

Calibrating LLM-as-a-Judge for Argument Quality Assessment through Prompting and Scoring Scale Design

Guilherme Aren Marchetti, Gil Rocha, Henrique Lopes Cardoso

LIACC, Faculty of Engineering, University of Porto, Portugal

Correspondence: up202408600@up.pt

Abstract

Although LLM-as-a-Judge approaches have been successful in several domains, recent studies show that LLMs still struggle with Computational Argumentation tasks, particularly Argument Quality Assessment. Motivated by recent results in Information Retrieval showing that scoring granularity can affect the effectiveness of LLM-based document rankers, this paper investigates whether the same effect holds in Argument Quality Assessment scenarios. Our results demonstrate that, in several cases, more granular scoring ranges yield substantially higher agreement after mapping predictions back to the original scale. This contrasts with how humans evaluate, where reducing the granularity of possible scores can foster clearer, more consistent decisions. Additionally, our error analysis reveals a recurring limitation: LLMs tend to underestimate arguments labeled by human experts as high quality, frequently assigning them medium scores instead. This suggests that, despite the gains obtained through prompt and scale design, current LLMs remain poorly calibrated at the upper end of the argument quality scale.

1 Introduction

The use of Large Language Models (LLMs) for evaluative tasks, often referred to as the LLM-as-a-Judge approach, has become increasingly common (Gu et al., 2026). This class of approaches has shown promising results in several domains (Bavaresco et al., 2025), but recent studies suggest that LLMs still face important limitations when applied to Computational Argumentation tasks (Ajjour et al., 2026). Argument Quality Assessment is one such task. This task remains challenging because argument quality is inherently multidimensional and partially subjective. Human annotators may consider several factors when judging an argument, such as whether its premises are acceptable, relevant to the claim, and sufficient to

support it. This complexity makes the task difficult not only for computational models but also for constructing reliable annotated datasets.

LLM-as-a-Judge for Argument Quality Evaluation

The judge compares an input argument against reference examples, then assigns a score from 1 to 3.

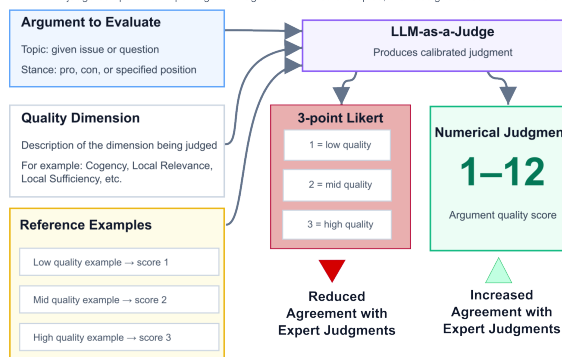


Figure 1: Illustration of the LLM-as-a-Judge process for Argument Quality Evaluation, using Few-Shot prompting and a 3-point Likert scoring.

Data availability is another major limitation for Argument Quality Assessment. Although several datasets exist for coarse-grained evaluation (Ivanova et al., 2024), fine-grained resources remain comparatively scarce. This is largely because annotating multiple dimensions of argument quality is labor-intensive and requires expert knowledge. As a result, datasets with detailed quality annotations usually contain a relatively small number of arguments. A relevant case is the Dagstuhl-15512 ArgQuality dataset (Wachsmuth et al., 2017b), one of the most complete resources in terms of the number of quality dimensions annotated. The dataset follows the taxonomy proposed by Wachsmuth et al. (2017a) and contains expert annotations for fifteen different aspects of argument quality. Each argument was evaluated by three expert annotators, resulting in high-quality annotations. However, the dataset contains only about 300 arguments, limiting the applicability of data-intensive methods such as supervised fine-tuning.

One way to address this limitation is through in-

context learning, which has been shown to improve LLM performance across a variety of tasks (Brown et al., 2020; Agarwal et al., 2024), while requiring a small amount of annotated data. A second strategy for improving LLM-as-a-Judge performance concerns the format of the evaluation scale. Recent work in Information Retrieval has shown that changing the scoring methodology can affect the effectiveness of LLM-based rankers and evaluators (Zhuang et al., 2024; Godfrey et al., 2025). These findings suggest that the way a model is asked to express its judgment may influence the quality and consistency of the resulting evaluation.

Motivated by these two lines of work, this paper investigates whether few-shot prompting and alternative scoring scales can improve LLM-based Argument Quality Assessment. Using the taxonomy proposed by Wachsmuth et al. (2017a), we evaluate four LLMs from different model families and parameter scales on five argument quality dimensions: Overall Quality, Cogency, Local Acceptability, Local Relevance, and Local Sufficiency. Although we refer to these as five dimensions, they form a hierarchy: Overall Quality comprises Cogency, Effectiveness, and Reasonableness (with the last two not included in this study), and Cogency is further decomposed into Local Acceptability, Local Relevance, and Local Sufficiency.

Our experiments, leveraging the Dagstuhl-15512 ArgQuality dataset (Wachsmuth et al., 2017b), show that prompt design has a substantial impact on model agreement with human annotations. In particular, including examples and expanding the scoring scale can improve agreement across several models and quality dimensions. This contrasts with how human annotators behaved during the creation of the dataset. Wachsmuth et al. (2017b) states that a 3-point scale was employed since annotators rarely chose the extreme values when a 4-point scale was employed in a validation study. This may indicate a clear difference in how LLMs and Humans perform evaluations: while humans tend to struggle to differentiate between too granular scores, LLMs perform better with a wide range of scores to choose from.

2 Related Work

Argument Quality Assessment has been studied as a central task in Computational Argumentation. One particularly challenging aspect of this task is that argument quality is not a single property.

It depends on several underlying criteria, including the plausibility of the premises, their relevance to the claim, among others. One influential work that proposes a framework for capturing this multidimensional nature is the taxonomy proposed by Wachsmuth et al. (2017b). This taxonomy comprises 15 dimensions, ranging from specific aspects of the argument, such as the Local Acceptability of the premises, to more general ones, such as the Overall Quality of the argument.

This taxonomy is the basis for the Dagstuhl-15512 ArgQuality dataset (Wachsmuth et al., 2017a). The dataset was constructed by having three expert annotators evaluate 304 arguments collected from debate portals. These arguments cover a broad range of topics, including historical figures, Atheism vs. Christianity, and the validity of banning plastic water bottles. Each argument was independently scored across all 15 dimensions by the three annotators, making the dataset one of the most detailed resources available for fine-grained Argument Quality Assessment. At the same time, its relatively small size reflects one of the field’s central challenges: obtaining expert annotations across multiple quality dimensions is expensive and labor-intensive.

Due to the strong text-processing and instruction-following capabilities of Large Language Models (LLMs), their use as evaluators has become increasingly common. This approach is often referred to as LLM-as-a-Judge (Gu et al., 2026; Li et al., 2025). In this setting, LLMs can be used to evaluate different properties of a text or response, producing outputs in several formats, such as point-wise scores, pairwise preferences, or ranked lists. For argument evaluation, LLMs are an attractive alternative to fully supervised approaches, especially when annotated data is scarce. However, the reliability of LLM judgments remains an open question, particularly for subjective and multidimensional tasks such as Argument Quality Assessment.

