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

Nicolás Benjamín Ocampo[1], Agnes Paullate Nyiranziza[2], Davide Ceolin[1]
[1]HCDA, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
n.b.ocampo@cwi.nl, paullate.nziza@gmail.com, d.ceolin@cwi.nl

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. We investigate the extent to which LLMs can effectively perform this task. We tested 12 open-weight LLMs of different sizes and families under zero-shot, few-shot, and chain-of-thought to approximate human pairwise comparisons of argument quality across three dimensions—logical, rhetorical, and dialectic—and used these comparisons in a Bradley-Terry model to infer latent strength scores and derive a ranking of arguments. Our insights show that LLMs have promising but moderate correlation with human judgment, with Llama-70B obtaining the strongest alignment, reaching moderate Cohen's $\kappa = 0.493$ and moderate correlations with Bradley–Terry scores derived from these annotations (Kendall, Pearson, and Spearman: 0.327–0.477). Other LLMs exhibit weak, moderate, or high alignment with Llama-70B while achieving comparable results against human judgment, suggesting partial but complementary understanding of underlying quality dimensions despite differences in model size and family. Moreover, LLM predictions are stable across trial runs, with fewer than 7.75% of cases yielding different labels. Remaining variability is handled via majority voting and few-shot prompting for large-size models.

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 (Habernal and Gurevych, 2016; Wachsmuth et al., 2017b; Toledo et al., 2019; Vecchi et al., 2021; Joshi et al., 2023; Wachsmuth et al., 2024). A particularly effective approach within the field has been the use of pairwise comparisons (Gienapp et al., 2020; Habernal and Gurevych, 2016; Toledo et al., 2019; Gretz et al., 2020), 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 PageRank (Habernal and Gurevych, 2016), Gaussian Process (Simpson and Gurevych, 2018), and Bradley-Terry (Gienapp et al., 2020), to produce quality scores and obtain a ranking of evaluated arguments. This design reduces annotation cost and variance while providing reliable multidimensional quality scores for benchmarking automated evaluation methods. As online platforms grow, the ability to identify high-quality arguments becomes increasingly important, especially for effective persuasion, decision-making, and participation in meaningful discourse (Wachsmuth et al., 2024). However, manually evaluating argument quality at scale remains a major bottleneck due to cost and scalability.

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 LLM is asked to evaluate text quality (Gu et al., 2026). This strategy could enable large-scale assessments with minimal cost and time. However, whether LLMs can reliably replicate human-level judgments is an open question. Existing research offers mixed results: some studies show that LLMs, such as GPT-3 and PaLM-2 can effectively assess certain quality aspects (Mirzakhmedova et al., 2024), while others found that GPT-3.5 underperformed specialized su-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 24842–24857
October 24-29, 2026 ©2026 Association for Computational Linguistics

pervised models on tasks that involve ranking of arguments (Wang et al., 2023). These contrasting findings motivate a deeper investigation into LLM-based argument quality evaluation. This study attempts to replicate the argument quality rankings from the Webis-ArgQuality-20 dataset (Gienapp et al., 2020) that were obtained by modeling, via Bradley-Terry, pairwise comparisons annotations done by human crowdsourcers. 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. We aim to address the following research question (RQ):

RQ: How well can LLM pairwise comparisons reproduce human rankings of argument quality?

We break it down through the following sub-research questions (SRQ):

1) **SRQ1**: *How do LLMs of different sizes and families, when using various prompting strategies, compare in their ability to evaluate arguments on quality dimensions compared against human judgment?*

2) **SRQ2**: *To what extent do LLMs of varying sizes and model families exhibit similar patterns of argument evaluation?*

3) **SRQ3**: *How consistent are LLM-based evaluations when repeated over multiple trials?*

Our contribution is an extensive evaluation of 12 open-weight LLMs from five model families (7B–104B parameters) under zero-shot, few-shot, and chain-of-thought for pairwise argument assessment across logical, rhetorical, and dialectical dimensions. Using LLM pairwise judgments, we derive Bradley-Terry rankings, being, to the best of our knowledge, the first application of Bradley-Terry ranking to LLM-based comparative judgments in argument mining. Leveraging this contribution, we derive three important insights: (i) Our results show that LLM-derived BT rankings achieve moderate correlation with human judgments (up to 0.327–0.477 across Kendall, Pearson, and Spearman), with moderate human-level pairwise agreement (up to Cohen's κ=0.493), highlighting both the promise and current limitations of LLMs as evaluators of argument quality. (ii) Other LLMs show weak-to-high alignment with the best performing model, Llama-70B, while matching human performance, indicating complementary understanding of quality dimensions across models. (iii) LLM predictions are stable, with label differences in at most 7.75% across three runs. Few-shot prompting further improves stability for larger models, whereas smaller models are inherently more stable but less accurate [1].

2 Related Work

Argument quality assessment is relevant, as people tend to seek the strongest arguments to form opinions, write persuasive texts, foster agreement, and improve mutual understanding when presenting ideas to an audience (Wachsmuth et al., 2024). To address this need, several frameworks conceptualize argument quality as a multidimensional construct. For example, the taxonomy proposed by Wachsmuth et al. (2017b) distinguishes logical, rhetorical, and dialectical dimensions of argument quality, providing a strong theoretical foundation for computational assessment. Complementing this perspective, Vecchi et al. (2021); Friess and Eilders (2015) emphasize deliberative norms that focus on the collaborative and constructive aspects of discourse, Habernal and Gurevych (2016); Toledo et al. (2019) focus on argument convincingness, and Joshi et al. (2023) operationalize argument quality through a recommendation criterion, namely whether an argument would be suitable for use, as-is, in a supporting or opposing speech.

Using these theoretical frameworks, several datasets for argument quality assessment have been proposed that vary in number of topics and arguments, ranking strategies, evaluation dimensions, research communities, number of annotators, and annotation methods (Romberg et al., 2025). Two main approaches have been adopted for evaluation: Likert or continuous scale judgments of individual arguments (Swanson et al., 2015; Wachsmuth et al., 2017b; Potthast et al., 2019; Lauscher et al., 2020) and, pairwise comparison setups in which annotators select the better of two arguments with respect to a given dimension (Habernal and Gurevych, 2016; Toledo et al., 2019; Gretz et al., 2020; Gienapp et al., 2020). Pairwise comparisons avoid the problems of absolute ratings by not requiring assessors to assign fixed labels that are often treated as interval-scale, and instead asking them to simply choose between two items, which is easier, more consistent, and produces more reliable data (Gienapp et al., 2020). Subsequent scores are later ob-

[1] Code, and LLM generations and rankings: https://github.com/benjaminocampo/arg_quality_with_llms.

tained using a ranking function such as PageRank (Page et al., 1999) (or variants of it (Habernal and Gurevych, 2016)), Gaussian Process (Simpson and Gurevych, 2018), or Bradley-Terry (Hunter, 2004; Chen et al., 2013).

As pairwise comparisons scale rapidly, recent studies have explored using LLMs to assess argument quality. Mirzakhmedova et al. (2024); Wang et al. (2023); Deshpande et al. (2024) investigated agreement between LLMs and human annotators using models like GPT-3, GPT-3.5, and PaLM-2. Rescala et al. (2024) examined LLMs' ability to assess argument convincingness, Elaraby et al. (2024) studied the role of rationales in pairwise quality judgments, Jin et al. (2025) proposed a multi-persona framework based on Llama models, and De la Iglesia et al. (2025) used LLM-based ranking with medical arguments. Still, these works evaluate only a small set of LLMs, rely on Likert-scale, or focus on convincingness rather than multidimensional argument quality. In contrast, we systematically compare multiple LLM families and sizes across prompts using the broader theoretical framework of Wachsmuth et al. (2017b), operationalized through the Webis-ArgQuality-20 dataset. We assess argument quality via pairwise comparisons and aggregate judgments with the Bradley-Terry model, a well-established and computationally efficient approach (Hunter, 2004; Chen et al., 2013).

3 Methodology

The overall methodology consists of performing comparisons of argument pairs using zero-shot, few-shot, and chain-of-thought. Arguments are aggregated per topic (only topic-related arguments are compared), and the comparison is performed across three dimensions of argument quality.

