# Argument Quality Assessment with Large Language Models: A Pairwise Bradley-Terry Approach

Anonymous ACL submission

## Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in tasks related to reasoning and judgment. However, assessing the quality of arguments requires a rigorous evaluation. This research investigates the extent to which LLMs can effectively perform this task. It focuses on the zero-shot capabilities of LLMs in approximating expert rankings of argument quality across three dimensions: logical, rhetorical, and dialectic. It also examines the model's specific strengths and weaknesses within a prompt-engineering and pairwise evaluation framework using a Bradley-Terry model to infer latent strength scores and obtain a ranking of arguments. Although none of the models tested (GPT-4, Gemini 2.0 Flash, and LLaMA 3.3) achieved strong alignment with the human-provided gold standard, GPT-4 demonstrated the most consistent overall performance, followed by Gemini 2.0 Flash, with LLaMA 3.3 ranking third across most dimensions. LLMs show promising potential but still fall short of replicating expert-level evaluations.

## 1 Introduction

Argumentation is fundamental to reasoning in various fields (Chalaguine and Schulz, 2017): scientists publish discoveries together with supporting evidence, lawyers structure legal arguments to solve disputes, and political debaters rely on argumentation to gain the approval of the public. Early computational work focused on mining and evaluating arguments in well-structured texts such as essays (Wachsmuth et al., 2016) until more research shifted toward assessing argument quality (Wachsmuth et al., 2017b). A particularly effective approach within the field has been the use of pairwise comparisons, in which the annotators judge which of the two arguments is stronger based on specific criteria. These judgments can then be aggregated using statistical models, such as the Bradley-Terry (BT) model, to produce quality scores (Gienapp et al., 2020) and obtain a ranking for each of the arguments evaluated. This design reduces annotation variance and cost while producing multidimensional quality scores, offering a rich, reliable foundation for benchmarking automated evaluation methods. With the rise of online platforms, the ability to identify high-quality arguments becomes increasingly important (Wachsmuth et al., 2024), especially for effective persuasion, decision-making, and participation in meaningful discourse. However, manually evaluating the quality of arguments at a large scale remains a significant bottleneck due to high costs and inconsistencies.

Large Language Models (LLMs) offer a promising avenue for scaling argument quality assessment. LLMs demonstrated impressive capabilities in language understanding tasks, and a growing line of work explores the use of LLMs as annotators (Mirzakhmedova et al., 2024)—an example of LLM-as-a-judge—where the language model is asked to evaluate text quality (AI, 2025). This strategy could enable large-scale assessments with minimal cost and time. However, whether LLMs can reliably replicate expert-level judgments is an open question. Existing research offers mixed results: some studies show that, properly guided LLMs , models such as GPT-3 and PaLM-2 can effectively assess certain quality aspects (Mirzakhmedova et al., 2024), while the other found that GPT-3.5 underperformed specialized supervised models on tasks that involve ranking of arguments (Wang et al., 2023). These contrasting findings motivate a deeper investigation into LLM-based evaluation of argument quality. This study attempts to replicate expert-derived argument quality rankings from the Webis-ArgQuality-20 dataset. Given that, argument quality is multidimensional (Wachsmuth et al., 2017b), this study will also examine the abilities of the LLMs in assessing the different quality dimensions. The overall scientific aim of this paper is to address the following research questions:

**Main Research Question**: *How well can zero-shot LLM pairwise comparisons reproduce expert rankings of argument quality?* This question examines whether LLMs can reliably assess argument quality in comparison to expert judgments. It is tackled through the following sub-questions: 1) **RQ-1.1**: *How does LLM performance vary across the different argument quality dimensions, and which dimensions are most accurately assessed?* 2) **RQ-1.2**: *How do the different LLMs compare in their ability to evaluate argument quality?* 3) **RQ-1.3**: *How consistent are LLM-based evaluations when repeated over multiple trials?* The contributions of this work are as follows: 1) A novel evaluation pipeline that combines prompt engineering and pairwise comparisons of arguments by LLMs with the Bradley-Terry Model (BT) to rank arguments. To the best of our knowledge, this is the first application of BT ranking to LLM-based comparative judgments in argument mining. 2) This statistical framework that evaluates LLM performance against an expert-derived gold standard, offering a robust assessment of their capabilities as evaluators. 3) A detailed analysis of model performance across the different dimensions, revealing which aspects of argument quality are more or less reliably assessed by the LLMs.

## 2 Related Work

Wachsmuth et al. (2017b) presented a framework for computational argument quality introducing a taxonomy of across logical, rhetorical, and dialectic dimensions. They collected 320 arguments, annotated by experts for all dimensions. The work laid a strong theoretical foundation, but the annotation process revealed a limited consistency across dimensions. We adopt the same framework but replace the expert annotators with LLMs.

Habernal and Gurevych (2016) proposed a convincingness-based argument ranking system using more than 16,000 pairwise comparisons in 1,052 arguments on 32 topics. The method used for our study demonstrated that reliable rankings could be inferred from crowdsourced pairwise comparisons, achieving a high agreement with the gold standard (Gienapp et al., 2020; Habernal and Gurevych, 2016). Our paper shifts from crowdsourcing to model-driven annotation, where LLMs are examined in their ability to serve as reliable judges without the need for manual labeling.