A study directly exploring the effectiveness of LLMs as argument evaluators was conducted by Mirzakhmedova et al. (2024). The authors evaluated two proprietary LLMs for argument quality assessment using the taxonomy proposed by Wachsmuth et al. (2017a), exclusively via zero-shot prompting, and with the same scoring methodology (3-point Likert scale). Their results show that LLMs can reach moderate agreement with human experts for some fine-grained dimensions. However, performance varies substantially across

dimensions, and the models struggle more with broader or more abstract judgments. These findings suggest that LLMs may be better suited to evaluating narrowly defined aspects of argument quality than to producing global quality judgments.

Another study that highlights the challenging nature of Argument Quality Assessment was conducted by Ajjour et al. (2026). The authors benchmarked different LLMs of varying families and sizes across multiple Computational Argumentation tasks, including Argument Mining (Stab et al., 2018), Perspective Assessment (Bar-Haim et al., 2017), Quality Assessment (Wachsmuth et al., 2017b), Argument Reasoning (Habernal et al., 2018), and Argument Generation (Hua et al., 2019). Their results show that Argument Quality Assessment remains one of the most difficult tasks for current LLMs, with models from different families and sizes achieving comparatively low scores. This also suggests that increasing model size alone may not be sufficient to solve the task.

These findings underscore the need to explore alternative approaches to improving LLM-based argument evaluation without relying on additional training data. Prompt engineering and in-context learning allow models to be guided through task descriptions and examples rather than fine-tuning. At the same time, recent work on LLM-based ranking and evaluation has shown that the scoring scale format can influence model behavior (Zhuang et al., 2024; Godfrey et al., 2025). Together, these lines of work suggest that both the examples included in a prompt and the way the model is asked to express its judgment may affect the reliability of LLM-as-a-Judge approaches.

Building on this prior work, our study investigates how few-shot prompting and alternative scoring scales affect LLM agreement with expert annotations in fine-grained Argument Quality Assessment. Unlike previous work that primarily evaluates LLMs under a fixed prompting strategy, we focus on how prompt design choices influence agreement across different models, model families, and quality dimensions.

3 Methodology

The prompt variants used in this study build on the design proposed by Mirzakhmedova et al. (2024). In that work, each prompt included the argument to be evaluated, a description of the target quality dimension, and the corresponding annotation ques-

tion originally presented to expert annotators in Wachsmuth et al. (2017b). For example, Figure 2 shows the description and annotation question for the Cogency dimension. The descriptions for the remaining four dimensions used in our study are provided in Appendix A.

Description:

An argument should be seen as cogent if it has individually acceptable premises that are relevant to the argument's conclusion and that are sufficient to draw the conclusion. Try to adequately weight your judgments about local acceptability, local relevance, and local sufficiency when judging about cogency - unless there is a particular dimension among these that dominates your view of an argument.

Accordingly, if you identify more than one argument, try to adequately weight the cogency of each argument when judging about their "aggregate" cogency - unless there is a particular argument that dominates your view of the author's argumentation.

Annotator Question:

How would you rate the success of the author's argumentation in creating credibility?

Figure 2: The original Cogency description and question presented to the experts annotators involved in the Dagstuhl-15512 ArgQuality dataset (Wachsmuth et al., 2017a) study.

Starting from this baseline template, we investigated the effect of two main prompt modifications. The first modification introduced in-context examples of argument evaluation through few-shot prompting. For each quality dimension, we selected three examples: one argument with a low score, one with a medium score, and one with a high score, as illustrated in Figure 1. To promote consistency across the examples, we selected only arguments for which the three expert annotators assigned the same quality rating.

The examples were selected independently for each quality dimension. This is important because scores across dimensions are not perfectly correlated: an argument rated as high in Cogency, for instance, is not necessarily rated as high in Local Acceptability (one of its sub-dimensions), and the converse may also hold. Therefore, using dimension-specific examples helps the prompt better reflect the criteria for each quality dimension.

The second modification concerned the scoring scale. The original dataset uses a three-point Likert scale, with the labels *Low*, *Medium*, and *High*. Motivated by the findings of Zhuang et al. (2024), we

explored whether alternative scoring formats could affect LLM-based argument quality evaluation. In particular, we tested two expanded scales with fixed response options: a five-point and a six-point scale. In addition, we evaluated three open-range scoring variants, in which the model was asked to assign a quality score Q within a predefined interval: $1 \leq Q \leq n$, $n \in \{5, 6, 12\}$.

Because the original annotations use a three-point scale, the outputs produced under the modified scales had to be mapped back to the original labels for evaluation. For the five-point scale and range, either fixed-list or range-based, the two lowest scores were mapped to Low, the middle score to Medium, and the two highest scores to High. For the six-point and twelve-point scales and ranges, the mappings were obtained by grouping consecutive scores into pairs for the six-point scale/range and quartets for the twelve-point scale/range.

Overall, we evaluated seven prompt variants: a zero-shot baseline prompt (as in [Mirzakhmedova et al. \(2024\)](#)), to serve as control for improvements resulting from language models training; one few-shot prompt, which adds dimension-specific examples but maintains the original scale (referred to as 3 Levels); two few-shot prompts using expanded scales with fixed choices, namely the five-point and six-point variants (referred to as 5 and 6 Levels); and three few-shot prompts using open scoring ranges, with upper bounds of 5, 6, and 12 (referred to as Range 1- n). For brevity, this section presents only the prompt variant with open-range scoring between 1 and 12, shown in Figure 3. The remaining prompt templates are included in Appendix B.

4 Experiments

To investigate the consistency and reliability of LLMs in Argument Quality assessment, we evaluated models from different families and with different parameter sizes. This design allows us to examine whether the observed effects are specific to individual models or whether they generalize across architectures.

All experiments were conducted using the HuggingFace Transformers library ([Wolf et al., 2020](#)), with the following models: *google/gemma-4-31B-it*¹, *Qwen/Qwen3.5-27B*², *microsoft/phi-4 (15B)*³,

¹<https://huggingface.co/google/gemma-4-31B-it>

²<https://huggingface.co/Qwen/Qwen3.5-27B>

³<https://huggingface.co/microsoft/phi-4>

```
Considering that
{quality_dimension_description}

{Using a scale from 1 (Low) to 12 (High)}
{quality_dimension_question}

The score should be denoted only using
numbers and enclosed within ** **.

For reference, the following argument {
medium_argument_example} has a score of
**6 (Medium)**, while the argument {
high_argument_example} has a score of
**12 (High)**, and the argument {
low_argument_example} has a score of **1 (
Low)**

Topic: {topic}
Stance : {stance}
Argument: {argument}.
```

Figure 3: Prompt variant with open-range scoring between 1 and 12.

and *meta-llama/Llama-3.3-70B-Instruct*⁴. For generation, all models used their respective default temperature and greedy decoding.