3.1 Argument Quality Dimensions

Theoretical work in argumentation distinguishes three dimensions: logical, rhetorical, and dialectical (Blair, 2012; Wachsmuth et al., 2017b). **Logical quality**, or *cogency* checks whether premises are relevant and sufficient to conclude. Wachsmuth et al. (2017a) defines 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?). **Rhetorical quality** or *effectiveness* measures whether the argument can persuade a target audience. Inspired by Aristotle's rhetorical appeals, Wachsmuth et al. (2017b) 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). **Dialectic quality**, or *reasonableness* measures the ability of the argument to help resolve a disagreement. 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 Bradley-Terry

Pairwise comparisons are a widely used method in statistical analysis. In argument quality assessment, this approach allows annotators to reduce cognitive load and annotation time by focusing on comparing two arguments at a time, instead of ranking a whole set. Once we obtain these comparisons, we use the Bradley-Terry model to derive continuous quality scores. Given a pair of arguments d_i and d_j, BT estimates the probability that d_i is preferred to d_j ($d_i \succ d_j$) through the formula (Hunter, 2004):

$$P(d_i \succ d_j) = \frac{\gamma_i}{\gamma_i + \gamma_j}$$

If ties are allowed in pairwise comparisons, the Bradley–Terry model can be extended in several ways, including the formulations proposed by Rao and Kupper (1967); Gienapp et al. (2020). In our study, we instead adopted a simpler approximation, treating each tie as two reciprocal comparisons, one in each direction.

4 Dataset

We used the Webis-ArgQuality-20 corpus, an argument quality dataset introduced by Gienapp et al. (2020), comprising 1271 arguments across 20 debate topics which were subjected to 13952 pairwise comparisons obtained via cyclic group sampling. Crowdsourcers annotated, for each topic and quality dimension (Logical, Rhetorical, or Dialectical Quality), which argument in each pair was better with respect to that specific quality dimension. These ground-truth labels were subsequently used to rank argument quality across dimensions using a Bradley-Terry model adjusted for ties. Only one

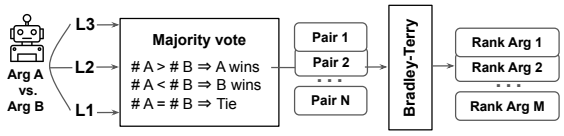


Figure 1: Pipeline overview of the ranking methodology. LLM pairwise comparison labels (L1–L3) are first resolved by majority vote. All the resulting pairwise labels are then aggregated with the Bradley–Terry model to produce a ranking of all arguments.

crowdsourcer per pair was used, so inter-annotator agreement for the comparisons in the dataset is not reported. Instead, to assess annotation quality, Gienapp et al. (2020) relied on the Amazon Mechanical Turk HIT approval rate, requiring each annotator to have an approval rate of at least 95%. This shifts the quality control mechanism to the annotator level. We used the entire set of pairwise comparisons provided in the dataset in our experiments.

5 Experimental setup

We evaluate 12 open-weight models from five families: Mistral, Llama 3*, Qwen2.5, Olmo2, and Command-R, spanning parameter scales from 7B to 104B [2]. Each family includes models of different sizes, enabling a controlled analysis of the effects of both model scale and architecture on output variability and inter-model agreement. We focus exclusively on open-weight models to ensure reproducibility and transparency.

We evaluate whether LLMs can replicate human-derived argument quality rankings published for the Webis-ArgQuality-20 corpus. The approach follows the original pairwise framework where each comparison task presents the two arguments—named *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.

We employed three prompting strategies: zero-shot, few-shot, and chain-of-thought. Zero-shot takes as input the dimension to judge and the pairwise comparison to assess. Chain-of-thought extends the zero-shot strategy by instructing the LLM to reason internally over sub-dimensions associated with each main dimension, following the framework of Wachsmuth et al. (2017b) described in Section 3.1. Logical quality is assessed through local acceptability, local relevance, and local sufficiency; rhetorical quality through credibility, emotional appeal, clarity, appropriateness, and arrangement; and dialectical quality through acceptability, relevance, and sufficiency. Each sub-dimension is reformulated as a comparative question asking which of the two arguments exhibits more of that property, accompanied by a brief definition. The model outputs only the final label for the target dimension, not intermediate sub-dimension judgments, as evaluation is conducted at the dimension level. Thus, the sub-dimensions function as structured guidance for the model's internal reasoning process rather than as prediction targets in their own right. Finally, few-shot uses the same definitions as zero-shot but dynamically adds, for the pair of arguments under evaluation, three annotated examples within their topic: one where "`A`" is chosen, one where "`B`" is chosen, and one labeled as "`Tie`". These examples remain the same across all comparisons within that topic.

Specifically, we consider arguments A and B, and for the label A, we select the pair with the largest score difference where the score of A is higher than that of B. Similarly, we choose an example for B by selecting the pair with the largest score difference where the score of B is higher than that of A. For the *Tie* example, we select, from the ground-truth comparisons labeled as *Tie*, the pair with the smallest quality score difference for that dimension and topic. In this way, we ensure that the few-shot examples are clear-cut and avoid ambiguous examples [3].

Regarding temperature, values in the 0–1 range show no statistically significant differences in performance, while values above 1 degrade it (Renze, 2024). We set temperature to 1 to balance stability and stochasticity for measuring intra-model reliability, with gains expected to generalize to lower settings (Renze, 2024; Li et al., 2025).

Given the probabilistic nature of LLMs, we evaluated each model across three independent runs using the same pairing strategy to measure prediction variability. The final label was determined by majority vote. The argument (A or B) with the highest frequency was selected; in cases where no single argument secured a majority (e.g., a three-

[2]The specific version of each LLM used in this study is provided in Appendix A.

[3]Each of the prompts is provided in Appendix B.