The reliability of pairwise annotation for argument quality continues with Toledo et al. (2019) who compiled a dataset of 14,000 annotated argument pairs and absolute quality scores in the range of 0-1 for 6,300 arguments. Crowd workers were asked to pick which of the arguments would have been preferred by most people to support the topic, yielding consistent results. Also, they suggested neural methods based on a language model for argument ranking and classification. Results from those models were considered state-of-the-art, validating the capabilities of supervised neural methods. Our study does not rely on large-scale annotation, but makes use of the LLMs out-of-the-box knowledge.

Gretz et al. (2019) created a corpus of 30,497 arguments with crowd-annotated quality scores covering 71 diverse topics. They analyze the quality dimensions that characterize the dataset, showing that the dimensions of Global Relevance and Effectiveness are the most indicative to overall quality scores. They present a BERT-based model for argument quality ranking, which outperformed state-of-the-art. Our paper, instead, investigates whether models can match the expert's ranking in a zero-shot setting without the need for labeled data.

The work of Gienapp et al. (2020) is especially relevant to this paper. They introduced an efficient annotation framework that combines pairwise comparisons and active sampling to label arguments on argument quality dimensions such as logical, rhetorical, and dialectical, as well as overall quality. The Webis-ArgQuality-20 corpus with 1,271 arguments resulting from their work is the data source used for this research, as it aligns with the objective of dealing with multidimensional quality assessment via pairwise comparisons. The key difference is in who/what is doing the comparison, as we replace crowd workers with LLMs.

Mirzakhmedova et al. (2024) study the potential of using state-of-the-art LLMs as proxies for annotators in argument quality assessment. They analyze the agreement between model, human expert, and novice human annotators based on an established taxonomy of argument quality dimensions. Their findings show that LLMs can produce consistent annotations with moderately high agreement that aligns with human experts. Their argument quality dimensions differ from ours and we will extend in the future our experiments with those.

## 3 Background

We introduce here key foundational concepts.

### 3.1 Argument Quality Dimensions

Theoretical work in argumentation often distinguish three dimensions: logical, rhetorical, and dialectical (Blair, 2011; Wachsmuth et al., 2017b). The **logical dimension** or *cogency*: A logically good argument has individually acceptable premises that are relevant and sufficient to draw the conclusion. Wachsmuth et al. (2017a) comprised the logical quality as: *local acceptability* (are the premises worthy of acceptance?), *local relevance* (do the premises support or attack the conclusion?), and *local sufficiency* (are the premises enough to justify the conclusion?). The **rhetorical dimension** or *effectiveness*: An argument is effective if it succeeds in persuading a target audience (Wachsmuth et al., 2017b). Starting from Aristotle's rhetorical appeals, which include pathos (emotional appeal), ethos (the arguer's credibility), and logos (appeals to the audience's reason), they further decomposes this dimension into: *credibility* (establishing the speaker's authority), *emotional appeal* (engaging the audience), *clarity* (using correct and unambiguous language), *appropriateness* (aligning the tone and language with the topic), and *arrangement* (structuring the argument). The **dialectic dimension**, or *reasonableness*: It focuses on the ability of the argument to contribute to resolving a disagreement, especially with regard to counterarguments. It is further decomposed into *acceptability* (the audience's acceptance of the argument), *relevance* (the ability of the argument to advance the discussion) and *sufficiency* (rebuttal to counter arguments) (Wachsmuth et al., 2017a).

### 3.2 Pairwise Comparison via BT

Pairwise comparisons are a widely used method in statistical analysis to evaluate preferences between pairs of items (Nordstokke and Stelnicki, 2014). In argument quality assessment, this approach involves presenting two arguments at a time and asking the annotators to judge which is better based on specific criteria. Pairwise comparisons reduce cognitive load and time for annotation compared to ranking. (Kyne, 2022). We use the Bradley Terry (BT) model to derive continuous quality scores from these comparisons. Given a pair of items (e.g., arguments) $d_i$ and $d_j$ drawn from the dataset, BT estimates the probability that the pairwise comparison $d_i \succ d_j$ turns out true through the formula (Hunter, 2004): $P(d_i \succ d_j) = \frac{\gamma_i}{\gamma_i+\gamma_j}$. $i > j$ denotes that argument $d_i$ is preferred to $d_j$.

### 3.3 Large Language Models (LLMs)

LLMs have emerged as powerful tools for the generation and understanding of natural language. Based on vast trained corpora of text and their architectures, these models capture statistical patterns inherent in human language, and use them to generate (Coronado-Blázquez, 2025) and evaluate (AI, 2025) text. We use: **GEMINI 2.0 Flash**, **Llama 3.3 70B Instruct Turbo**, **GPT-4.1 mini**.

## 4 Research Methods

This study adopts a controlled experiment to evaluate whether LLMs can replicate expert-derived argument quality rankings published for the Webis-ArgQuality-20 corpus (Gienapp et al., 2020). The approach follows the original pairwise framework, except for replacing the annotators with LLMs and using a different sampling procedure.

### 4.1 Prompt Engineering

Each pairwise comparison task is initiated with a prompt that presents the two arguments—labeled *Argument A* and *Argument B* - and instructs the LLM to decide which argument is better with respect to a single quality dimension. The response is restricted to the following: `'A'` if A is better, `'B'` if B is better, or `'Tie'` to indicate a tie. No additional text is allowed and a temperature of 0.2 was applied to ensure accurate responses (Ujawane, 2024). A zero-shot prompt engineering technique is used, so no prior examples are provided (Gadesha, 2025). No additional training is conducted. The prompts are provided in the appendix A.