The models were evaluated on the Dagstuhl-15512 ArgQuality dataset ([Wachsmuth et al., 2017b](#)), which contains arguments annotated by human experts according to several dimensions of argument quality. Evaluation was performed on the entire dataset (which lacks a train/test split), excluding only the arguments included as examples in the prompt, resulting in 301 arguments per dimension.

For each prompt configuration described in Section 3, every model was asked to evaluate each argument with respect to one of the following dimensions at a time:

- Overall Quality
- Cogency - An argument is cogent if it has acceptable premises, which are relevant and sufficient to reach the conclusion;
- Local Acceptability - The premises are worthy of being believed;
- Local Relevance - The premises contribute to accepting or rejecting the argument’s conclusion;
- Local Sufficiency - The premises together make it rational to reach the argument’s conclusion.

⁴<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

These dimensions were selected to examine whether LLMs differ in their ability to assess global and local aspects of argument quality. Overall Quality represents a broad judgment of the argument as a whole and may require the model to implicitly integrate several criteria. In contrast, dimensions such as Local Acceptability, Local Relevance, and Local Sufficiency focus on more specific properties of argumentative components and their relations. Cogency occupies an intermediate position, as it is closely related to the logical and structural adequacy of the argument.

Model predictions were compared against the gold-standard annotations provided in the dataset. Since all predictions were converted to the original three-point scale before evaluation, results across prompt configurations remain directly comparable. Agreement was measured using Krippendorff- α against the majority vote of the expert annotators.

5 Results

As described in Section 4, we evaluated each model across the five selected argument quality dimensions and across the prompting strategies introduced in Section 3. Table 1 reports the agreement between the model predictions and the majority-vote expert annotations from the Dagstuhl-15512 ArgQuality dataset. Agreement is measured using Krippendorff’s α , with higher values indicating stronger agreement with the expert labels. All prompt variants are few-shot, with the only exception being the zero-shot prompt.

The table also includes some external reference points. The row labeled Expert Inter-annotator Agreement reports the agreement among the original expert annotators, as reported by Wachsmuth et al. (2017b). The GPT-3 and PaLM 2 rows report the results obtained by Mirzakhmedova et al. (2024) using their expert prompt variant. These values provide external baselines against which the models and prompt variants evaluated in this study can be compared. The zero-shot row was included as a control for improvements driven by more recent language models’ reasoning/language understanding.

Overall, the results show that several of the tested models reach agreement similar to that observed among expert annotators, especially for the more specific quality dimensions. The strongest average results are obtained by Gemma 4 31B, which remains stable across most prompting variants, with

average agreement scores ranging from 0.44 to 0.49. Qwen 3.5 27B and Phi 4 15B follow, with average scores generally in the range of 0.35 to 0.45. Llama 3.3 70B achieves the lowest agreement across most prompt variants, though its performance improves substantially when using the 1-12 scoring range.

The first interesting result relates to the effect of including examples in the prompt. Few-shot leads to stronger agreement than the GPT-3 and PaLM 2 expert-prompt baselines reported by Mirzakhmedova et al. (2024) and the zero-shot variant, particularly for dimensions where those baselines showed weak agreement. This suggests that providing concrete examples of expert-rated arguments can help align LLM judgments with the annotation scheme. This improvement is especially relevant because the models evaluated in this study are smaller than the proprietary models used in the previous baselines, yet they often achieve higher agreement.

At the same time, the results do not suggest a simple relationship between model size and agreement with expert annotations. For example, Phi 4 15B outperforms Llama 3.3 70B under most prompting variants, despite being substantially smaller. Similarly, while Gemma 4 31B achieves the best overall results, Qwen 3.5 27B and Phi 4 15B remain competitive across several dimensions. These findings indicate that model family, training data, instruction-following behavior, and sensitivity to prompt format may be more important than parameter count alone. This observation aligns with Ajjour et al. (2026), who also report limited or inconsistent gains from larger models in argument quality assessment tasks.

The choice of scoring scale also clearly affects agreement. When aggregating by prompt variant, the 1-12 range variant has the highest mean agreement, with an average score of 0.43 (calculated with all four models). This is higher than the averages obtained by the other prompt variants, which range from approximately 0.36 to 0.38. The improvement is especially noticeable for Llama 3.3 70B: its average agreement increases from 0.22 across the other five prompt variants to 0.42 with the 1-12 range. This represents the largest prompt-related gain observed in the table. Qwen 3.5 27B and Phi 4 15B also obtain their best average results with the 1-12 range. Gemma 4 31B is the main exception: it performs consistently well across most configurations, but its average agreement decreases slightly with the 1-12 range.

Table 1: Agreement between LLM judgments and majority-vote expert annotations, measured using Krippendorff’s α and rounded to two decimal places. The Expert Inter-annotator Agreement row reports the inter-annotator agreement from Wachsmuth et al. (2017b). GPT-3 and PaLM 2 results are taken from the expert-prompt variant reported by Mirzakhmedova et al. (2024). Bold denotes the best score for a given dimension or average, while underline denotes the second best. The last row contains the average agreement obtained per prompt variant across all models, with the best agreement highlighted as before. All prompt variants, with the exception of the zero-shot rows, are few-shot.

	Prompting	Average	Overall Quality	Cogency	Local Acceptability	Local Relevance	Local Sufficiency
Expert Inter-annotator Agreement (Humans)		0.46	0.51	0.44	0.46	0.47	0.44
GPT-3 (Zero-shot) (Mirzakhmedova et al., 2024)		0.11	0.12	-0.12	0.49	0.40	-0.33
PaLM 2 (Zero-shot) (Mirzakhmedova et al., 2024)		0.29	0.19	0.02	0.56	0.47	0.23
Gemma 4 31B	Zero-shot	0.30	0.10	0.11	0.46	0.42	0.43
	3 Levels	0.49	<u>0.47</u>	0.47	0.52	0.56	0.43
	5 Levels	<u>0.47</u>	0.36	0.52	0.48	0.51	<u>0.46</u>
	6 Levels	0.49	0.46	<u>0.51</u>	0.56	0.51	0.40
	Range (1-5)	0.49	0.35	0.52	0.54	<u>0.54</u>	0.47
	Range (1-6)	0.49	0.44	0.50	<u>0.55</u>	<u>0.54</u>	0.41
	Range (1-12)	0.44	0.34	0.49	0.47	0.46	<u>0.46</u>
Qwen 3.5 27B	Zero-shot	0.39	0.37	0.25	0.40	0.49	0.42
	3 Levels	0.39	0.39	0.36	0.38	0.47	0.38
	5 Levels	0.35	0.33	0.40	0.33	0.41	0.31
	6 Levels	0.41	0.35	0.43	0.42	0.47	0.38
	Range (1-5)	0.43	0.40	0.41	0.40	0.46	0.47
	Range (1-6)	0.39	0.33	0.37	0.44	0.48	0.34
	Range (1-12)	0.45	0.34	0.46	0.50	0.51	0.41
Phi 4 15B	Zero-shot	0.36	0.34	0.38	0.39	0.36	0.32
	3 Levels	0.38	0.28	0.38	0.48	0.46	0.29
	5 Levels	0.37	0.35	0.40	0.46	0.37	0.29
	6 Levels	0.36	0.22	0.36	0.48	0.45	0.29
	Range (1-5)	0.37	0.30	0.39	0.42	0.38	0.34
	Range (1-6)	0.38	0.24	0.41	0.48	0.43	0.36
	Range (1-12)	0.43	0.40	0.42	0.50	0.48	0.34
Llama 3.3 70B	Zero-shot	0.20	0.22	0.11	0.27	0.29	0.07
	3 Levels	0.26	0.17	0.18	0.41	0.42	0.13
	5 Levels	0.23	0.22	0.11	0.38	0.28	0.14
	6 Levels	0.16	0.12	-0.03	0.34	0.33	0.04
	Range (1-5)	0.13	0.18	-0.02	0.25	0.18	0.08
	Range (1-6)	0.21	0.24	0.04	0.36	0.29	0.13
	Range (1-12)	0.41	0.35	0.38	0.53	0.41	0.38
Prompt Variant Aggregated Average	Zero-shot	0.31	0.26	0.21	0.38	0.39	0.31
	3 Levels	<u>0.38</u>	0.33	0.35	0.45	0.48	0.31
	5 Levels	0.36	0.31	<u>0.36</u>	0.41	0.39	0.30
	6 Levels	0.35	0.29	0.32	0.45	0.44	0.28
	Range (1-5)	0.35	0.31	0.32	0.40	0.39	<u>0.34</u>
	Range (1-6)	0.37	0.31	0.33	<u>0.46</u>	0.43	0.31
	Range (1-12)	0.43	0.36	0.44	0.50	<u>0.46</u>	0.40