Model	Prompt	Pearson (r)	Spearman (ρ)	Kendall	MAE	RMSE	κ IAA	% Agreement
Mistral-7B	zero	0.232 ± 0.021	0.232 ± 0.019	0.156 ± 0.012	0.973 ± 0.003	1.239 ± 0.017	0.162 ± 0.060	26.88 ± 5.16
	few	0.082 ± 0.183	0.082 ± 0.181	0.055 ± 0.123	1.046 ± 0.112	1.350 ± 0.138	0.332 ± 0.133	39.49 ± 9.85
	cot	0.221 ± 0.034	0.213 ± 0.033	0.143 ± 0.022	0.967 ± 0.021	1.248 ± 0.028	0.095 ± 0.051	22.39 ± 4.47
Llama-8B	zero	0.258 ± 0.019	0.263 ± 0.018	0.176 ± 0.011	0.968 ± 0.010	1.218 ± 0.015	0.373 ± 0.016	42.68 ± 1.33
	few	0.155 ± 0.035	0.155 ± 0.024	0.104 ± 0.016	1.020 ± 0.009	1.300 ± 0.027	0.424 ± 0.006	46.62 ± 0.67
	cot	0.057 ± 0.051	0.056 ± 0.047	0.037 ± 0.032	1.080 ± 0.022	1.373 ± 0.037	0.303 ± 0.012	37.01 ± 1.01
Olmo2-7B	zero	-0.034 ± 0.044	-0.032 ± 0.031	-0.021 ± 0.021	1.121 ± 0.029	1.438 ± 0.031	0.331 ± 0.017	39.12 ± 1.36
	few	-0.087 ± 0.032	-0.103 ± 0.015	-0.069 ± 0.010	1.149 ± 0.016	1.474 ± 0.022	0.401 ± 0.007	44.51 ± 0.74
	cot	0.002 ± 0.060	-0.003 ± 0.084	-0.002 ± 0.056	1.105 ± 0.048	1.412 ± 0.042	0.356 ± 0.031	41.07 ± 2.44
Qwen2.5-7B	zero	0.322 ± 0.031	0.326 ± 0.033	0.221 ± 0.023	0.918 ± 0.018	1.164 ± 0.026	0.431 ± 0.010	47.51 ± 0.87
	few	0.354 ± 0.042	0.359 ± 0.032	0.244 ± 0.021	0.882 ± 0.022	1.137 ± 0.038	0.456 ± 0.004	49.57 ± 0.34
	cot	0.188 ± 0.155	0.201 ± 0.158	0.135 ± 0.108	1.007 ± 0.103	1.270 ± 0.123	0.399 ± 0.035	44.87 ± 2.96
Command-R-7B	zero	0.370 ± 0.067	0.363 ± 0.068	0.247 ± 0.048	0.887 ± 0.051	1.122 ± 0.059	0.417 ± 0.007	46.34 ± 0.62
	few	0.301 ± 0.146	0.302 ± 0.145	0.205 ± 0.099	0.924 ± 0.097	1.178 ± 0.121	0.376 ± 0.011	43.04 ± 1.01
	cot	0.345 ± 0.031	0.344 ± 0.043	0.233 ± 0.030	0.903 ± 0.029	1.144 ± 0.027	0.340 ± 0.012	40.25 ± 0.99
Mixtral-8x7B	zero	0.174 ± 0.049	0.172 ± 0.047	0.115 ± 0.032	1.011 ± 0.021	1.285 ± 0.038	0.339 ± 0.033	40.09 ± 2.56
	few	0.233 ± 0.170	0.246 ± 0.150	0.166 ± 0.103	0.951 ± 0.094	1.233 ± 0.137	0.392 ± 0.063	44.22 ± 4.88
	cot	0.093 ± 0.119	0.084 ± 0.107	0.057 ± 0.073	1.047 ± 0.077	1.345 ± 0.088	0.277 ± 0.060	35.69 ± 3.80
Mistral-22B	zero	-0.086 ± 0.035	-0.094 ± 0.016	-0.063 ± 0.011	1.142 ± 0.009	1.473 ± 0.024	0.381 ± 0.003	42.90 ± 0.20
	few	-0.047 ± 0.019	-0.053 ± 0.012	-0.035 ± 0.008	1.128 ± 0.009	1.447 ± 0.013	0.378 ± 0.007	42.65 ± 0.47
	cot	-0.072 ± 0.060	-0.084 ± 0.045	-0.056 ± 0.030	1.141 ± 0.031	1.464 ± 0.041	0.370 ± 0.013	42.14 ± 0.94
Olmo2-32B	zero	0.402 ± 0.018	0.377 ± 0.024	0.257 ± 0.017	0.860 ± 0.010	1.094 ± 0.017	0.366 ± 0.036	42.41 ± 2.87
	few	0.296 ± 0.081	0.295 ± 0.083	0.199 ± 0.057	0.937 ± 0.049	1.185 ± 0.068	0.392 ± 0.023	44.25 ± 1.86
	cot	0.398 ± 0.026	0.380 ± 0.028	0.258 ± 0.020	0.869 ± 0.020	1.097 ± 0.023	0.402 ± 0.028	45.29 ± 2.30
Mixtral-8x22B	zero	0.237 ± 0.035	0.227 ± 0.023	0.152 ± 0.016	0.978 ± 0.022	1.235 ± 0.029	0.311 ± 0.055	37.95 ± 4.42
	few	0.290 ± 0.091	0.273 ± 0.083	0.185 ± 0.058	0.938 ± 0.068	1.190 ± 0.076	0.412 ± 0.016	45.95 ± 1.39
	cot	0.218 ± 0.029	0.217 ± 0.011	0.145 ± 0.008	1.000 ± 0.018	1.250 ± 0.024	0.371 ± 0.032	42.66 ± 2.28
Llama-70B	zero	0.442 ± 0.005	0.441 ± 0.012	0.301 ± 0.010	0.836 ± 0.002	1.056 ± 0.004	0.487 ± 0.008	52.33 ± 0.77
	few	**0.473 ± 0.004**	**0.477 ± 0.012**	**0.327 ± 0.008**	**0.811 ± 0.006**	**1.027 ± 0.004**	**0.493 ± 0.001**	**52.93 ± 0.14**
	cot	0.465 ± 0.019	0.465 ± 0.016	0.320 ± 0.012	0.812 ± 0.017	1.034 ± 0.019	0.487 ± 0.011	52.43 ± 0.98
Qwen2.5-72B	zero	0.463 ± 0.003	0.466 ± 0.006	0.319 ± 0.005	0.817 ± 0.009	1.036 ± 0.003	0.467 ± 0.006	50.82 ± 0.40
	few	0.453 ± 0.024	0.459 ± 0.029	0.314 ± 0.021	0.824 ± 0.026	1.046 ± 0.023	0.445 ± 0.016	49.04 ± 1.30
	cot	0.435 ± 0.017	0.443 ± 0.019	0.302 ± 0.013	0.846 ± 0.022	1.063 ± 0.016	0.451 ± 0.023	49.55 ± 1.83
Command-R-104B	zero	0.365 ± 0.041	0.362 ± 0.044	0.246 ± 0.031	0.876 ± 0.029	1.127 ± 0.037	0.423 ± 0.011	47.07 ± 0.95
	few	0.392 ± 0.012	0.381 ± 0.014	0.260 ± 0.012	0.864 ± 0.015	1.103 ± 0.011	0.406 ± 0.040	45.64 ± 3.21
	cot	0.354 ± 0.027	0.357 ± 0.025	0.242 ± 0.019	0.885 ± 0.020	1.136 ± 0.023	0.428 ± 0.032	47.42 ± 2.58
Majority Baseline		0.103 ± 0.029	0.121 ± 0.028	0.081 ± 0.019	1.023 ± 0.010	1.339 ± 0.021	0.427 ± 0.014	46.87 ± 1.37
Random Baseline		-0.021 ± 0.024	-0.014 ± 0.029	-0.009 ± 0.019	1.127 ± 0.017	1.429 ± 0.017	0.286 ± 0.005	35.83 ± 0.26

Table 1: LLM and prompting configurations scores and annotations compared against human judgment. Scores were averaged across logic, rhetoric, and dialectic dimensions, with standard deviations (±) reported across them. Pearson, Spearman, and Kendall coefficients measure correlation between LLM BT rankings and human BT rankings. MAE and RMSE measure LLM-human BT ranking differences, where lower values are better. Cohen's κ and percent agreement (%) compare LLM annotations directly rather than BT scores. The best configuration per column is in **bold**; the second-best is underlined.

way split or a tie involving an explicit 'Tie' model response), the final label was Tie. Figure 1 shows the overall ranking pipeline. First, the LLM evaluates each pairwise comparison three independent times, choosing between "A," "B," or "Tie." Majority voting is then used to obtain a single label for each pair. Once all 13952 pairs are annotated, the resulting labels are used to fit the Bradley–Terry model and obtain rankings for the 1271 arguments in the dataset.

5.1 Evaluation metrics

To assess the reliability and consistency of LLM-based argument quality evaluation, we did three complementary analyses. First, in the **LLM vs. Human** analysis, we compared LLM-generated pairwise preferences and the resulting Bradley-Terry scores against those derived from the Webis-ArgQuality-20 human annotations across the logic, rhetoric, and dialectic dimensions. Second, in the **Best LLM vs. Other LLMs** analysis, we selected the highest-performing model and compared its pairwise annotations and resulting BT scores against those produced by other LLMs across the same argument pairs. Third, in the **Prediction Variability** analysis, we assessed the stability of each LLM by repeating the evaluation process across three independent runs and measuring the agreement between the resulting pairwise annotations.

We evaluated Bradley-Terry score alignment using Pearson (r), Spearman (ρ), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Annotation agreement was measured using Cohen's κ, Krippendorff's α, and percent agreement.

To provide reference values for the evaluation metrics, we defined two baselines: the Random Baseline, which randomly selects between "A", "B", and "Tie" for each comparison, and the Majority Baseline, which always predicts the label most frequently chosen by the human annotators,

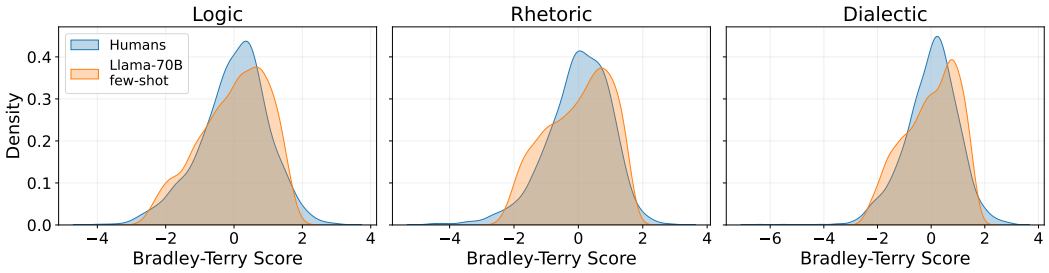


Figure 2: Distribution of Bradley-Terry scores assigned by humans and Llama-70B (few-shot).

Dim	Model	Pearson (r)	Spearman (ρ)	Kendall	MAE	RMSE	κ IAA	% Agreement
Logic	Random Baseline	-0.041	-0.027	-0.018	1.139	1.443	0.288	35.64
	Majority Baseline	0.125	0.153	0.103	1.014	1.323	0.442	48.28
	Llama-70B Few-shot	0.474	0.479	0.329	0.806	1.026	0.493	53.00
Rhetoric	Random Baseline	0.006	0.019	0.013	1.107	1.410	0.281	35.72
	Majority Baseline	0.071	0.104	0.069	1.033	1.363	0.414	45.55
	Llama-70B Few-shot	0.476	0.488	0.333	0.808	1.024	0.493	52.76
Dialectic	Random Baseline	-0.029	-0.034	-0.023	1.135	1.434	0.290	36.12
	Majority Baseline	0.113	0.104	0.070	1.021	1.332	0.426	46.78
	Llama-70B Few-shot	0.469	0.464	0.318	0.818	1.031	0.494	53.02

Table 2: Results per dimension of the best performing configuration LLama-70B with few-shot.

namely A for the logic, rhetoric, and dialectic dimensions (see data distribution in Appendix D). We then derived the corresponding BT scores from the comparisons produced by these baselines.