### 4.2 Pairing Strategy

Arguments are grouped by the topic identification number provided in the dataset, to avoid argument comparison between unrelated issue. Within each topic, cyclic overlapping pairings are generated using a step size of two. This means that the argument $i$ is compared to its nearest (2) neighbors $i{+}1$ and $i{+}2$. For the main results, each model completed 1,500 comparisons per dimension, yielding 4,500 comparisons per model across the three dimensions. With three models evaluated (GPT-4.1 Mini, Gemini 2.0 Flash, and LLaMA 3.3) and three trial runs conducted (further explained in the next sub-section 4.2.1), this resulted in 13,500 comparisons per trial run. The total number of comparisons amounted to 40,500. To explore the effect of additional comparisons, an alternative configura-

tion with a step size of three (where the argument $i$ is compared to $i+1$, $i+2$, and $i+3$ ) was also tested. Also, a second alternative configuration was introduced in which the step size of three was combined with a modified prompt. This variation aimed to assess whether the prompt could improve model performance. Both configurations are analyzed to determine their effect on the evaluation metrics.

#### 4.2.1 Trial Execution and Reliability

LLMs possess inherently probabilistic structures (Coronado-Blázquez, 2025); their outputs can exhibit deterministic qualities even when identical inputs are provided. This is substantially influenced by training data, which may not truly cover the evaluation context. In some cases, LLMs hallucinate, generating responses that appear coherent but are not supported by evidence(Ni et al., 2024). To address this, each model underwent three independent evaluation trials in each dimension using the same pairing strategy. For each trial, the model was prompted separately, generating one output file per dimension, resulting in a total of nine output files across the three trials. These outputs were analyzed to assess inter-run consistency and the reliability of each model's judgments. Given the minimal variation among trials, the scores for each pairwise comparison were averaged among runs.

### 4.3 Data Collection

The study relies on the argument quality corpus known as the Webis-ArgQuality-20 corpus introduced by Gienapp et al. (Gienapp et al., 2020). This gold standard dataset contains 1,610 texts covering 20 debate topics. Each piece of text is ensured to contain arguments (checked via crowdsourcing). Arguments were then subjected to pairwise comparison analysis by experts, per quality dimension, and then ranked via a BT model.

#### 4.3.1 Data Preparation Steps

The first step was to remove all the non-argumentative items from the dataset. This was done by adding a filter to the is_argument metadata, which left 1,271 usable rows. From this filtered set, 750 arguments were randomly pre-selected. This sample was used consistently across all models to ensure fairness and comparability in the evaluation. The choice to use 750 arguments was based on practical constraints that include experimental validity, API token limits, rate limitations, and other costs. The arguments ran across all 20 topics.

#### 4.3.2 Evaluation metrics

To assess how well the LLM-derived rankings/scores aligned with the expert gold standard, different metrics were applied: 1) Pearson correlation ($r$) measures the strength of the linear relationship between two variables (BT scores) produced by both the LLM and the expert scores. (Williams et al., 2020)., 2) Spearman rank correlation ($\rho$) assesses the degree of similarity between the LLM rankings and the expert gold standard (Easily, 2024)., 3) MAE (Mean Absolute Error) and RMSE (Root Mean Square Error): quantify the deviation of the predicted scores (LLM-based BT scores) from the expert gold standard in terms of average of the absolute differences and squared average differences (Mondal, 2024; Choudhary, 2024).

#### 4.3.3 Inter-Rater Agreement Analysis

To assess the reliability of the evaluation results, the inter-rater agreement was measured using Cohen's Kappa ($\kappa$). This statistical metric quantifies the agreement between two raters on categorical decisions. Higher inter-rater reliability indicates that the collected data consistently reflects the variables being measured (McHugh, 2012). **LLM vs. Expert Annotations:** Pairwise comparisons from each LLM were evaluated against the expert comparisons in the Webis-ArgQuality-20 dataset, assessing alignment with the expert judgments across logic, rhetoric and dialectic. **LLM vs. LLM (Comparisons):** Agreement between the different LLMs on the same argument pairs was measured to assess the consistency across models with varying architectures. **LLM Comparison in Three trials:** To assess model reliability, pairwise comparisons from the repeated runs are examined using $\kappa$. By analyzing the latent strength scores derived from the BT model, together with the $\kappa$ values, the study offered more comprehensive evaluation of the internal coherence and reliability of LLM based assessments.

### 4.4 Link to Research Questions

**Main RQ: How well can zero-shot LLM pairwise comparisons reproduce expert rankings of argument quality?** To answer this, three different LLMs were applied to a dataset of 750 arguments sampled from the Webis-ArgQuality-20 gold standard. Each model generated pairwise comparisons of the different arguments, which were then further processed using the BT model to derive latent strength scores. These scores provided a ranking of the arguments that could be directly

compared against expert rankings. To further evaluate the alignment between the LLM generated and the expert rankings, a combination of different metrics were used. This approach ensured that the analysis addressed both the strength of linear and rank-order associations, as well as the absolute magnitude of deviation from expert judgments.

**RQ-1.1: How does LLM performance vary across the different dimensions of argument quality, and which dimensions are most accurately assessed?** To address this question, the evaluation was conducted across the three key dimensions of argument quality: logical, rhetorical, and dialectic. Dimension-specific prompts were carefully designed with definitions based on criterion from the argumentation literature (Wachsmuth et al., 2017b). By fitting the BT model for each dimension, the analysis reveals whether LLMs demonstrate any particular strengths/weaknesses related to the different facets of argument quality.