These results suggest that a more fine-grained scoring range can help some models produce judgments that are better aligned with expert annotations after conversion back to the original three-point scale. We explore a possible explanation for this effect in Section 5.1.

Taken together, the results indicate that prompt design has a substantial effect on LLM-based argument quality assessment. Few-shot prompting improves agreement relative to previous expert-prompt baselines, and expanded scoring ranges, particularly the 1-12 range, can further improve agreement for several models. However, the gains are not uniform across model families and quality dimensions.

5.1 Error Analysis

To better understand the factors limiting model performance, we conducted a complementary er-

ror analysis. While the previous section evaluated agreement between LLM predictions and expert annotations using Krippendorff’s alpha, here we reformulate the task as a three-class classification problem. Each model prediction is compared with the expected score from the majority vote of the expert annotators, and the resulting errors are represented in confusion matrices.

For this analysis, all predictions were first mapped back to the original three-point scale whenever necessary, as described in Section 3. The confusion matrices were normalized by the gold reference class. This allows interpreting each row as the distribution of model predictions for arguments belonging to a given gold class.

Due to space limitations, this section reports only the confusion matrices for Gemma 4 31B with 3 Levels, shown in Figure 4. This configuration was selected because it achieved one of the strongest

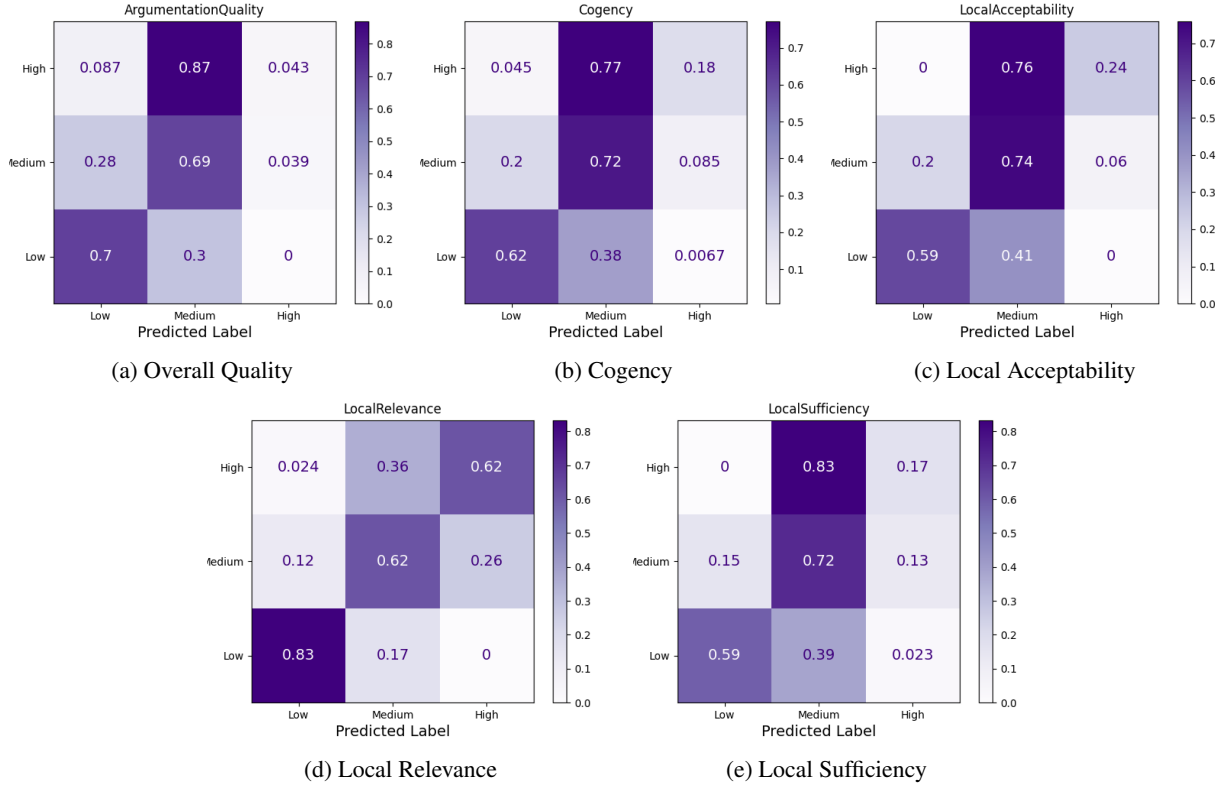


Figure 4: Confusion Matrices for the Gemma 4 31B model while analyzing the different quality dimensions with the 3 Levels prompt variant.

overall performances among the evaluated settings, making its remaining errors particularly informative. The matrices for the other models are provided in Appendix C.

The confusion matrices reveal a consistent pattern, with the model’s errors concentrated in cases where human-annotated arguments labeled High are predicted as Medium. This suggests that, while the model identifies low- and medium-quality arguments with reasonable consistency, it assigns the highest quality label less frequently.

This pattern is especially visible in the Overall Quality dimension. For Low and Medium arguments, the model shows comparatively stable class-wise recall, around 70%. However, for arguments with a gold label of High, recall drops to only 4%. In this class, 87% of the arguments are instead classified as Medium. Thus, the main source of disagreement in Overall Quality is not that the model confuses high-quality arguments with low-quality ones, but rather that it systematically underestimates them by one level.