6 Results & Discussion

To address **SRQ1**, we begin by identifying the best-performing model and prompt configuration for evaluating argument quality. Table 1 compares rankings and annotations produced by humans with those generated by LLMs under zero-shot (zero), few-shot (few), and chain-of-thought (cot) prompting strategies, across 36 configurations comprising 12 models and 3 prompting strategies. Llama-70B with few-shot prompting achieves the highest average correlations across the Pearson, Spearman, and Kendall metrics over the logic, rhetoric, and dialect dimensions, with averaged scores of 0.473, 0.477, and 0.327, respectively. It is followed by Qwen2.5-72B under zero-shot prompting, which attains corresponding averaged scores of 0.463, 0.466, and 0.319. Both models also exhibit low variability across the three dimensions and obtain scores well superior to random and majority baselines. Both configurations achieve the lowest error metrics, with MAE and RMSE of 0.811 and 1.027 for LLaMA-70B with few-shot, and 0.817 and 1.036 for Qwen2.5-72B with zero-shot, respectively. The highest Cohen's κ is obtained also by LLama-70B with few-shot with 0.493, while the second best is the same model under both zero-shot and chain-of-thought. Similarly, these configurations achieve a percent agreement of 52.93%, 52.43%, and 52.33% for few-shot, chain-of-thought, and zero-shot, respectively. To test for statistical significance, we run a one-tailed bootstrap test (N=10000) across all dimensions and metrics. At the 95% confidence level, Llama-70B few-shot is significantly better than all other configurations, except Llama-70B with chain-of-thought and Qwen-2.5-72B with zero and few-shot.

Note also that, Olmo2-7B and Mistral-22B, obtain negative correlation values, although these values remain close to zero. Nevertheless, both models correctly achieve a moderate IAA with respect to the ground truth. Olmo2-7B achieves an average Cohen's Kappa ranging from 0.331 to 0.401 across the quality dimensions, while Mistral-22B achieves an average Cohen's Kappa ranging from 0.366 to 0.381. Looking at the comparison distributions (Table A in Appendix D), both Olmo2-7B and Mistral-22B predict argument B more often across pairwise comparisons, regardless of the prompting configuration or quality dimension. Since B is the preferred label in approximately 40% of the ground-truth comparisons, this tendency can result in moderate inter-annotator agreement. However, these predictions are not sufficiently informative

Model	Prompt	Pearson (r)	Spearman (ρ)	Kendall	MAE	RMSE	κ IAA	% Agreement
Mistral-7B	zero	0.521 ± 0.032	0.531 ± 0.027	0.370 ± 0.020	0.787 ± 0.022	0.978 ± 0.033	0.234 ± 0.089	26.32 ± 8.30
	few	0.057 ± 0.242	0.066 ± 0.249	0.044 ± 0.168	1.085 ± 0.139	1.366 ± 0.181	0.354 ± 0.150	39.47 ± 15.40
	cot	0.476 ± 0.063	0.492 ± 0.058	0.340 ± 0.042	0.823 ± 0.034	1.023 ± 0.061	0.134 ± 0.067	15.95 ± 6.04
Llama-8B	zero	0.497 ± 0.014	0.497 ± 0.021	0.344 ± 0.014	0.797 ± 0.011	1.003 ± 0.014	0.498 ± 0.035	53.22 ± 3.34
	few	0.228 ± 0.089	0.235 ± 0.072	0.158 ± 0.050	0.991 ± 0.058	1.241 ± 0.071	0.566 ± 0.028	60.09 ± 2.54
	cot	0.194 ± 0.095	0.196 ± 0.094	0.132 ± 0.064	1.020 ± 0.066	1.268 ± 0.073	0.388 ± 0.014	42.54 ± 1.37
Olmo2-7B	zero	-0.033 ± 0.088	-0.035 ± 0.079	-0.022 ± 0.053	1.140 ± 0.032	1.436 ± 0.062	0.328 ± 0.078	37.42 ± 7.29
	few	-0.064 ± 0.116	-0.074 ± 0.110	-0.050 ± 0.074	1.162 ± 0.057	1.457 ± 0.080	0.415 ± 0.077	46.14 ± 7.12
	cot	0.008 ± 0.070	0.005 ± 0.078	0.004 ± 0.052	1.121 ± 0.033	1.408 ± 0.049	0.371 ± 0.089	41.62 ± 8.51
Qwen2.5-7B	zero	0.640 ± 0.051	0.640 ± 0.052	0.459 ± 0.043	0.658 ± 0.050	0.847 ± 0.059	0.640 ± 0.023	66.57 ± 2.18
	few	0.615 ± 0.057	0.626 ± 0.056	0.450 ± 0.046	0.667 ± 0.055	0.876 ± 0.066	0.693 ± 0.019	71.64 ± 1.69
	cot	0.496 ± 0.164	0.494 ± 0.176	0.348 ± 0.132	0.777 ± 0.133	0.995 ± 0.164	0.542 ± 0.103	57.36 ± 9.65
Command-R-7B	zero	0.453 ± 0.098	0.451 ± 0.102	0.311 ± 0.076	0.830 ± 0.085	1.044 ± 0.093	0.505 ± 0.055	54.17 ± 5.10
	few	0.382 ± 0.203	0.384 ± 0.203	0.263 ± 0.141	0.875 ± 0.137	1.102 ± 0.176	0.452 ± 0.075	49.07 ± 6.91
	cot	0.441 ± 0.041	0.438 ± 0.043	0.299 ± 0.031	0.845 ± 0.028	1.057 ± 0.038	0.419 ± 0.035	45.64 ± 3.17
Mixtral-8x7B	zero	0.343 ± 0.067	0.338 ± 0.058	0.231 ± 0.042	0.904 ± 0.043	1.145 ± 0.059	0.396 ± 0.068	43.35 ± 6.69
	few	0.313 ± 0.243	0.327 ± 0.224	0.224 ± 0.155	0.920 ± 0.164	1.161 ± 0.200	0.491 ± 0.111	52.61 ± 11.22
	cot	0.181 ± 0.138	0.164 ± 0.128	0.111 ± 0.086	1.010 ± 0.080	1.277 ± 0.107	0.291 ± 0.055	32.93 ± 6.10
Mistral-22B	zero	-0.050 ± 0.106	-0.053 ± 0.100	-0.035 ± 0.067	1.146 ± 0.046	1.448 ± 0.074	0.385 ± 0.080	43.15 ± 7.41
	few	0.062 ± 0.025	0.055 ± 0.021	0.037 ± 0.014	1.092 ± 0.011	1.370 ± 0.018	0.435 ± 0.044	47.78 ± 4.20
	cot	-0.050 ± 0.135	-0.055 ± 0.137	-0.037 ± 0.093	1.149 ± 0.071	1.447 ± 0.095	0.369 ± 0.059	41.53 ± 5.33
Olmo2-32B	zero	0.666 ± 0.024	0.659 ± 0.025	0.472 ± 0.021	0.646 ± 0.027	0.817 ± 0.030	0.500 ± 0.048	53.17 ± 4.77
	few	0.605 ± 0.069	0.594 ± 0.079	0.420 ± 0.062	0.700 ± 0.067	0.886 ± 0.076	0.529 ± 0.042	56.32 ± 4.03
	cot	0.685 ± 0.019	0.673 ± 0.023	0.486 ± 0.019	0.623 ± 0.019	0.793 ± 0.024	0.553 ± 0.033	58.40 ± 3.19
Mixtral-8x22B	zero	0.552 ± 0.040	0.545 ± 0.044	0.380 ± 0.033	0.747 ± 0.035	0.946 ± 0.042	0.429 ± 0.094	46.22 ± 8.98
	few	0.509 ± 0.112	0.497 ± 0.109	0.348 ± 0.085	0.775 ± 0.097	0.986 ± 0.117	0.544 ± 0.055	57.66 ± 5.11
	cot	0.502 ± 0.019	0.513 ± 0.026	0.356 ± 0.019	0.774 ± 0.018	0.997 ± 0.019	0.499 ± 0.021	53.20 ± 2.30
Llama-70B	zero	**0.892 ± 0.003**	**0.892 ± 0.004**	**0.718 ± 0.005**	**0.341 ± 0.003**	**0.464 ± 0.007**	**0.831 ± 0.011**	**84.45 ± 1.05**
	few	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	100.00 ± 0.00
	cot	0.857 ± 0.026	0.861 ± 0.026	0.677 ± 0.032	0.396 ± 0.035	0.533 ± 0.049	0.795 ± 0.030	81.06 ± 2.83
Qwen2.5-72B	zero	0.829 ± 0.022	0.824 ± 0.021	0.632 ± 0.023	0.445 ± 0.027	0.584 ± 0.037	0.741 ± 0.023	76.04 ± 2.17
	few	0.819 ± 0.023	0.818 ± 0.021	0.625 ± 0.023	0.456 ± 0.028	0.600 ± 0.037	0.708 ± 0.030	72.83 ± 2.84
	cot	0.796 ± 0.005	0.793 ± 0.002	0.600 ± 0.004	0.485 ± 0.010	0.638 ± 0.007	0.691 ± 0.025	71.29 ± 2.36
Command-R-104B	zero	0.577 ± 0.077	0.592 ± 0.073	0.423 ± 0.060	0.705 ± 0.067	0.917 ± 0.086	0.608 ± 0.043	63.65 ± 4.06
	few	0.624 ± 0.034	0.629 ± 0.038	0.452 ± 0.033	0.672 ± 0.036	0.866 ± 0.040	0.593 ± 0.058	62.22 ± 5.59
	cot	0.563 ± 0.044	0.577 ± 0.046	0.412 ± 0.037	0.712 ± 0.042	0.934 ± 0.047	0.621 ± 0.057	64.90 ± 5.41