**RQ-1.2: How do the different LLMs compare in their ability to evaluate argument quality?** This sub-question was addressed through a cross-model evaluation. The three selected LLMs were run on the same set of 750 arguments using the same sampling and prompting techniques to ensure comparability. By analyzing the results produced by the three models under equivalent conditions, the study was able to observe performance trends specific to each model. This not only allowed for fair comparison evaluation, but also provided insight into how the different models and conditions influence their assessment of argument quality.

**RQ-1.3: How consistent are LLM-based evaluations when repeated over multiple trials?** LLMs may produce (slightly) different output when prompted repeatedly; therefore, reliability was a key consideration in study design. To account for this, each model was executed in three separate trial runs over the entire argument set. The performance indicators were then averaged to produce stable final results. This step is essential to address the reliability and consistency of LLM performance.

## 5 Experiment Set-Up

To check consistency across LLMs, the final experimental setup involved three separate evaluation runs, one for each LLM. Each model was provided with the same number of arguments (750). Results were very consistent across the three runs. Furthermore, in order to produce the final reported results, the performance indicators of the three trials were averaged into a single value per model and dimension. However, the correlation agreements were generally weak between the LLM's results and the expert-derived scores and ranks. For the first three trials, a cyclic step size of 2 was used - each argument was compared to its next two neighbors within the same topic. These trials used prompt version 1, which included sub-criteria for each quality dimension. Two other exploratory trials are conducted where the step size is increased to three, and a second version of the prompt is used.

**Cost and Computational Evaluation** Preliminary runs helped determine a configuration to balance output quality, pairwise comparison coverage, and budget, yielding a sample size of 750 arguments. Each trial required about 2.5 - 3.5 hours of computational time depending on batch size, number of comparisons, rate limits, and API errors.

## 6 Results

| Model | Dimension | Pearson | Spearman | MAE | RMSE |
|---|---|---|---|---|---|
| GPT-4.1 Mini | Logic | 0.323 | 0.341 | 0.921 | 1.164 |
| | Rhetoric | 0.351 | 0.379 | 0.898 | 1.139 |
| | Dialectic | 0.376 | 0.387 | 0.883 | 1.117 |
| Gemini 2.0 Flash | Logic | 0.287 | 0.300 | 0.946 | 1.195 |
| | Rhetoric | 0.319 | 0.351 | 0.918 | 1.167 |
| | Dialectic | 0.332 | 0.338 | 0.918 | 1.155 |
| Meta LLaMA 3.3 | Logic | 0.308 | 0.326 | 0.920 | 1.177 |
| | Rhetoric | 0.312 | 0.340 | 0.917 | 1.173 |
| | Dialectic | 0.317 | 0.330 | 0.925 | 1.168 |

Table 1: Performance of LLMs per argument quality dimensions (averaged over three trials).

**Agreement with Human Assessment** Across the three models, the performance varied both by model and by quality dimension (see Table 1.) GPT 4.1 mini consistently achieved the highest correlations with human-derived rankings, particularly in the dialectic dimension. GPT 4.1 mini prevails with the lowest error in the dialectic dimension.

| Dimension | GPT-4.1 Mini | Gemini 2.0 Flash | LLaMA 3.3 |
|---|---|---|---|
| Logic | 0.96 | 0.93 | 0.96 |
| Rhetoric | 0.97 | 0.96 | 0.97 |
| Dialectic | 0.96 | 0.91 | 0.97 |

Table 2: Average $\kappa$ values measuring agreement between repeated runs of each LLM.

**Agreement Across Repeated Runs** Within-model agreement was further evaluated using $\kappa$

across the three runs. Each trial run was compared to other runs and the results were averaged in Table 2. All three models achieved high levels of consistency. The BT model does not capture whether the same pairwise outcomes are consistently reproduced across trials. This analysis offers a complementary perspective of within-model reliability.

### 6.1 Exploratory Trial Runs

Two exploratory trial runs were conducted to examine the effects of the sampling strategy and prompt design on model performance. These trials served as a test to see whether adjustments in prompt formulation and number of comparisons would influence the correlations between the LLM-generated ranks and expert decisions. Through these tests, valuable information was obtained on how methodological choices might impact results. One trial keeps the cyclic step size at 3 while using the version 1 prompt (including dimensions criteria) used for the table 1, but the other uses a different prompt with step size 3 that asks the model to decide based only on the name of the dimension without detailing the criteria. This helps us to explore whether the models could rely on their internal understanding of the dimensions. However, no repetitions were performed due to resource constraints. Each trial generated 2,250 comparisons each per LLM and per dimension (6,750 overall).

| Model | Dimension | Pearson $r$ | Spearman $\rho$ | MAE | RMSE |
|---|---|---|---|---|---|
| GPT-4.1 Mini | Logic | 0.377 | 0.400 | 0.886 | 1.116 |
| | Rhetoric | 0.386 | 0.415 | 0.867 | 1.109 |
| | Dialectic | 0.434 | 0.445 | 0.842 | 1.064 |
| Gemini 2.0 Flash | Logic | 0.329 | 0.344 | 0.935 | 1.158 |
| | Rhetoric | 0.373 | 0.402 | 0.880 | 1.120 |
| | Dialectic | 0.390 | 0.396 | 0.883 | 1.104 |
| LLaMA 3.3 70B | Logic | 0.339 | 0.359 | 0.901 | 1.149 |
| | Rhetoric | 0.348 | 0.374 | 0.898 | 1.142 |
| | Dialectic | 0.362 | 0.357 | 0.906 | 1.129 |