The same tendency appears, although less severely, in the other quality dimensions. Local Relevance, which is the dimension with the strongest agreement for this model, shows a more

balanced distribution across the three classes. In this case, class-wise recall ranges from 62% to 83%, indicating that the model is better calibrated across Low, Medium, and High instances. The same pattern is observed across the remaining models, as shown in Appendix C.

These results suggest that the model is conservative in assigning quality scores. Rather than overestimating weak arguments, it tends to avoid assigning the highest score, even when expert annotators agree that the argument is high-quality. One possible interpretation is that the model applies a stricter threshold for the High label than the human annotators. In practice, this could mean the model penalizes minor weaknesses more heavily, downgrading otherwise strong arguments to Medium.

This conservative bias helps explain the agreement results reported in Table 1. Increasing the scoring range allows the models to penalize minor weaknesses in an argument while still keeping it within the High-quality range. This effect can be observed by comparing the change in recall between using the 3 Levels and the 1-12 range prompting variants for Local Acceptability evaluation, with the Qwen 3.5 27B model (Figure 5), which displays a more pronounced impact of the

scoring range while maintaining a comparable performance as the Gemma 4 31B model. This comparison shows that increasing the scoring range during assessment shifts some previously evaluated arguments from Medium to High, with no significant decrease in recall for the other labels. The unconverted scores for the 1-12 range prompt variant are included in Appendix D.

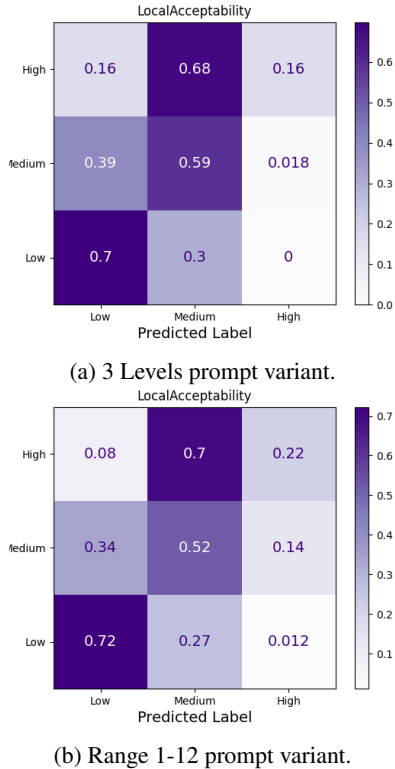


Figure 5: Confusion matrices for different prompting variants, while assessing the Local Acceptability dimension with the Qwen 3.5 27B model.

The error analysis, therefore, indicates that the main limitation is not simply random misclassification. Instead, the errors are structured and ordinal: when the model disagrees with the gold label, it usually predicts an adjacent class rather than the opposite extreme. The increased number of possible labels gives models more room to penalize or reward different aspects of arguments, while still aligning with the coarser scores used by human annotators.

6 Conclusion

Despite recent improvements in reasoning capabilities, LLMs still struggle to generate argument evaluations aligned with human expert annotators. In this paper, we investigated whether prompt engineering strategies can improve the agreement be-

tween LLM judgments and human expert annotations. Specifically, we evaluated the effect of few-shot prompting and alternative scoring scales across multiple models and five argument quality dimensions.

Our results show that carefully selected in-context examples can substantially improve agreement with expert annotations. By using examples drawn from unanimously annotated arguments, few-shot prompting provided the models with concrete instances of how each quality dimension should be evaluated. This led to stronger results than previously reported expert-prompt baselines, suggesting that example-based prompting is an effective strategy for aligning LLM judgments with human annotation criteria.

In addition, the scoring format had a substantial impact on model performance. Expanding the evaluation scale, especially through the 1–12 range variant, improved agreement for several models and, in some cases, nearly doubled Krippendorff’s α . However, this effect was not uniform across all models, indicating that the usefulness of a more fine-grained scale depends on the interaction between model and prompt design.

This effect may have some interesting implications. While humans seem to struggle when score granularity is high, the opposite is true for LLMs. In the study that built the Dagstuhl-15512 dataset, the authors had to reduce the possible scores from 4 to 3 to improve annotation quality, as annotators rarely chose the extreme values. On the other hand, in our study, the wider range of scores allows the LLMs to include minor weaknesses without overpenalizing the final score. Regardless, even with this increased range, the error analysis reveals a persistent limitation: the models frequently misclassified high-quality arguments as medium ones, suggesting a conservative bias in their evaluations.

Overall, the findings suggest that altering how scores are assigned can improve LLM-based argument quality assessment. While some aspects of argument evaluation improve when LLMs retrace the same steps humans use, such as using the same descriptions of quality dimensions, for instance, it can be beneficial to diverge in others, such as the scoring scale. Future work will focus on further exploring alternative scoring methodologies and their impact on Argument Quality evaluation, as well as exploring other prompt designs that could benefit from diverging from the instructions suited for human evaluation.

Limitations

Although we selected models from different families in this study, it would be beneficial to expand the evaluation to include additional families and models of varying sizes within each family. Similarly, while we have selected a set of dimensions from different levels of the hierarchy, expanding the study to include the complete set of dimensions proposed in the taxonomy could expose interesting correlations not observed in this study.

Additionally, we have not explored the impact of altering the score anchoring (i.e., the numerical score assigned to the low/medium/high labels) on LLMs' argument evaluations, nor have we assessed sensitivity to different sets of examples in the prompt.

Acknowledgments

This work was financially supported by UID/00027/2025 of the Artificial Intelligence and Computer Science Laboratory (LIACC), with DOI <https://doi.org/10.54499/UID/00027/2025>, funded by Fundação para a Ciência e a Tecnologia, I.P. / MECI through national funds.

References

- Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pages 76930–76966, Red Hook, NY, USA. Curran Associates Inc.
- Yamen Ajjour, Carlotta Quensel, Nedim Lipka, and Henning Wachsmuth. 2026. [ArgBench: Benchmarking LLMs on Computational Argumentation Tasks](#). *arXiv preprint*. ArXiv:2604.17366 [cs.CL] version: 1.
- Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. [Stance Classification of Context-Dependent Claims](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in neural information processing systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Charles Godfrey, Ping Nie, Natalia Ostapuk, David Ken, Shang Gao, and Souheil Inati. 2025. [Likert or Not: LLM Absolute Relevance Judgments on Fine-Grained Ordinal Scales](#). *arXiv preprint*. ArXiv:2505.19334 [cs.LG].
- 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*, 7(6).
- Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. [The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.
- Xinyu Hua, Zhe Hu, and Lu Wang. 2019. [Argument Generation with Retrieval, Planning, and Realization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.
- Rositsa V Ivanova, Thomas Huber, and Christina Niklaus. 2024. [Let's discuss! Quality Dimensions and Annotated Datasets for Computational Argument Quality Assessment](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20749–20779, Miami, Florida, USA. Association for Computational Linguistics.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. [From Generation to Judgment: Opportunities and Challenges](#)

of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

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.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. [Cross-topic Argument Mining from Heterogeneous Sources](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [HuggingFace’s Transformers: State-of-the-art Natural Language Processing](#). *arXiv preprint. ArXiv:1910.03771 [cs]*.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. [Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels](#). 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 358–370, Mexico City, Mexico. Association for Computational Linguistics.

