, (2) GPT-3.5-Turbo, a large proprietary model, (3) Mistral-7B-Instruct, and (4) Mistral-small-latest (approximately 24B parameters). These models cover both open-weight and commercial systems and represent a range of model capacities. To ensure comparability across models, we use the same prompts and experimental settings ($temperature = 0.7$, and $top_p = 0.9$) for all models, allowing us to better isolate the effects of model scale and architecture on argument

quality assessment.

4 Surface Alignment

The first research question examines surface alignment, which measures **the degree of agreement between LLM-predicted argument quality scores and human annotations**. Our goal is to determine whether structural rationale prompting improves annotation performance compared with the prompting strategy used in prior work. To this end, we compare our model predictions with the results reported by [Mirzakhmedova et al. \(2024\)](#) on the perfect-agreement subset, where all human annotators assigned the same score and the ground truth is therefore unambiguous. Figure 2 reports Krippendorff’s α between human annotations and LLM predictions across models and prompting strategies. Overall, structural rationale prompting leads to consistently higher agreement across the evaluated dimensions. Notably, despite having fewer parameters than GPT-3.5-Turbo and Llama-3.3-70B-Instruct, Mistral-small-latest achieves the strongest overall agreement.

To complement the agreement analysis, Table 4 reports the RMSE of the four models with respect to the aggregated reference labels on the whole dataset. The results largely corroborate the findings from Krippendorff’s α : Mistral-small-latest achieves the lowest overall prediction error across the evaluated dimensions. Among the remaining

Models	LA	LR	LS	Cr	Em	Cl	Ap	Ar	GA	GR	GS	Avg.
Llama-3.3-70B-Instruct	0.74	0.73	0.71	0.64	0.76	0.73	0.73	0.76	0.61	0.71	0.77	0.72
GPT-3.5-Turbo	0.78	0.83	0.91	0.77	0.79	0.80	0.79	0.86	0.73	0.89	0.66	0.80
Mistral-7B-Instruct	0.70	0.88	0.82	0.80	0.85	0.75	0.72	0.87	0.78	0.79	0.67	0.78
Mistral-small-latest 2-shots	0.73	0.71	0.71	0.66	0.70	0.67	0.67	0.74	0.66	0.68	0.58	0.68
Mistral-small-latest 0-shots	0.85	0.79	0.74	0.69	0.79	0.69	0.75	0.84	0.67	0.72	0.80	0.74

Table 4: RMSE of the four models and `mistral-small-latest` under zero-shot setting. The abbreviations denote the following dimensions: **L**ocal **A**cceptability, **L**ocal **R**elevance, **L**ocal **S**ufficiency, **C**redibility, **E**motional Appeal, **C**larity, **A**ppropriateness, **A**rrangement, **G**lobal Acceptability, **G**lobal **R**elevance, and **G**lobal **S**ufficiency. The best-performing setting for each dimension is highlighted in bold.

Models	LA	LR	LS	Cr	Em	Cl	Ap	Ar	GA	GR	GS	Avg.
Llama-3.3-70B-Instruct	2.54	2.93	2.98	3.00	2.40	3.00	3.00	2.96	2.46	2.94	2.17	2.76
GPT-3.5-Turbo	1.98	1.93	2.00	2.14	2.25	1.99	2.32	2.77	2.29	2.00	2.25	2.17
Mistral-7B-Instruct	1.93	1.92	2.02	2.01	2.00	1.97	2.06	2.33	2.02	2.00	1.72	2.00
Mistral-small-latest	2.84	2.99	3.00	2.70	3.00	2.80	2.62	2.98	2.32	3.00	2.49	2.79

Table 5: Average instructional alignment scores of the generated rationales. Dimension abbreviations follow those defined in Table 4. The best-performing model for each dimension is highlighted in bold.

models, Llama-3.3-70B shows the second-best performance and even achieves the lowest RMSE on the dimensions of local sufficiency and credibility. This suggests that larger models can still outperform smaller ones on specific dimensions, although the overall trend favors Mistral-small-latest.

We further examine whether providing demonstration examples improves annotation performance. The last row in Table 4 shows the performance of Mistral-small-latest under zero-shot setting. Across all evaluated dimensions, the two-shot configuration consistently yields lower RMSE than the zero-shot setting. This result indicates that including a small number of annotated examples helps the model better calibrate its scoring behavior, leading to more accurate predictions.