Table 3: Expl. Run Results (Prompt v.1, Step = 3)

**Step Count 3 - Prompt v. 1** This exploratory trial used prompt version 1 with a cyclic step size of three. The purpose of this run was to investigate whether increasing the number of comparisons would lead to stronger correlations with the expert rankings. Compared to the final averaged results in Table 1, the exploratory results shown in Table 3 demonstrate slightly higher correlations. For example, GPT 4.1 mini in the dialectic dimension achieved a $r$ of 0.434 and a $\rho$ of 0.445, both higher than those in the averaged results. Similar patterns were observed for Gemini 2.0 Flash and LLaMA 3.3. Thus, increasing the number of comparisons may improve the reliability of the model rankings.

| Model | Dimension | Pearson $r$ | Spearman $\rho$ | MAE | RMSE |
|---|---|---|---|---|---|
| GPT-4.1 Mini | Logic | 0.392 | 0.416 | 0.873 | 1.103 |
| | Rhetoric | 0.417 | 0.450 | 0.843 | 1.080 |
| | Dialectic | 0.432 | 0.447 | 0.855 | 1.066 |
| Gemini 2.0 Flash | Logic | 0.358 | 0.377 | 0.894 | 1.133 |
| | Rhetoric | 0.418 | 0.442 | 0.851 | 1.079 |
| | Dialectic | 0.385 | 0.394 | 0.866 | 1.109 |
| LLaMA 3.3 70B | Logic | 0.346 | 0.365 | 0.894 | 1.144 |
| | Rhetoric | 0.359 | 0.387 | 0.879 | 1.133 |
| | Dialectic | 0.364 | 0.365 | 0.889 | 1.128 |

Table 4: Expl. Run Results (Prompt v. 2, Step = 3)

**Step Count 3 - Prompt v. 2** Table 4 presents the performance metrics of the exploratory trial run using a revised prompt and a step count of 3. GPT-4.1 mini led in performance, particularly in the rhetoric dimension, with the results in the dialectic dimension closely following. Gemini 2.0 Flash showed a notable improvement in the rhetoric dimension. LLaMA 3.3 maintained stable performance with a slight increase in all dimensions. Interestingly, Gemini 2.0 Flash also showed gains in the logical dimension compared to previous trials. Compared with Table 1, these results indicate that prompt refinement and increasing the number of comparisons may significantly improve the model reliability.

### 6.2 Inter-Rater Agreement Analysis

To assess the reliability of LLMs in reproducing expert pairwise comparisons of argument quality, $\kappa$ was computed. Expert comparisons were compared to LLM-based judgments obtained from the exploratory trial conducted with prompt version 2 using a step size of three which produced the results in the table below, since this configuration provided a much broader coverage of comparisons.

| Dimension | GPT vs Expert | Gemini vs Expert | LLaMA vs Expert |
|---|---|---|---|
| Logic | 0.492 | 0.446 | 0.466 |
| Rhetoric | 0.485 | 0.473 | 0.467 |
| Dialectic | 0.531 | 0.525 | 0.509 |

Table 5: $\kappa$ values measuring agreement between LLMs and human experts across quality dimensions.

**Agreement Between LLMs and Experts** Table 5 presents the agreement ($\kappa$) between each LLM and the expert annotations across dimensions.

| Dimension | GPT vs Gemini | GPT vs LLaMA | Gemini vs LLaMA |
|---|---|---|---|
| Logic | 0.737 | 0.798 | 0.737 |
| Rhetoric | 0.808 | 0.820 | 0.812 |
| Dialectic | 0.757 | 0.786 | 0.730 |

Table 6: $\kappa$ values measuring inter-LLM agreement.

**Agreement Between LLMs** The inter-model agreement was considerably higher than the agreement with experts, as shown in Table 6. In the logical dimension, agreement was lowest overall, while rhetoric showed the strongest agreement.

## 7 Discussion

We discuss here the results obtained per RQ.

**RQ-1.1: How does LLM performance vary across the different argument quality dimensions and which dimensions are most accurately assessed?** The results indicate that performance varied across all dimensions and between models. Based on Spearman correlations, the rhetorical dimension emerged as the strongest for two of three models (Gemini 2.0 Flash and LLaMA 3.3). However, GPT 4.1 mini achieved its highest scores in both Pearson and Spearman correlations for the dialectical dimension. In Table 4, which reports results from the trial run using prompt version 2 with a cyclic step size of 3, the rhetorical dimension again produced the highest spearman correlation, confirming its robustness under the different prompts. Similar patterns were observed with $\kappa$ analysis of inter-LLM agreement, where rhetoric achieved values between 0.808 and 0.820, and in inter-trial reliability, where rhetoric consistently produced values ranging from 0.96 to 0.97. These findings suggest that the models were particularly effective in capturing signals of persuasiveness, clarity, and emotional appeal. However, examining $\kappa$ values for agreement between LLMs and human experts, the dialectic dimension emerged as the strongest. Additionally, dialectical quality also led in Pearson correlations in several tables. In contrast, logical quality was consistently the most challenging across all models, showing slightly lower correlations and higher error values. This may reflect the inherent difficultly of logical assessment: criteria such as sufficiency of evidence, acceptance, and relevance in justifying a conclusion require deep reasoning and structured evaluation. Research shows that some models lack a deep understanding of logical fallacies (Payandeh et al., 2023).