A Quality Dimensions Information

These are the quality dimension descriptions and questions originally presented to the annotators involved in the study that originated the Dagstuhl-15512 ArgQuality dataset (Wachsmuth et al.,

2017a), which was included in the prompts used in this study.

A.1 Overall Quality

Description: A definition of overall argumentation quality does not exist so far. Try to judge about the overall quality based on all those of your ratings that you think influence the overall quality of the given argumentation. If there is anything not covered by these ratings that influences your view of the author’s argumentation, also take that into account.

Question: How would you rate the overall quality of the author’s argumentation?

A.2 Cogency

Description: An argument should be seen as cogent if it has individually acceptable premises that are relevant to the argument’s conclusion and that are sufficient to draw the conclusion. Try to adequately weight your judgments about local acceptability, local relevance, and local sufficiency when judging about cogency—unless there is a particular dimension among these that dominates your view of an argument. Accordingly, if you identify more than one argument, try to adequately weight the cogency of each argument when judging about their “aggregate” cogency—unless there is a particular argument that dominates your view of the author’s argumentation.

Question: How would you rate the success of the author’s argumentation in creating credibility?

A.3 Local Acceptability

Description: A premise of an argument should be seen as acceptable if it is worthy of being believed, i.e., if you rationally think it is true or if you see no reason for not believing that it may be true. If you identify more than one premise in the comment, try to adequately weight the acceptability of each premise when judging about their “aggregate” acceptability—unless there are particular premises that dominate your view of the author’s argumentation.

Question: How would you rate the acceptability of the premises of the author’s argument?

A.4 Local Relevance

Description: A premise of an argument should be seen as relevant if it contributes to the acceptance or rejection of the argument’s conclusion, i.e.,

if you think it is worthy of being considered as a reason, evidence, or similar regarding the conclusion. If you identify more than one premise in the comment, try to adequately weight the relevance of each premise when judging about their “aggregate” relevance—unless there are particular premises that dominate your view of the author’s argumentation. You should be open to see a premise as relevant even if it does not match your own stance on the issue.

Question: How would you rate the relevance of the premises of the author’s argument?

A.5 Local Sufficiency

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

Question: How would you rate the sufficiency of the premises of the author’s argument?

B Prompt Templates

```
Considering that
{quality_dimension_description}

{quality_dimension_question}
•3 (High)
•2 (Medium)
•1 (Low)

The score should be denoted only using
numbers and enclosed within ** **.

For reference, the following argument
{medium_argument_example} has a score of **2
(Medium)**, while the argument
{high_argument_example} has a score of **3 (
High)**, and the argument
{low_argument_example} has a score of **1 (
Low)**

Topic: {topic}
Stance : {stance}
Argument: {argument}.
```

Figure 6: Few-shot prompting using the original 3-point Likert scale.

```
Considering that
{quality_dimension_description}

{quality_dimension_question}
•5 (High)
•4 (Medium-High)
•3 (Medium)
•2 (Medium-Low)
•1 (Low)

The score should be denoted only using
numbers and enclosed within ** **.

For reference, the following argument
{medium_argument_example} has a score of **3
(Medium)**, while the argument
{high_argument_example} has a score of **5 (
High)**, and the argument
{low_argument_example} has a score of **1 (
Low)**

Topic: {topic}
Stance : {stance}
Argument: {argument}.
```

Figure 7: The 5 Level Few-shot prompt variant.

```
Considering that
{quality_dimension_description}

{quality_dimension_question}
•6 (High)
•5
•4
•3
•2
•1 (Low)

The score should be denoted only using
numbers and enclosed within ** **.

For reference, the following argument
{medium_argument_example} has a score of
**4**, while the argument
{high_argument_example} has a score of **6 (
High)**, and the argument
{low_argument_example} has a score of **1 (
Low)**

Topic: {topic}
Stance : {stance}
Argument: {argument}.
```

Figure 8: The 6 Level Few-shot prompt variant.

```

Considering that
{quality_dimension_description}

Using a scale from 1 (Low) to 5 (High)
{quality_dimension_question}

The score should be denoted only using
numbers and enclosed within ** **.

For reference, the following argument "{
medium_argument_example}" has a score of **3
(Medium)**, while the argument "{
high_argument_example}" has a score of **5 (
High)**, and the argument "{
low_argument_example}" has a score of **1 (
Low)**

Topic: {topic}
Stance : {stance}
Argument: {argument}.

```

Figure 9: The Range 1-5 few-shot prompt variant.

```

Considering that
{quality_dimension_description}

Using a scale from 1 (Low) to 6 (High)
{quality_dimension_question}

The score should be denoted only using
numbers and enclosed within ** **.

For reference, the following argument
{medium_argument_example} has a score of **4
(Medium)**, while the argument
{high_argument_example} has a score of **6 (
High)**, and the argument
{low_argument_example} has a score of **1 (
Low)**

Topic: {topic}
Stance : {stance}
Argument: {argument}.

```

Figure 10: The Range 1-6 few-shot prompt variant.

```

Considering that
{quality_dimension_description}

Using a scale from 1 (Low) to 12 (High)
{quality_dimension_question}

The score should be denoted only using
numbers and enclosed within ** **.

For reference, the following argument
{medium_argument_example} has a score of **6
(Medium)**, while the argument
{high_argument_example} has a score of **12
(High)**, and the argument
{low_argument_example} has a score of **1 (
Low)**

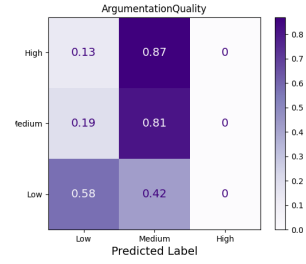
Topic: {topic}
Stance : {stance}
Argument: {argument}.

```

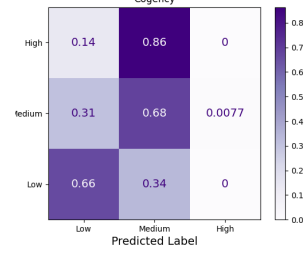
Figure 11: The Range 1-12 few-shot prompt variant.