Table 3: LLM and prompt configurations scores and annotations compared against Llama-70B few-shot. The configuration most aligned with Llama-70B few-shot per column is in **bold**; the second-most aligned is underlined.

for BT because they provide little discrimination between arguments. As a result, the models produce BT rankings that correlate poorly with the ground-truth rankings, leading to negative values.

Performance by model size. Small and medium size models do not exhibit higher agreement with humans than large size ones. Across prompts, chain-of-thought may lead to lower correlation scores and agreement with human rankings and annotations compared to zero-shot. Overall, zero-shot, few-shot, and chain-of-thought exhibit minimal variability across the evaluated dimensions.

Performance by model family. Command-R-7B and Command-R-104B achieve comparable performance, with Command-R-104B obtaining slightly higher correlation and agreement scores. In most other model families, smaller variants perform worse than their larger counterparts. The main exception is the Mistral family, where Mistral-7B achieves higher correlation with human rankings across all prompting configurations than Mistral-22B, while also obtaining lower MAE and RMSE values.

Performance by prompt strategy. There is no systematic preference for a single prompting strategy across all models. Among small models, except for Qwen2.5-7B, few-shot prompting decreases correlation and agreement with humans relative to zero-shot prompting. A similar pattern is observed in medium-sized models, where few-shot prompting is generally detrimental, except for Mixtral-8x7B and Mixtral-8x22B. In contrast, among large models, few-shot prompting benefits Llama-70B and Command-R-104B.

Performance by quality dimension. Overall, configurations showed low variability across the logical, quality, and rhetorical dimensions, indicating that no specific dimension was harder to evaluate. Only Mixtral-8x7B and Command-R-7B under the few-shot setting presented higher variability compared to other configurations across Pearson, Spearman, Kendall, MAE, and RMSE. However, this variability was not reflected in Cohen's κ or percent agreement. Table 2 shows the highest performance achieved by the best configuration, Llama-70B with few-shot, highlighting that the three dimensions had similar difficulty levels [4]. Moreover, Figure 2 shows that, despite its performance, Llama-70B with few-shot produces a BT score distribution that differs from that of humans, consis-

[4]The full table of results for each model, prompt, and dimension can be found in the repository of the paper.

Model	Prompt	All equal		2 equal		All unequal		kripp. α
		#	%	#	%	#	%	
Mistral-7B	zero	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	few	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	cot	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
Llama-8B	zero	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	few	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	cot	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
Olmo2-7B	zero	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	few	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	cot	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
Qwen2.5-7B	zero	13565.67 ± 669.15	97.23 ± 4.80	386.33 ± 669.15	2.77 ± 4.80	0.00 ± 0.00	0.00 ± 0.00	0.97 ± 0.05
	few	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	cot	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
Command-R-7B	zero	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	few	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	cot	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
Mixtral-8x7B	zero	13105.00 ± 188.65	93.93 ± 1.35	847.00 ± 188.65	6.07 ± 1.35	0.00 ± 0.00	0.00 ± 0.00	0.92 ± 0.01
	few	12025.33 ± 1331.71	86.19 ± 9.54	1898.00 ± 1298.63	13.60 ± 9.31	28.67 ± 48.79	0.21 ± 0.35	0.78 ± 0.02
	cot	12567.00 ± 543.52	90.07 ± 3.90	1379.33 ± 533.88	9.89 ± 3.83	5.67 ± 9.81	0.04 ± 0.07	0.87 ± 0.06
Mistral-22B	zero	6787.00 ± 1466.48	48.65 ± 10.51	6739.67 ± 1419.55	48.31 ± 10.17	425.33 ± 61.09	3.05 ± 0.44	0.16 ± 0.04
	few	4899.67 ± 1359.89	35.12 ± 9.75	7971.67 ± 778.22	57.14 ± 5.58	1080.67 ± 584.11	7.75 ± 4.19	0.16 ± 0.07
	cot	5320.00 ± 999.99	38.13 ± 7.17	7830.00 ± 753.81	56.12 ± 5.40	802.00 ± 374.15	5.75 ± 2.68	0.11 ± 0.00
Olmo2-32B	zero	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	few	11310.00 ± 4576.08	81.06 ± 32.80	2642.00 ± 4576.08	18.94 ± 32.80	0.00 ± 0.00	0.00 ± 0.00	0.81 ± 0.33
	cot	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
Mixtral-8x22B	zero	9584.00 ± 585.15	68.69 ± 4.19	4229.00 ± 527.10	30.31 ± 3.78	139.00 ± 59.77	1.00 ± 0.43	0.63 ± 0.01
	few	6598.67 ± 847.45	47.30 ± 6.07	6379.33 ± 667.24	45.72 ± 4.78	974.00 ± 194.13	6.98 ± 1.39	0.39 ± 0.09
	cot	7869.00 ± 645.82	56.40 ± 4.63	5667.67 ± 632.36	40.62 ± 4.53	415.33 ± 80.08	2.98 ± 0.57	0.50 ± 0.06
Llama-70B	zero	13268.00 ± 64.63	95.10 ± 0.46	684.00 ± 64.63	4.90 ± 0.46	0.00 ± 0.00	0.00 ± 0.00	0.93 ± 0.01
	few	13952.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	cot	13210.33 ± 645.63	94.68 ± 4.63	740.00 ± 643.95	5.30 ± 4.62	1.67 ± 2.89	0.01 ± 0.02	0.93 ± 0.06
Qwen2.5-72B	zero	12743.33 ± 1053.37	91.34 ± 7.55	1208.67 ± 1053.37	8.66 ± 7.55	0.00 ± 0.00	0.00 ± 0.00	0.90 ± 0.09
	few	13347.67 ± 1046.74	95.67 ± 7.50	582.00 ± 1008.05	4.17 ± 7.23	22.33 ± 38.68	0.16 ± 0.28	0.95 ± 0.09
	cot	13433.00 ± 898.93	96.28 ± 6.44	519.00 ± 898.93	3.72 ± 6.44	0.00 ± 0.00	0.00 ± 0.00	0.96 ± 0.07
Commandr-104B	zero	10567.67 ± 3030.76	75.74 ± 21.72	3256.33 ± 2878.99	23.34 ± 20.63	128.00 ± 221.70	0.92 ± 1.59	0.67 ± 0.30
	few	12596.67 ± 2347.51	90.29 ± 16.83	1355.33 ± 2347.51	9.71 ± 16.83	0.00 ± 0.00	0.00 ± 0.00	0.86 ± 0.24
	cot	12541.00 ± 2443.92	89.89 ± 17.52	1315.33 ± 2278.22	9.43 ± 16.33	95.67 ± 165.70	0.69 ± 1.19	0.82 ± 0.31

Table 4: Variability across three runs per LLM and prompting configuration in the pairwise comparison annotation task. Reported are the proportions of instances with identical labels across all three runs (all equal), agreement between two of three runs (2 equal), and different labels across all runs (unequal), along with Krippendorff's α. Results are averaged across dimensions (± standard deviation).

tent with the moderate correlation observed. In particular, the LLM's distribution appears slightly more right-skewed.