**RQ-1.2: How do the different LLMs compare in their ability to evaluate argument quality?** Clear differences were observed between the three LLMs evaluated. GPT 4.1 Mini delivered the strongest overall performance, consistently outperforming Gemini 2.0 Flash and LLaMA 3.3, although the correlations with experts remain low. GPT 4.1 Mini achieved moderate Spearman correlations in the different trials carried out, and the highest across all dimensions, indicating relatively strong alignment. Gemini 2.0 Flash ranked second, leading in two out of the three dimensions (rhetoric and dialectic) of which it performed best in the rhetoric for both Spearman and Pearson, as well as the error metrics. It also outperformed GPT 4.1 Mini in the rhetorical dimension by a margin of 0.001 ($r$). It outperformed LLaMA 3.3 in the logical dimension ($r$) compared to other tables where it lagged behind. Notably, Gemini's $\kappa$ values suggested stronger agreement with experts for pairwise comparison decisions in the dialectic than rhetoric. LLaMA 3.3 came last, showing the lowest correlations and greater errors, though in some cases it performed comparably to Gemini in the logical dimension ($r$). The LLM inter-agreement analysis ($\kappa$) revealed strong consistency among the models, particularly in rhetoric. Although LLaMA's agreement with experts was lowest, it did not fall far behind the other models. Also, its $\kappa$ agreement across the repeated runs was slightly higher than the other two models. These findings confirm that although zero-shot LLMs can not yet match expert level evaluations, they can approximate them to a degree that is both measurable and meaningful, particularly when evaluating rhetoric and dialectic. GPT 4.1 Mini, the newest model among those evaluated, delivered the strongest performance. Continued advances in LLM development translate into improvements in assessing argument quality.

**RQ-1.3: How consistent are LLM-based evaluations when repeated over multiple trials?** Given that LLMs are inherently probabilistic, their output may vary across runs even when presented with the same inputs. To assess this, each model was run three times on the same set of arguments and pairwise comparisons, and inter-trial agreement was measured using $\kappa$. The results revealed considerably higher consistency for all LLMs, as the values ranged between 0.91 and 0.97, usually indicating almost perfect agreement. In addition to this, the BT scores further confirmed the stability of the model output across the three trials, with consistency evident even before averaging the results for final reporting. Taken together, the agreement metrics and BT scores suggest that all three models demonstrated particularly consistent rankings, despite differences in their overall alignment with

expert judgments. This is important as a model that performs well in a single run but fluctuates across repetitions may be less reliable in practice. Although all LLMs produced consistent rankings, GPT 4.1 Mini and Gemini 2.0 Flash demonstrated strong reproducibility, while LLaMA 3.3 also performed reliably. Nevertheless, future work could further enhance reliability checks by running additional trials and applying a majority vote strategy to understand the effect of output variations.

**Main Research Question** The results of this study indicate that zero-shot LLMs can partially approximate expert rankings of argument quality, particularly in rhetoric and dialectic. Among the models evaluated, GPT 4.1 mini consistently demonstrated the strongest alignment with expert scores, followed by Gemini 2.0 Flash, with LLaMA 3.3 trailing. Although none of the models replicated perfectly expert rankings, moderate correlations suggest that current LLMs have the ability to capture some cues that experts use when evaluating arguments. Two of the three models (Gemini 2.0 Flash and LLaMA 3.3) performed best in rhetoric, suggesting that LLMs are particularly sensitive to cues related to this dimension. GPT 4.1 Mini also had good results in rhetoric but dialectic had better results, whereas logical quality posed the greatest challenge across the models. Overall, the results indicate that zero-shot LLMs cannot yet serve as full substitutes for expert evaluation, but they can provide valuable and cost-effective approximations. Their ability to capture rhetorical and dialectical quality is particularly promising for applications like debate moderation or discourse analysis, as seen from the exploratory trials. The limitations in assessing logical quality suggest the need for hybrid approaches combining LLM-based evaluation with logical reasoning or human oversight.

## 8 Conclusion

This paper explored the potential of zero-shot LLMs in assessing argument quality across multiple argument quality dimensions using pairwise comparisons and the application of Bradley Terry modeling. Through prompt engineering and evaluation of three LLMs - GPT-4.1 mini, Gemini, and LLaMa 3.3 - in multiple trials, this research offers a detailed analysis of how well LLM generated rankings align with the expert's judgments and how reliably these rankings are produced across runs. The findings demonstrate that while LLMs do not yet match expert-level evaluation, they are capable of approximating expert rankings to a measurable and meaningful degree, especially in rhetorical and dialectical dimensions. GPT 4.1 Mini consistently outperformed the other models in the trials, showing the highest (but moderate) correlation with the expert scores. Gemini 2.0 Flash followed closely, particularly excelling in the rhetorical dimension, while LLaMA 3.3, though the weakest overall, showed occasional strength in logical comparisons during earlier trials, according to Pearson correlations. These performance differences highlight how the model, scale, and potentially their training data influence zero-shot evaluative capabilities. Analysis of repeated trials using $\kappa$ demonstrated that all three models produced highly consistent pairwise judgments, despite the nature of LLM outputs. Exploratory trials further suggested that prompt refinement and adjustments to comparison parameters such as the cyclic step size can influence performance, indicating opportunities for optimizing performance without compromising reliability. This work contributes a replicable framework for evaluating the quality of arguments with LLMs, integrating prompt engineering, pairwise comparisons methodology and statistical modeling. It reinforces the importance of multidimensional assessment in argument evaluation and shows that with human oversight and careful design, LLMs can assist in complex judgment tasks on a scale.