C Confusion Matrices for different LLMs

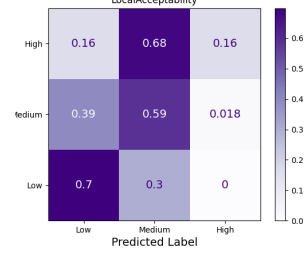
C.1 Few-shot prompting with 3 Levels



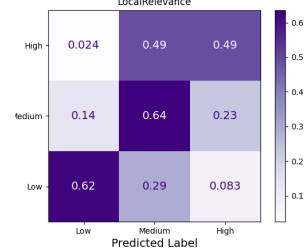
(a) Overall Quality



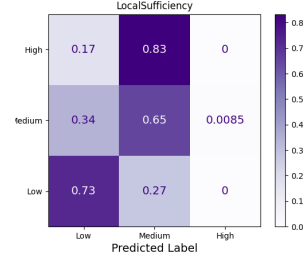
(b) Cogency



(c) Local Acceptability

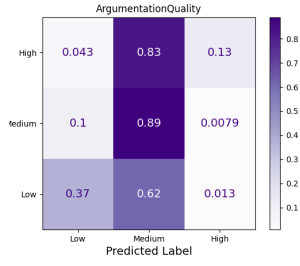


(d) Local Relevance

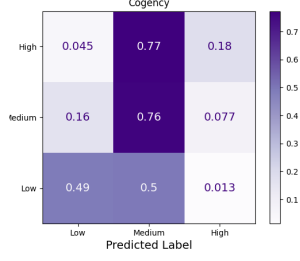


(e) Local Sufficiency

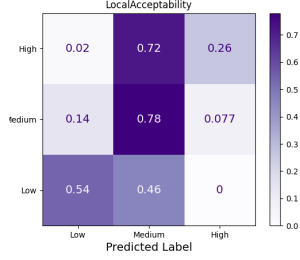
Figure 12: Confusion Matrices for the Qwen 3.5 27B model while analyzing the different quality dimensions with 3 Levels few-shot prompting.



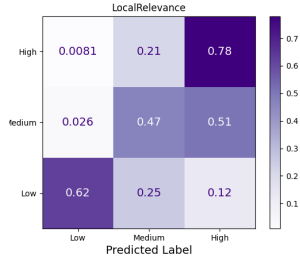
(a) Overall Quality



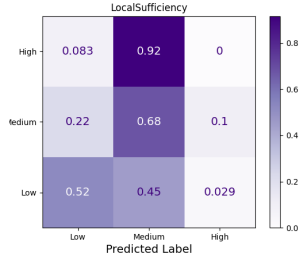
(b) Cogency



(c) Local Acceptability

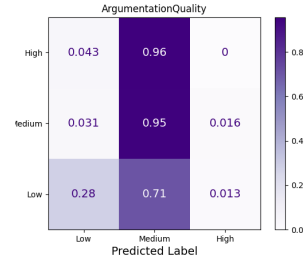


(d) Local Relevance

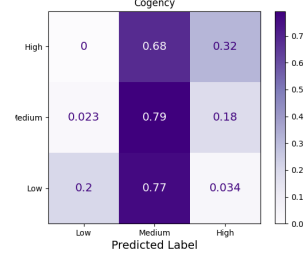


(e) Local Sufficiency

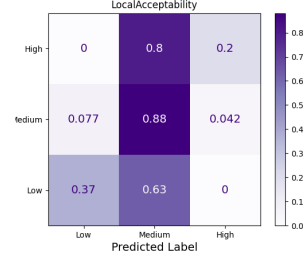
Figure 13: Confusion Matrices for the Phi 4 15B model while analyzing the different quality dimensions with the 3 Levels few-shot prompting.



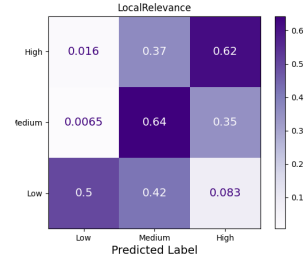
(a) Overall Quality



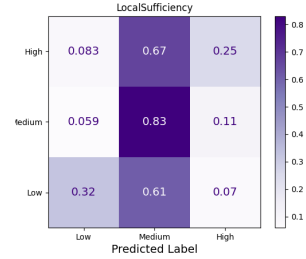
(b) Cogency



(c) Local Acceptability



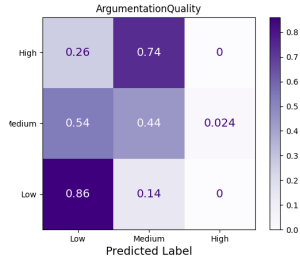
(d) Local Relevance



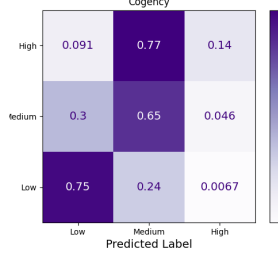
(e) Local Sufficiency

Figure 14: Confusion Matrices for the Llama 3.3 70B model while analyzing the different quality dimensions with the 3 Levels few-shot prompting.

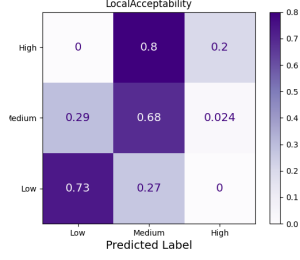
C.2 Few-shot prompting with Range 1-12



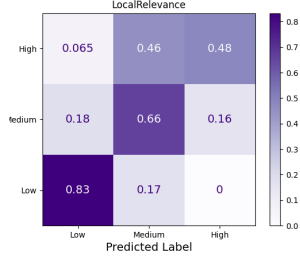
(a) Overall Quality



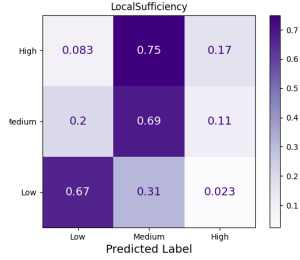
(b) Cogency



(c) Local Acceptability

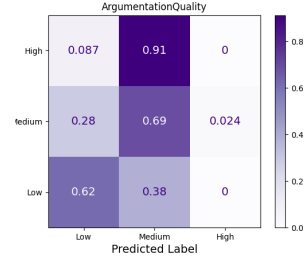


(d) Local Relevance

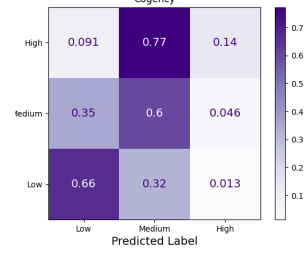


(e) Local Sufficiency

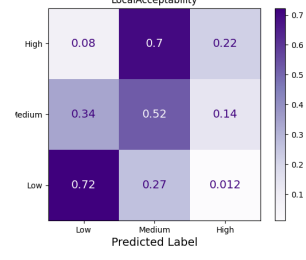
Figure 15: Confusion Matrices for the Gemma 4 31B model while analyzing the different quality dimensions with the Range 1-12 few-shot prompt variant.