Overall, the results indicate that structural rationale prompting substantially improves the alignment between LLM predictions and human annotations. Across multiple models and evaluation metrics, the prompting framework consistently increases agreement and reduces prediction error, suggesting that guiding LLMs through structured reasoning can enhance their reliability as automated evaluators of argument quality.

5 Instructional Alignment

The second research question investigates instructional alignment, focusing on the quality of the rationales generated by LLMs. In particular, we examine whether the models **follow the structured evaluation instructions provided in the prompts**.

Score Difference	Fully C.	Partially C.	Non-C.
$\Delta = 0$	575	149	2
$\Delta = 1$	255	60	1
$\Delta = 2$	12	4	1

Table 6: Distribution of instances across instructional alignment levels (*fully compliant*, *partially compliant*, *non-compliant*) together with the corresponding prediction errors of Mistral-small-latest.

To this end, we randomly sample five arguments for each quality dimension and generate rationales using the four models described in Section 3.4. A total of 5 samples x 11 dimensions x 4 models = 220 annotations were conducted. Two human experts from the authors then evaluate how well each generated rationale follows the given instructions using a three-point Likert scale (1 = non-compliant, 2 = partially compliant, 3 = fully compliant). In addition to human evaluation, we prompt an independent LLM (GPT-5.2) to perform the same assessment. The two human annotators achieve a Krippendorff’s α of 0.69, indicating substantial agreement. When the LLM judge is included as a third annotator, the overall inter-annotator agreement is 0.56, suggesting moderate agreement.

Table 5 reports the instructional alignment scores assigned by GPT-5.2 across all quality dimensions. Overall, the results indicate that the evaluated models generally follow the structured instructions well. Several models even achieve perfect scores in specific dimensions—for example, Mistral-small-latest obtains full compliance in local sufficiency,

Models	LA	LR	LS	Cr	Em	Cl	Ap	Ar	GA	GR	GS	Avg.
Llama-3.3-70B-Instruct	2.87	2.95	2.99	3.00	3.00	2.96	2.91	2.99	2.98	2.98	2.99	2.96
GPT-3.5-Turbo	2.77	2.74	2.80	2.70	2.84	2.76	2.79	2.77	2.92	2.85	2.91	2.80
Mistral-7B-Instruct	2.70	2.71	2.80	2.82	2.81	2.91	2.75	2.96	2.95	2.85	2.58	2.80
Mistral-small-latest	2.79	2.93	2.99	2.93	2.98	2.99	2.77	2.96	2.98	2.98	2.86	2.92

Table 7: Faithfulness alignment scores across models and dimensions, computed from the absolute difference between model-predicted scores and faithful scores inferred from the generated rationales. Dimension abbreviations follow those defined in Table 4. The best-performing model for each dimension is highlighted in bold.

Models	LA	LR	LS	Cr	Em	Cl	Ap	Ar	GA	GR	GS	Avg.
Mistral-small-latest end-to-end	0.73	0.71	0.71	0.66	0.70	0.67	0.67	0.74	0.66	0.68	0.58	0.68
Mistral-small-latest 2-step	0.90	0.64	0.70	0.48	0.40	0.61	0.64	0.58	0.52	0.43	0.58	0.59

Table 8: RMSE comparison between the original model-predicted scores and the faithful scores inferred from the generated rationales using the two-step evaluation pipeline. Dimension abbreviations follow those defined in Table 4. The best-performing setting for each dimension is highlighted in bold.

emotional appeal, and global relevance. Among the four models, Mistral-small-latest again achieves the strongest overall performance, ranking first in 7 out of the 11 evaluated dimensions.

Beyond model comparison, we further examine the relationship between instructional alignment and prediction accuracy. Table 6 analyzes the outputs of the best-performing model, Mistral-small-latest, by grouping instances according to their instructional alignment level. The results reveal a general tendency: rationales that more closely follow the provided instructions are associated with lower prediction errors. This observation suggests that adherence to the structured reasoning procedure may contribute to more reliable argument quality assessments.