Future work will build on this study in several directions. First, we will explore the fine-tuning of LLMs specifically for argument quality assessment, possibly focusing on specific dimensions. Tailoring models through domain specific training or few shot learning may help them better capture reasoning validity and detect fallacies. Second, we will increase the scale of pairwise comparisons aiming at reducing variability in the probabilistic outputs of LLMs and improving the robustness of the BT rankings. Third, prompt strategies also merit further investigation, since less rigid prompting led to small gains in the different dimension evaluations. We will employ other prompt engineering techniques such as chain-of-thought or other structured approaches to assess whether they improve the reliability of pairwise evaluations. Fourth, we will broaden the scope of argument quality dimensions, including additional dimensions and aiming at an overarching quality dimension. All code and the dataset used are available at: https://anonymous.4open.science/r/arg_quality_llm-12E9.

## 9 Limitations

In this section, the limitations of the study are discussed as well as the potential threats to the validity of the findings. The discussion is structured according to the common validity categories inspired by Wohlin et al. (Wohlin et al., 2012) guidelines, namely, internal validity, external validity, construct validity and conclusion validity.

### 9.1 Internal Validity

One potential threat was the unpredictable nature of how LLMs respond. Different runs could yield different pairwise judgments, which could affect the results. To mitigate this, three separate trials were conducted for each model and dimension. As reported, the results remained stable and did not significantly skew the findings. Although averaging across the trial runs provided a stable and cost-efficient measure of model performance. An alternative approach would be to apply a majority vote strategy where each pairwise comparison trial runs are conducted 3 - 5 times or more and the most frequent outcome is selected. This method could further mitigate the probabilistic nature of LLMs and yield more reliable pairwise judgments. Another factor would be prompt formulation. Although the same prompt templates were used across all models, it is possible that there are differences in how each LLM interprets the instructions, which could therefore affect the results of the study. The third concern would be data filtering and missing comparisons. 5 out of 750 arguments were dropped after fitting the BT model, due to undefined scores. There is the possibility that removing these cases could lead to bias in the results. However, given that the data was less than 1%, the impact could be minimal on the overall outcome.

### 9.2 External Validity

First, the evaluation was based on a specific dataset consisting of arguments drawn from an online debate portal on a set of different topics. These arguments are mostly a few sentences. Therefore, the results achieved may not generalize to other more complex forms, such as essays. There is a possibility that it could be (more or less) challenging for the LLM. Another aspect is that the expert scores used as ground truth came from a specific annotation process and reflect the judgments of the annotators shaped by their own interpretation of the dimensions. In another context, the different dimensions could be understood differently, meaning the LLM's alignment with one set of experts might not hold if used in another set. For this research, the zero-shot approach was used where the model is given instructions to compare the arguments without fine-tuning or additional training. While this approach is practical, it does not take into account the performance gains from fine-tuning the LLM or perhaps through few-shot prompting. The results do not explain how well the models would do in those particular settings. In future work, it would be valuable to assess whether alternative techniques could improve LLM accuracy, especially in dimensions like logical soundness, where current performance is limited. Finally, LLMs are evolving rapidly. The LLM versions used in this study are not the latest and could probably have been surpassed by newer versions. Although comparative and qualitative insights might remain relevant, metric values may shift even further with advances in LLM.

### 9.3 Construct Validity

Argument quality was split into three dimensions (logical, rhetorical, dialectical) according to the dataset's scheme for annotation. There is a possibility that the LLMs comparisons do not align with the intended construct. For example the LLM might rate the argument's logical quality on confident phasing even if the actual argument itself is logically flawed. To mitigate this, phrasing the prompts clear with criteria to steer the model towards the intended idea was conducted; however, LLMs are known to be black boxes. Comparing them to human judgment without 100%certainty that they have the same understanding of the 'quality' dimensions is very challenging. Furthermore, argument quality overall and other qualities were not assessed, since the study focused on the three dimensions of argument quality. While this approach provided insights that were valuable, it did not encompass every aspect related to argument quality such as precision and completeness. The dataset used for the gold standard comes from different experts who made pairwise comparisons. Any noise and/or bias could affect the results. The WebisArgQuality dataset is assumed as a valid measure for argument quality, but it is possible that there could have been inconsistencies in the annotation process or maybe the experts had their own biases. Another consideration is the separation of quality into three dimensions. Each dimension is treated

independently for evaluation purposes, but there was no explicit reminder in the prompt to ignore the other dimensions, except the specific prompt template that highlighted the criteria. Therefore, there is a possibility that the LLM might try to bring in other dimensions despite the intention to isolate. If the LLM for example took rhetorical into account for logical dimension analysis, then results could be blurry.

### 9.4 Conclusion Validity

One limitation related to conclusion validity is seen through the number of pairwise comparisons generated for this study. Although 1,500 comparisons across 750 arguments provided a substantial dataset, prior work by Gienapp (Gienapp et al., 2020) demonstrated the benefits of using far larger comparisons sets (over 41,000). Larger datasets may increase the statistical robustness of the BT model, and improve the stability of resulting rankings. In this study, the smaller comparison set was chosen to balance computational and financial strains, however it is possible that a larger set of comparisons would have produced more rankings closely aligned with expert decisions. Future research could explore scaling up number of comparisons.