6.1 Cross-LLM Comparison

In order to address **SRQ2**, Table 3 compares the rankings and annotations produced by Llama-70B with few-shot against those generated by the remaining models. While configurations other than Llama-70B with few-shot show lower results with respect to the ground truth, we also evaluate how closely other LLMs agree with this best configuration.

Among all configurations, Llama-70B consistently shows the highest correlations and agreement scores across zero-shot and chain-of-thought. For zero-shot it achieves 0.892, 0.892, and 0.718 for Pearson, Spearman, and Kendall, respectively, while 0.831 and 84.45% for Cohen's κ and percent agreement, respectively. Qwen2.5-72B follows closely under all prompting settings. Although Command-R-104B still shows substantial agreement, it consistently performs below the other large models, Llama-70B and Qwen2.5-72B.

Within the medium-sized models, Olmo2-32B is the closest to the best configuration. It surpasses Command-R-104B under zero-shot and chain-of-thought, while falling slightly behind under few-shot. The remaining medium-sized models show only moderate agreement. An exception is Mistral-22B, which obtains correlation scores close to 0 together with higher MAE and RMSE values.

The small models generally show moderate agreement and correlation scores. Qwen2.5-7B is the closest among the small models, reflecting the strong results of its larger 72B counterpart across all prompting settings. In contrast, Llama-8B achieves only moderate IAA and moderate-to-low correlation scores, together with higher MAE and RMSE values compared to the models with the strongest agreement. Similarly, Olmo2-7B does not align as well with the best configuration as its 32B counterpart, producing correlations close to 0. Command-R-7B, however, remains partially aligned as its larger version. The low variability across dimensions indicates that these findings are

consistent for each of them[5].

These results, together with those in Table 1, show that although Llama-70B is the best configuration and reaches moderate agreement, some groups of LLMs closely align with it and *share its vision*, while others are less aligned with it despite still agreeing with humans, reflecting differing but potentially accurate interpretations of the dimensions.

6.2 Prediction Variability

In order to address **SRQ3**, Table 4 shows how LLMs exhibit low prediction variability in the pairwise comparison annotation task. The worst-case scenario is Mistral-22B with few-shot prompting, where 7.75% of cases resulted in all three possible labels ("`A`", "`B`", and "`Tie`") across runs. Mistral-22B also achieved the lowest Krippendorff's α scores overall (0.16, 0.16, and 0.11 for zero-shot, few-shot, and chain-of-thought, respectively). This is followed by Mixtral-8x22B, which obtained Krippendorff's α scores of 0.63, 0.39, and 0.50 under zero-shot, few-shot, and chain-of-thought, respectively. Among the less stable models, Command-R-104B achieved α values of 0.67 and 0.82 with zero-shot and chain-of-thought, respectively, which improved to 0.86 with few-shot. All remaining models and prompting strategies exhibited strong agreement, achieving Krippendorff's α values above 0.78.

Although small models do not achieve the highest performance relative to humans, they are generally the most stable in the pairwise comparison annotation task. They consistently attain perfect or almost perfect, Krippendorff's α scores, while producing few or no "2 equal" and "All unequal" cases. Large models show slightly lower stability. Still, benefiting from few-shot prompting, resulting in fewer "All unequal", "2 equal" cases, or better inter-annotator agreement. In contrast, few-shot prompting appears detrimental for medium-sized models, which consistently achieve lower Krippendorff's α scores than their zero-shot and chain-of-thought counterparts.

7 Conclusion & Future Work

This paper investigated the potential of 12 open-weight LLMs to evaluate argument quality across the logical, rhetorical, and dialectical dimensions using zero-shot, few-shot, and chain-of-thought. Through pairwise comparisons and Bradley–Terry modeling, we analyzed both the alignment of LLM-generated rankings with human judgments and their consistency across runs.

We show that, although LLMs do not yet reach human-level performance, they can approximate BT rankings derived from human pairwise comparisons to a meaningful degree. Among all evaluated models, Llama-70B with few-shot prompting achieved the strongest results, obtaining the highest (but moderate) agreement with human annotations, with Pearson, Spearman, and Kendall correlations of 0.473, 0.477, and 0.327, respectively, as well as a Cohen's κ of 0.493 and 52.93 percent agreement, across the three quality dimensions. Some models align more closely with Llama-70B, such as Qwen-2.5-72B, or smaller models like Qwen-2.5-7B and Olmo2-32B, while others are less consistent with it but still agree with humans. This suggests that different models may capture complementary perspectives, and that combining diverse LLMs may outperform relying on Llama-70B alone.

Finally, we show that, across only three runs per model, prediction variability remains low. Disagreements ranged from cases where two out of three labels matched to instances where all labels differed (up to 7.75% of the total comparisons for the worst configuration Mistral-22B). While still low, this variability motivates the use of aggregation strategies such as majority voting. Few-shot prompting was also found to enhance stability, although its benefits were primarily observed in larger models.

Future work will explore fine-tuning (Stahl et al., 2025) and ensembling LLMs for argument quality assessment, including dimension-specific models. Domain-specific training and ensembling may improve their ability to evaluate reasoning validity and detect fallacies. Second, we will broaden the scope of argument quality assessment by incorporating additional dimensions and working toward a comprehensive overall quality measure.

8 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. (2000) guidelines, namely internal, external, construct and conclusion validity.

[5]The full results table across models, prompts, and dimensions can be found in the repository of the paper.

8.1 Internal Validity

One limitation of this study is the inherent variability of LLM-based evaluations, as different runs may produce different pairwise judgments. While the results remained stable across runs and majority voting was used to improve reliability, stability itself remains an important factor to assess. In particular, we did not evaluate the models at lower temperature settings. Based on the hypothesis proposed by Renze (2024), which identifies a plateau beyond which model outputs may become unreliable, stability at a temperature of 1 may indicate comparable or greater stability at lower temperatures. Nevertheless, an ablation study across different temperature settings would be necessary to verify this assumption and confirm that all models produce comparable results for our task in line with the findings of Renze (2024).

Another limitation concerns prompt formulation. Although the same prompt templates were used for all models, differences in how individual LLMs interpret the instructions may have influenced the resulting judgments.

8.2 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.

Second, the 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 humans might not hold if used in another set.

Third, zero-shot, few-shot, and chain-of-thought approaches were used where the model is given instructions to compare the arguments without fine-tuning or additional training. While these approaches are practical, it does not take into account the performance gains from fine-tuning the LLM. The results do not explain how well the models would do in that particular setting.

Fourth, while the Webis-ArgQuality-20 corpus has been used to evaluate LLMs in recent studies of computational argumentation, it is difficult to reliably assess the extent of potential data contamination as the training data of most modern LLMs is not publicly disclosed, being a current issue in LLM-based studies. Among the 12 models evaluated, only the OLMo family provides full transparency regarding its training data. The remaining open-weight models (Llama, Qwen, Command-R, and Mistral) do not disclose their training datasets or sources (except Llama that reports using public online data pre-December-2023).

Finally, LLMs are evolving rapidly. The LLM versions used in this study are not the latest and may have already been surpassed by newer releases. At the same time, to ensure a fair comparison, we accounted for model size (small, medium, and large) and family, while avoiding substantial version differences (e.g., comparing a version 3 model with a version 1 model from the same family). Although comparative and qualitative insights might remain relevant, metric values may shift even further with advances in LLM.

8.3 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. Comparing them to human judgment without 100% certainty that they have the same understanding of the 'quality' dimensions is very challenging. Furthermore, 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 human crowdsourcers who made pairwise comparisons. Any noise and/or bias could affect the results. The Webis-ArgQuality 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 annotators 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. 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.

8.4 Conclusion Validity

One limitation related to conclusion validity is seen through the number of pairwise comparisons generated for this study. Although 13952 comparisons across 1271 arguments provided a substantial dataset, prior work by Gienapp et al. (2020) demonstrated the benefits of using far larger comparisons sets (over 41000). Larger datasets may increase the statistical robustness of the Bradley-Terry model, and improve the stability of resulting rankings. In this study, the comparison set was chosen to maintain consistency with the comparison corpus released as part of the Webis-ArgQuality-20, which relies on a cyclic group sampling strategy. Future research could explore the impact of scaling up the number of comparisons.

9 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, this work intends 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 harmful 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 involving other information quality aspects. 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 LLM-in-the-loop systems.

Acknowledgments

This work used the Dutch national e-infrastructure with the support of the SURF Coop-erative using grant no. EINF-15500. The authors would also like to thank Greta Damo for her helpful comments during the revision of this paper.