Overall, the results indicate that modern LLMs can largely follow the structured evaluation instructions provided in the structural rationale prompts. The consistently high instructional alignment scores suggest that the generated rationales generally adhere to the intended evaluation criteria. Moreover, the observed association between higher instructional compliance and lower prediction error indicates that closely following the rational procedure may contribute to more reliable argument quality assessments.

6 Faithfulness Alignment

While the previous two research questions examine predicted scores and generated rationales separately, the third research question investigates faithfulness alignment, that is, **the internal consistency between the scores produced by a model and the**

rationales it generates. To assess this relationship, we use the same randomly selected examples described in Section 5, and conducted another 220 annotations. Two human experts and one LLM judge (GPT-5.2) are shown only the generated rationales, without access to the original arguments or the model-predicted scores. Based solely on the rationales, they are asked to assign argument quality scores, which we refer to as *faithful scores*. The agreement between the two human annotators reaches a Krippendorff’s α of 0.65, while the overall inter-annotator agreement across all three annotators is 0.53, indicating moderate agreement.

Using the annotations produced by GPT-5.2, we define faithfulness alignment based on the absolute difference between the faithful score and the original model-predicted score. A value of 3 indicates perfect alignment (no difference), 2 indicates a difference of one point, and 1 indicates a difference of two points. Table 7 reports the resulting faithfulness alignment scores across all dimensions and models. Overall, the predicted scores are highly consistent with the generated rationales. The lowest average faithfulness score is 2.80 for Mistral-7B-Instruct, while Llama-3.3-70B-Instruct achieves the highest score of 2.96, indicating very strong internal alignment.

Finally, we extend the analysis by asking GPT-5.2 to infer faithful scores for the entire dataset and comparing these scores with the human annotations, following the evaluation protocol used in Section 4. This procedure effectively implements a two-step evaluation pipeline: the model first generates a rationale and then derives a score from that rationale. Such a setup resembles

multi-step reasoning approaches such as Least-to-Most prompting (Zhou et al., 2023). Table 8 reports RMSE for each dimension, compared with the best-performing configuration from Section 4 (Mistral-small-latest with structural rationale prompting). The results show that this two-step approach further reduces RMSE, suggesting that explicitly grounding the score in the generated rationale can lead to more accurate predictions.

Overall, the results indicate that the scores produced by LLMs are largely supported by their generated rationales, demonstrating strong faithfulness alignment. Moreover, deriving scores explicitly from rationales in a two-step evaluation pipeline can further improve prediction accuracy. These findings suggest that incorporating structured reasoning not only improves interpretability but can also enhance the reliability of LLM-based argument quality assessment.

7 Conclusion

In this paper, we proposed a three-level audit framework for evaluating the reliability of LLMs as annotators for argument quality assessment, covering surface alignment, instructional alignment, and faithfulness alignment. To operationalize this framework, we introduced structural rationale prompting, which guides models to generate structured justifications before assigning scores. Experiments on the Dagstuhl-15512 argument quality corpus show that this prompting strategy substantially improves agreement with human annotations, while the generated rationales generally follow the intended evaluation criteria and remain highly consistent with the predicted scores. Overall, our findings highlight the importance of auditing LLM evaluators beyond surface score agreement in order to better understand the transparency and reliability of LLM-based argument evaluation.

8 Limitations

This study has several limitations. First, our experiments focus on argument quality assessment using the Dagstuhl-15512 corpus, which contains a limited number (320 instances) of annotated arguments from a specific debate dataset. As a result, the findings may not fully generalize to other domains or argumentation settings. Second, we evaluate a limited set of LLMs and prompting strategies. Although the selected models represent different model families and scales, the results may vary for

other models or alternative prompting approaches.

Third, part of our evaluation relies on an LLM-based judge to assess instructional and faithfulness alignment. While we validate the judge against human annotations on a subset of instances, automated evaluation may still introduce biases or inaccuracies. Finally, the structural rationale prompts used in this work are designed primarily to facilitate the proposed auditing framework rather than to represent an optimal prompting strategy. Future work could explore alternative prompt designs and examine whether similar alignment patterns hold across different reasoning structures.