## 10 Ethical Considerations

The assessment of argument quality aims, in general, at the societal benefit since better, stronger, higher quality arguments do not contain fallacies and spurious reasoning. Therefore, the intention of this work is to promote the development of high-quality, grounded discussions. At the same time, two ethical risks are inherent to this work. The first is that, by strengthening the quality of arguments supporting them, unethical or armful claims are spread. For instance, hateful messages could be supported by rhetorically high-quality arguments. One possible solution to this issue is to accompany argument quality assessment with hate speech detection or similar analyses. The second risk is that, by demanding quality assessment to LLMs, intentional or unintentional distortions are introduced in the analysis and moderation of public discourse. For example, models could show biases that favor some types of arguments over others. For this reason, as indicated also above, it is important to include human oversight in these moderation systems, and employ automated models in LLMs-in-the-loop systems.

## References


Evidently AI. 2025. Llm-as-a-judge: a complete guide to using llms for evaluations.

J Anthony Blair. 2011. *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair*, volume 21. Springer Science & Business Media.

Lisa Andreevna Chalaguine and Claudia Schulz. 2017. Assessing convincingness of arguments in online debates with limited number of features. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 75–83.

Alok Choudhary. 2024. All about: Mean squared error (mse), mean absolute error (mae) and rmse.

Javier Coronado-Blázquez. 2025. Deterministic or probabilistic? the psychology of llms as random number generators. *arXiv preprint arXiv:2502.19965*.

Learn Statistics Easily. 2024. Kendall tau-b vs spearman: Which correlation coefficient wins?

Vrunda Gadesha. 2025. Prompt engineering techniques.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781, Online. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. *CoRR*, abs/1911.11408.

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, Berlin, Germany. Association for Computational Linguistics.

David R Hunter. 2004. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406.

Daniel Kyne. 2022. Pairwise comparison (definition, methods, examples, tools).

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

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.

Sabbha Mondal. 2024. Understanding mae, mse, and rmse: Key metrics in machine learning.

Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. 2024. Are large language models more honest in their probabilistic or verbalized confidence? In *China Conference on Information Retrieval*, pages 124–135. Springer.

David Nordstokke and Andrea M. Stelnicki. 2014. *Pairwise Comparisons*, pages 4575–4576. Springer Netherlands, Dordrecht.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K Gurbani. 2023. How susceptible are llms to logical fallacies? *arXiv preprint arXiv:2308.09853*.

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. *CoRR*, abs/1909.01007.

Malhar Ujawane. 2024. Understanding llm settings for optimal performance. *Medium*.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*, pages 1680–1691.

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*.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. 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.

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

Bianca Williams, Caroline Halloin, Wiebke Löbel, Ferdous Finklea, Elizabeth Lipke, Robert Zweigerdt, and Selen Cremaschi. 2020. Data-driven model development for cardiomyocyte production experimental failure prediction. In Sauro Pierucci, Flavio Manenti, Giulia Luisa Bozzano, and Davide Manca, editors, *30th European Symposium on Computer Aided Process Engineering*, volume 48 of *Computer Aided Chemical Engineering*, pages 1639–1644. Elsevier.

C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, and A. Wesslén. 2012. *Experimentation in Software Engineering - An Introduction*. Kluwer Academic Publishers, International Series in Software Engineering.

# A APPENDIX

## A.1 Prompt Structure

Below are the structured prompts used in this study:

### A.1.1 Prompt Version 1

: (used in all three evaluation trials): This version explicitly includes the definitions of the relevant sub-criteria for each argument quality dimension. The objective was to ensure that the models had a clearer understanding of what each dimension entailed, potentially improving their ability to focus on the intended evaluation aspect.

**Logical**: "*"You are given two arguments: Argument A and Argument B. Decide which one is logically stronger based on these criteria only:"*

- *"- which is more acceptable/credible"*
- *"- which is more relevant to a conclusion"*
- *"- which is more sufficient to justify a conclusion"*

*"Reply with only one of the following options: A, B, or tie. Do NOT add any other text."*

**Rhetorical**: " *"You are given two arguments: Argument A and Argument B. Decide which one is rhetorically stronger based on these criteria only:*

- *"- which appears more authoritative/trust worthy."*
- *"- which makes a stronger emotional appeal."*

- *"- which is clear and more appropriate in tone."*

*"Reply with only one of the following options: A, B, or tie. Do NOT add any other text."*
**Dialectical**: " *"You are given two arguments: Argument A and Argument B. Decide which one is dialectically stronger based on these criteria only:*

- *"- which would be acceptable to the audience."*
- *"- which contributes more to constructive dialogue."*
- *"- which better anticipates or refutes counterarguments."*

*"Reply with only one of the following options: A, B, or tie. Do not provide any explanation."*

### A.1.2 Prompt Version 2

: (used in a single exploratory run): This version omits the explicit criteria and refers only to the dimension (e.g., Decide which one is logically stronger). This was intended to stimulate a more open-ended setting in which the model is allowed to rely on its internal representation of the dimension. However, due to time and resource constraints, this prompt version was tested in a single trial and is considered exploratory. Below are the prompt examples:
**Logical**: "*"You are given two arguments: Argument A and Argument B. Decide which one is logically stronger." "Reply with only one of the following options: A, B, or tie. Do NOT add any other text." "*
**Rhetorical**: " *"You are given two arguments: Argument A and Argument B. Decide which one is rhetorically stronger. "Reply with only one of the following options: A, B, or tie. Do NOT add any other text."*
**Dialectical**: " *"You are given two arguments: Argument A and Argument B. Decide which one is dialectically stronger. "Reply with only one of the following options: A, B, or tie. Do not provide any explanation."*