References

J. Anthony Blair. 2012. *Groundwork in the Theory of Argumentation*, volume 21 of *Argumentation Library*. Springer Netherlands, Dordrecht.

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, Valencia, Spain. Association for Computational Linguistics.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 193–202, New York, NY, USA. Association for Computing Machinery.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, and 207 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Iker De la Iglesia, Iakes Goenaga, Johanna Ramirez-Romero, Jose Maria Villa-Gonzalez, Josu Goikoetxea, and Ander Barrena. 2025. Ranking over scoring: Towards reliable and robust automated evaluation of LLM-generated medical explanatory arguments. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9456–9471, Abu Dhabi, UAE. Association for Computational Linguistics.

Darshan Deshpande, Zhivar Sourati, Filip Ilievski, and Fred Morstatter. 2024. Contextualizing argument quality assessment with relevant knowledge. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 316–326, Mexico City, Mexico. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mohamed Elaraby, Diane Litman, Xiang Lorraine Li, and Ahmed Magooda. 2024. Persuasiveness of generated free-text rationales in subjective decisions: A case study on pairwise argument ranking. In *Findings of the Association for Computational Linguistics:*

EMNLP 2024, pages 14311–14329, Miami, Florida, USA. Association for Computational Linguistics.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

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. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on llm-as-a-judge. *The Innovation*, page 101253.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bojun Jin, Jianzhu Bao, Yufang Hou, Yang Sun, Yice Zhang, Huajie Wang, Bin Liang, and Ruifeng Xu. 2025. A multi-persona framework for argument quality assessment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12148–12170, Vienna, Austria. Association for Computational Linguistics.

Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. ArgAnalysis35K : A large-scale dataset for argument quality analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13916–13931, Toronto, Canada. Association for Computational Linguistics.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lujun Li, Lama Sleem, Niccolo' Gentile, Geoffrey Nichil, and Radu State. 2025. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science*, 264:242–251. International Neural Network Society Workshop on Deep Learning Innovations and Applications 2025.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Robust Argumentation Machines*, pages 129–146, Cham. Springer Nature Switzerland.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*.

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument search: Assessing argument relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1117–1120, New York, NY, USA. Association for Computing Machinery.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

P. V. Rao and L. L. Kupper. 1967. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? In *Findings of the*

Association for Computational Linguistics: EMNLP 2024, pages 8826–8837, Miami, Florida, USA. Association for Computational Linguistics.

Julia Romberg, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa. 2025. Towards a perspectivist turn in argument quality assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7458–7485, Albuquerque, New Mexico. Association for Computational Linguistics.

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

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, Vienna, Austria. Association for Computational Linguistics.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

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. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

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, Vancouver, Canada. Association for Computational Linguistics.

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, Valencia, Spain. Association for Computational Linguistics.

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.

Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2000. *Experimentation in Software Engineering: An Introduction*. The Kluwer International Series In Software Engineering. Springer, Germany.

A Specific Version of LLMs Evaluated

We use Mistral (Jiang et al., 2023), Llama 3* (Dubey et al., 2024), Qwen2.5 (Qwen et al., 2025), Olmo2 (OLMo et al., 2025), and Command-R (Cohere et al., 2025) open-access Instruct-based models from Hugging Face [6]. All models are publicly available and traceable through their repository identifiers:

Small Size Models:

- Mistral-7B: `mistralai/Mistral-7B-Instruct-v0.3`
- Llama-8B: `meta-llama/Llama-3.1-8B-Instruct`
- Olmo2-7B: `allenai/OLMo-2-1124-7B-Instruct`
- Qwen2.5-7B: `Qwen/Qwen2.5-7B-Instruct`
- Command-r-7B: `CohereLabs/c4ai-command-r7b-12-2024`

Medium Size Models:

- Mixtral-8x7B: `mistralai/Mixtral-8x7B-Instruct-v0.1`
- Mistral-22B: `mistralai/Mistral-Small-Instruct-2409`
- Olmo2-32B: `allenai/OLMo-2-0325-32B-Instruct`
- Mixtral-8x22B: `mistralai/Mixtral-8x22B-Instruct-v0.1`

Large Size Models:

- Llama-70B: `meta-llama/Llama-3.3-70B-Instruct`
- Qwen2.5-72B: `Qwen/Qwen2.5-72B-Instruct`
- Command-r-104B: `CohereLabs/c4ai-command-r-plus-08-2024`

Small models were run on a single H100 GPU, medium-sized models on up to two H100 GPUs, and large models on up to four H100 GPUs.

[6] https://huggingface.co/models

B Prompts

We tested three prompting strategies, zero-shot, few-shot, and chain-of-thought using a system-user prompt setup with structured outputs through the vLLM Python library [7].

B.1 Zero-shot Prompt

System Prompt: `You are given two arguments: Argument A and Argument B. Decide which one is {dim_adverb} stronger: Reply with only one of the following options: A, B, or Tie.`

User Prompt:

```
Argument A: {a}
Argument B: {b}
```

Output Labels: `["A", "B", "Tie"]`

Where `dim_adverb` are replaced with the strings `logically`, `dialectically`, `rhetorically`, for the logic, dialectic, and rhetoric dimension, respectively. `a`, and `b` are replaced with the arguments A, and B, respectively, to be judged by the LLM.

B.2 Few-shot Prompt

System Prompt: `You are given two arguments: Argument A and Argument B. Decide which one is {dim_adverb} stronger: Reply with only one of the following options: A, B, or Tie.`

```
Examples:
Argument A: {ex_A_argA}
Argument B: {ex_A_argB}
A

Argument A: {ex_B_argA}
Argument B: {ex_B_argB}
B

Argument A: {ex_tie_argA}
Argument B: {ex_tie_argB}
tie
```

User Prompt:

```
Argument A: {a}
Argument B: {b}
```

Output Labels: `["A", "B", "Tie"]`

Where `dim_adverb`, `a`, and `b` are instantiated as in the zero-shot prompt. `ex_A_argA` and `ex_A_argB` are few-shot examples where `A` is preferred on the evaluated dimension; `ex_B_argA` and `ex_B_argB` are examples where `B` is preferred;

[7] https://docs.vllm.ai/en/latest/

and ex_tie_argA and ex_tie_argB are examples where the preferred label is Tie.

We used 3 shots for each of the 20 topics in the Webis-ArgQuality-20 corpus, leading to 60 examples per dimension [8].

B.3 Chain-of-thought

System Prompt: You are given two arguments: Argument A and Argument B. Decide which one is {dim_adverb} stronger considering these criteria:
{dim_criterias}
Let's think step by step and reply with only one of the following options: A, B, or Tie.

User Prompt:
Argument A: {a}
Argument B: {b}

Output Labels: ["A", "B", "Tie"]

Where dim_adverb, a, and b are instantiated as in the zero-shot prompt. dim_criterias is replaced with the chain-of-thought questions for each quality dimension based on the theoretical work of Wachsmuth et al. (2017b):

Criteria for logic:
- which is more acceptable? A premise of an argument is acceptable if it is rationally worthy of being believed to be true.
- which is more relevant to a conclusion? A premise of an argument is relevant if it contributes to the acceptance or rejection of the argument's conclusion.
- which is more sufficient to justify a conclusion? An argument's premises are sufficient if, together, they give enough support to make it rational to draw its conclusion.

Criteria for dialectic:
- which would be acceptable to the audience? Argumentation is acceptable if the target audience accepts both the consideration of the stated arguments for the issue and the way they are stated.
- which contributes more to constructive dialogue? Argumentation is relevant if it contributes to the issue's resolution, i.e., if it states arguments or other information that help to arrive at an ultimate conclusion.
- which better anticipates or refutes counterarguments? Argumentation is sufficient if it adequately rebuts those counterarguments to it that can be anticipated.

Criteria for rhetoric:
- which appears more authorative/trust worthy? Argumentation creates credibility if it conveys arguments and similar in a way that makes the author worthy of credence.
- which makes a stronger emotional appeal? Argumentation makes a successful emotional appeal if it creates emotions in a way that makes the target audience more open to the author's arguments.
- which has a clearer style? Argumentation has a clear style if it uses correct and widely unambiguous language as well as if it avoids unnecessary complexity and deviation from the issue.
- which has a more appropiate style? Argumentation has an appropriate style if the used language supports the creation of credibility and emotions as well as if it is proportional to the issue.
- Which is better arranged? Argumentation is arranged properly if it presents the issue, the arguments, and its conclusion in the right order.

[8] The full list of selected examples is available at https://github.com/benjaminocampo/arg_quality_with_llms.