9 Ethical Concerns

The increasing use of LLMs as automated annotators raises ethical considerations regarding the role of human expertise in data annotation and corpus creation. While LLM-based annotation may reduce the cost and time required to construct labeled datasets, it should not be viewed as a direct replacement for human annotators, particularly for complex tasks such as argument quality assessment that involve nuanced judgment and contextual understanding. The goal of this work is therefore not to advocate replacing human annotators, but rather to examine whether and how LLM outputs align with human evaluation criteria.

Another concern is the potential over-reliance on LLM-generated annotations in downstream research. If LLM outputs are used without careful validation, systematic biases or reasoning errors may propagate into training data and evaluation benchmarks. Our proposed auditing framework aims to mitigate this risk by encouraging researchers to evaluate LLM annotators along multiple alignment dimensions, including their reasoning processes. By promoting transparency in how LLM evaluators generate scores and explanations, we hope to support more responsible use of LLMs in annotation and evaluation pipelines.

Acknowledgments

This paper is based in part on results obtained from the AIST policy-based budget project “R&D on Generative AI Foundation Models for the Physical Domain.” This work was also supported by the AISafety Project, funded by the BMBF under grant proposal 05D23PD1, and by the State of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

- Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, Translator). Clarendon Aristotle series. Oxford University Press.
- J Anthony Blair. 2011. *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair*, volume 21. Springer Science & Business Media.
- WK Chan, YT Yu, Jacky W Keung, and Victor CS Lee. 2023. Toward ai-assisted exercise creation for first course in programming through adversarial examples of ai models. In *2023 IEEE 35th International Conference on Software Engineering Education and Training (CSEE&T)*, pages 132–136. IEEE.
- Wei-Fan Chen, Milad Alshomary, Maja Stahl, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2024. Reference-guided style-consistent content transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13754–13768.
- Wei-Fan Chen, Zhixue Zhao, Akbar Karimi, and Lucie Flek. 2025. Explainable hallucination through natural language inference mapping. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1888–1896.
- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Asaf 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.
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.
- Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.
- Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574.
- John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational linguistics*, 45(4):765–818.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang 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.
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.
- Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Conference on Advances in Robust Argumentation Machines*, pages 129–146. Springer.
- Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here’s my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1364–1373.
- Zahra Rahimi, Diane J Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. 2014. Automatic scoring of an analytical response-to-text assessment. In *International conference on intelligent tutoring systems*, pages 601–610. Springer.

- Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.
- Maja Stahl, Timon Ziegenbein, Joonsuk Park, and Henning Wachsmuth. 2025. Arginstruct: specialized instruction fine-tuning for computational argumentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11103–11127.
- 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.
- Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.
- Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.
- Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.
- Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, and 1 others. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yuxiao Ye and Simone Teufel. 2021. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 669–678.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and 1 others. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Full Example

Table 9 provides an argument together with its issue and stance. Using this argument, the tested four models generate the rationales and predict the quality score as shown in Table 10 and 11. Besides, the alignment scores by GPT-5.2 are also included in the tables.

Human Annotation: 1 (lowest)

Issue: ban-plastic-water-bottles

Stance: no-bad-for-the-economy

Argument: Western Europe consumes almost 50% of the world's bottled water. Many claim that this is due to European culture, since the continent has had very polluted waters due to agriculture and industry dating back to the Industrial Revolution. The water bottles they buy help us.

Table 9: An example argument together with its issue and stance.

Llama-3.3-70B-Instruct

Predicted Score: 1

Instructional Alignment: 3

Faithfulness Alignment: 3

Rationales:

1. **Conclusion**: The argument aims to support the stance “no-bad-for-the-economy” — i.e., banning plastic water bottles is not bad for the economy. However, the stated conclusion is not clearly articulated. Instead, the argument makes a claim about Western Europe’s bottled water consumption and cultural reasons for it, concluding with “The water bottles they buy help us.” This is vague and does not logically connect to the economic impact of a ban.
 2. **Premises**:
 - Western Europe consumes nearly 50% of the world’s bottled water.
 - This consumption is attributed to historical pollution from agriculture and industry (dating to the Industrial Revolution).
 - The purchase of bottled water by Europeans “helps us” (implying economic benefit to others, likely producers or exporters).
 3. **Evaluation of Logical Connections**:
 - The premises do not address the economic consequences of banning plastic water bottles.
 - The claim that “the water bottles they buy help us” is ambiguous: it’s unclear who “we” refers to, what kind of help is meant (e.g., export revenue, job creation), and how this relates to a ban.
 - Even if Western Europe’s consumption supports some industries, this does not justify the claim that a ban would be *bad for the economy*. In fact, a ban might shift demand to alternatives (e.g., reusable bottles, filtered tap water), which could still support economic activity.
 - There is no reasoning linking the current consumption pattern to the economic viability of maintaining plastic bottle use under a ban.
 4. **Assessment of Premise Quality**:
 - The premise about Western Europe consuming 50% of bottled water is plausible and supported by data (e.g., from Euromonitor or OECD reports), so it is factually acceptable.
 - The historical pollution claim is broadly consistent with known environmental history of Western Europe.
 - However, the causal link between pollution and bottled water use is oversimplified and lacks nuance (e.g., many European countries have high-quality tap water but still consume bottled water due to convenience or marketing).
 - The final claim — “The water bottles they buy help us” — is unsupported, vague, and lacks evidence or explanation.
 5. **Global Judgment**:
 - The premises fail to provide sufficient support for the conclusion that banning plastic water bottles is bad for the economy.
 - There is a significant gap between the evidence presented and the economic claim being defended.
 - The argument appears to confuse current consumption patterns with long-term economic impacts of regulation.
 - It does not address alternatives, market shifts, or potential economic benefits of a ban (e.g., reduced waste management costs, growth in sustainable industries).
- Therefore, the argument suffers from a fundamental lack of logical sufficiency. The premises are either irrelevant or too weak to justify the conclusion, despite some plausible factual claims.

GPT-3.5-Turbo

Predicted Score: 2

Instructional Alignment: 2

Faithfulness Alignment: 3

Rationales:

The premises provided in the argument are relevant and logical, linking European culture and historical context to the consumption of bottled water. However, the argument lacks depth and evidence to fully support the conclusion that water bottles help the economy. The connection between cultural history and economic benefit is not explicitly detailed, leading to a partially sufficient local sufficiency score.

Table 10: The output rationales together with the predicted local sufficiency quality score and the alignment scores by Llama-3.3-70B-Instruct and GPT-3.5-Turbo.

Mistral-7B-Instruct

Predicted Score: 2

Instructional Alignment: 2

Faithfulness Alignment: 3

Rationales:

While the premises are relevant and logically connected, they lack depth and specificity in supporting the conclusion that the Western European consumption of bottled water is not bad for the economy. The argument relies on generalizations about European culture and history without providing concrete evidence of the economic benefits of bottled water consumption. Additionally, the premises could be strengthened by including specific data or examples that demonstrate the positive impact of bottled water on the economy.

Mistral-small-latest

Predicted Score: 1

Instructional Alignment: 3

Faithfulness Alignment: 3

Rationales:

1. **Conclusion**: The argument aims to support the stance that banning plastic water bottles is bad for the economy by claiming that plastic water bottles are beneficial to the economy (implied by "The water bottles they buy help us").
 2. **Premises**:
 - (1) Western Europe consumes almost 50% of the world's bottled water.
 - (2) This consumption is due to European culture, as the continent has had polluted waters due to agriculture and industry since the Industrial Revolution.
 - (3) The water bottles they buy help us (presumably the global economy or the author's country).
 3. **Evaluation**:
 - The premises are somewhat relevant to the conclusion but contain significant gaps in reasoning.
 - Premise (1) provides a factual claim about consumption but does not directly explain how this supports the economy.
 - Premise (2) attempts to justify the consumption by historical context but does not logically connect to economic benefits.
 - Premise (3) is vague ("help us") and lacks specificity about how the economy benefits. It assumes a positive economic impact without evidence.
 - The argument does not explain why banning plastic water bottles would harm the economy, nor does it provide data or reasoning to support the claim that bottled water is economically beneficial.
 4. **Overall Judgment**:
 - The premises are weak, lack depth, and fail to provide sufficient logical or evidentiary support for the conclusion. The argument relies on unsupported assumptions and broad generalizations.
-

Table 11: The output rationales together with the predicted local sufficiency quality score and the alignment scores by Mistral-7B-Instruct and Mistral-small-latest.