C Licenses and Intended Use

The dataset used in our experiments, Webis-ArgQuality-20, was created by Gienapp et al. (2020) and is available under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. It is publicly accessible at https://zenodo.org/records/3780049.

The models employed in our experiments, as listed in Appendix A, were used in accordance with their respective licenses and applicable usage policies. The code used in our experiments, together with the LLM-generated pairwise judgments and the Bradley-Terry rankings derived from them, is released under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. Redistribution of model-generated outputs is believed to be permitted under the applicable licenses of the models used; users remain responsible for complying with the original model licenses and usage policies.

		Logic						Rhetoric						Dialectic					
		Zero		Few		CoT		Zero		Few		CoT		Zero		Few		CoT	
		#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Mistral-7B	A	1265	9.07	605	4.34	458	3.28	4163	29.84	11901	85.30	1274	9.13	2776	19.90	3432	24.60	2075	14.87
	B	1443	10.34	13204	94.64	650	4.66	2057	14.74	1860	13.33	1144	8.20	1445	10.36	1734	12.43	1553	11.13
	Tie	11244	80.59	143	1.02	12844	92.06	7732	55.42	191	1.37	11534	82.67	9731	69.75	8786	62.97	10324	74.00
Llama-8B	A	6246	44.77	9038	64.78	6407	45.92	7401	53.05	10510	75.33	7601	54.48	5891	42.22	11318	81.12	5645	40.46
	B	4663	33.42	4617	33.09	2687	19.26	4868	34.89	3372	24.17	3027	21.70	5470	39.21	2589	18.56	3768	27.01
	Tie	3043	21.81	297	2.13	4858	34.82	1683	12.06	70	0.50	3324	23.82	2591	18.57	45	0.32	4539	32.53
Olmo2-7B	A	512	3.67	711	5.10	1079	7.73	749	5.37	1679	12.03	1532	10.98	354	2.54	761	5.45	1945	13.94
	B	9846	70.57	13065	93.64	9409	67.44	11072	79.36	12028	86.21	11569	82.92	10476	75.09	12993	93.13	9840	70.53
	Tie	3594	25.76	176	1.26	3464	24.83	2131	15.27	245	1.76	851	6.10	3122	22.38	198	1.42	2167	15.53
Qwen2.5-7B	A	7016	50.29	11118	79.69	2902	20.80	6911	49.53	9676	69.35	6839	49.02	7544	54.07	9888	70.87	4586	32.87
	B	4917	35.24	2094	15.01	8101	58.06	5826	41.76	4165	29.85	4568	32.74	5146	36.88	3552	25.46	8496	60.89
	Tie	2019	14.47	740	5.30	2949	21.14	1215	8.71	111	0.80	2545	18.24	1262	9.05	512	3.67	870	6.24
Commandr-7B	A	3569	25.58	2430	17.42	3893	27.90	5300	37.99	4664	33.43	5005	35.87	4532	32.48	5280	37.84	5088	36.47
	B	8868	63.56	8960	64.22	5789	41.49	7434	53.28	6869	49.23	5278	37.83	7987	57.25	6267	44.92	5697	40.83
	Tie	1515	10.86	2562	18.36	4270	30.60	1218	8.73	2419	17.34	3669	26.30	1433	10.27	2405	17.24	3167	22.70
Mixtral-8x7B	A	1681	12.05	8348	59.83	600	4.30	1151	8.25	13239	94.89	393	2.82	1015	7.27	2529	18.13	1089	7.81
	B	7577	54.31	4937	35.39	8618	61.77	10509	75.32	692	4.96	5322	38.15	8221	58.92	6204	44.47	8563	61.37
	Tie	4694	33.64	667	4.78	4734	33.93	2292	16.43	21	0.15	8237	59.04	4716	33.80	5219	37.41	4300	30.82
Mistral-22B	A	319	2.29	3930	28.17	287	2.06	468	3.35	3456	24.77	1085	7.78	306	2.19	4568	32.74	474	3.40
	B	12492	89.54	8299	59.48	12454	89.26	12782	91.61	9939	71.24	10818	77.54	12428	89.08	7969	57.12	12379	88.73
	Tie	1141	8.18	1723	12.35	1211	8.68	702	5.03	557	3.99	2049	14.69	1218	8.73	1415	10.14	1099	7.88
Olmo2-32B	A	5725	41.03	5244	37.59	6709	48.09	6542	46.89	6602	47.32	5961	42.73	5694	40.81	6841	49.03	5983	42.88
	B	5078	36.40	5578	39.98	5209	37.34	4782	34.27	6009	43.07	6316	45.27	3750	26.88	5626	40.32	4913	35.21
	Tie	3149	22.57	3130	22.43	2034	14.58	2628	18.84	1341	9.61	1675	12.01	4508	32.31	1485	10.64	3056	21.90
Mixtral-8x22B	A	2019	14.47	3965	28.42	3768	27.01	3020	21.65	9143	65.53	2547	18.26	2619	18.77	6700	48.02	3266	23.41
	B	4859	34.83	8357	59.90	8029	57.55	7172	51.40	3009	21.57	7448	53.38	7692	55.13	5657	40.55	8429	60.41
	Tie	7074	50.70	1630	11.68	2155	15.45	3760	26.95	1800	12.90	3957	28.36	3641	26.10	1595	11.43	2257	16.18
Llama-70B	A	8434	60.45	9033	64.74	9001	64.51	7628	54.67	7282	52.19	8643	61.95	9171	65.73	8555	61.32	6698	48.01
	B	5175	37.09	4384	31.42	4808	34.46	6187	44.34	6590	47.23	5188	37.18	4390	31.47	5198	37.26	6664	47.76
	Tie	343	2.46	535	3.83	143	1.02	137	0.98	80	0.57	121	0.87	391	2.80	199	1.43	590	4.23
Qwen2.5-72B	A	7678	55.03	7368	52.81	7331	52.54	8250	59.13	8105	58.09	6544	46.90	7865	56.37	8221	58.92	6525	46.77
	B	4609	33.03	3773	27.04	5598	40.12	5124	36.73	4011	28.75	5375	38.52	4646	33.30	4344	31.14	5128	36.75
	Tie	1665	11.93	2811	20.15	1023	7.33	578	4.14	1836	13.16	2033	14.57	1441	10.33	1387	9.94	2299	16.48
Commandr-104B	A	10146	72.72	8455	60.60	11761	84.30	10095	72.36	9768	70.01	10976	78.67	11133	79.80	10760	77.12	10011	71.75
	B	2072	14.85	1594	11.42	1357	9.73	1909	13.68	2547	18.26	2001	14.34	1342	9.62	2176	15.60	1361	9.75
	Tie	1734	12.43	3903	27.97	834	5.98	1948	13.96	1637	11.73	975	6.99	1477	10.59	1016	7.28	2580	18.49

Table A: Distribution of LLM Comparisons across each quality dimension and prompt configuration.

D Distribution of Human and LLM Comparisons

The Webis-ArgQuality-20 corpus contains 13952 comparisons. For the logic dimension, 6222 are predicted as A (44.60%), 5719 as B (40.99%), and 2011 as Tie (14.41%); for rhetoric, 5833 are A (41.81%), 5660 are B (40.57%), and 2459 are Tie (17.62%); and for dialectic, 6005 are A (43.04%), 5767 are B (41.33%), and 2180 are Tie (15.62%). This means that, across the comparisons in the dataset, humans choose A and B in a more or less comparable and balanced quantity of times, while Ties occur less frequently.

Table A reports the LLM comparisons across each quality dimension and prompting method, reporting the majority-vote distribution across the three runs described in Section 5 for each LLM configuration [9]. Note that some models, such as Mistral-7B for small sizes and Mixtral-8x7B and Mixtral-8x22B for medium sizes, predict a higher quantity of Ties under zero-shot and CoT prompting, reaching up to 92.06% for Mistral-7B with CoT prompting. When using few-shot prompting, this tendency is reversed; however, the model tends to favor one argument more strongly. For example, in the logic dimension, B is usually chosen in 94.64% of the cases, while in rhetoric, A is chosen in 85.30% of the cases. Tie is also chosen more often under few-shot prompting, with 62.97% of the cases, although this is lower than the Tie rates observed with zero-shot and CoT prompting, which are 69.75% and 74%, respectively, for these dimensions. Larger models predict fewer Ties across all prompts and dimensions while predicting A more often than B.

[9] Given the amount of configurations (12 LLMs, 3 prompts, and 3 quality dimensions) We report the distribution (`distplots`) of the LLM-derived BT scores in the repository of the paper.