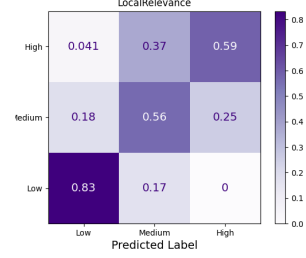
(a) Overall Quality



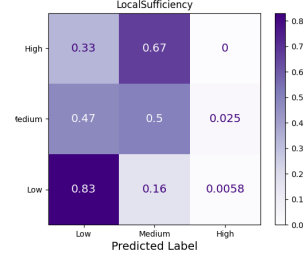
(b) Cogency



(c) Local Acceptability

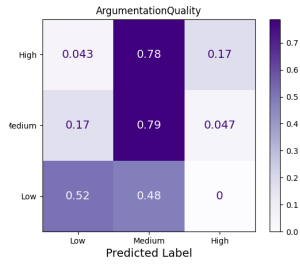


(d) Local Relevance

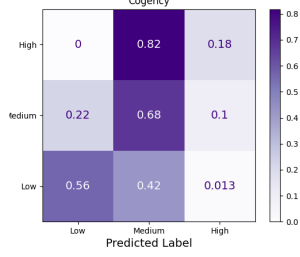


(e) Local Sufficiency

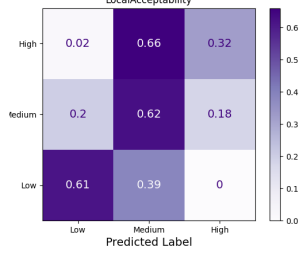
Figure 16: Confusion Matrices for the Qwen 3.5 27B model while analyzing the different quality dimensions with the Range 1-12 few-shot prompt variant.



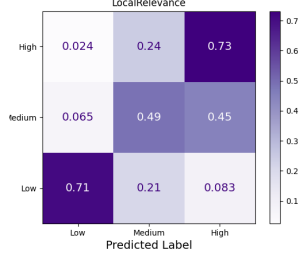
(a) Overall Quality



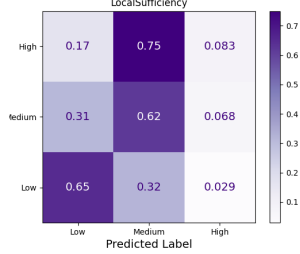
(b) Cogency



(c) Local Acceptability

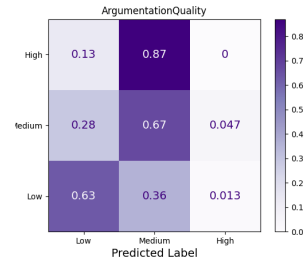


(d) Local Relevance

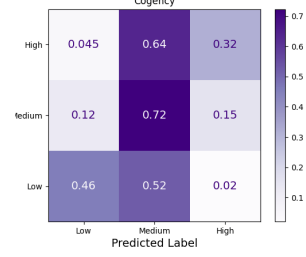


(e) Local Sufficiency

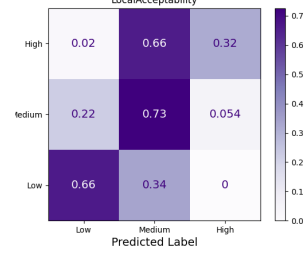
Figure 17: Confusion Matrices for the Phi 4 15B model while analyzing the different quality dimensions with the Range 1-12 few-shot prompt variant.



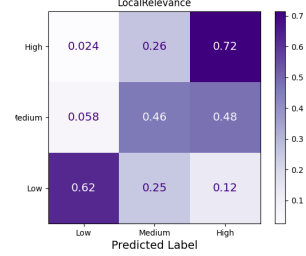
(a) Overall Quality



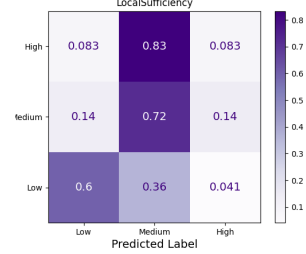
(b) Cogency



(c) Local Acceptability



(d) Local Relevance



(e) Local Sufficiency

Figure 18: Confusion Matrices for the Llama 3.3 70B model while analyzing the different quality dimensions with the Range 1-12 few-shot prompt variant.

D Distribution Matrices of Unconverted Scores for different LLMs

We present in this section the distribution of scores predicted by each LLM under the 1-12 range prompt variant, before converting them to the original 3-point range. It is possible to observe that most LLMs can leverage this extended range, as at least one argument is scored with the new possible scores. One interesting thing to note is that, even with this increased scoring range, LLMs are still reluctant to attribute the maximum score for an argument, with very few examples scored with the maximum score (as seen in Figures 20d, 21c, and 21d).

One interesting behavior is shown by the Llama 3.3 70B model (Figure 22). We observe a tendency for scores to cluster around the decision boundaries between quality classes. In particular, many arguments receive scores of 4 and 8, which correspond to the highest scores still assigned to the low- and medium-quality classes, respectively. Conversely, scores for high-quality arguments are concentrated near the lower end of the high-quality range, especially around 9. This pattern suggests that the model often places arguments close to class frontiers rather than distributing scores evenly within each class. The clustering around the 8–9 boundary may partly explain the observed improvement in model performance, as small score differences in this region can directly affect class assignment. However, further experiments are needed to determine why this boundary-focused scoring behavior emerges more strongly in this model.

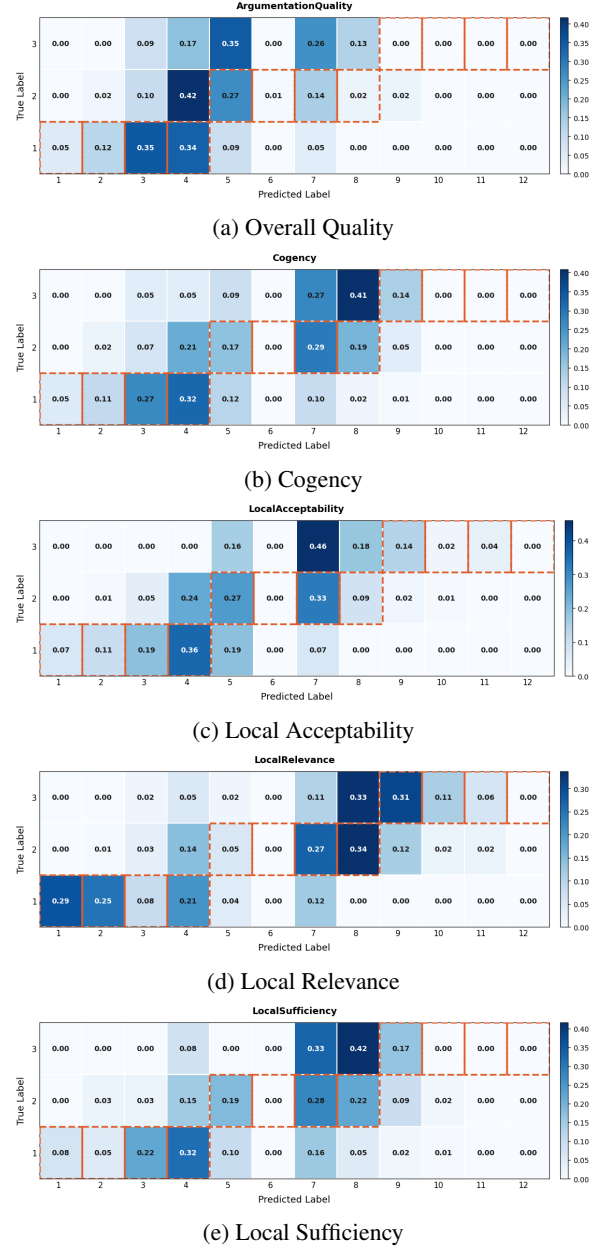


Figure 19: Matrices representing the unconverted scores for the Gemma 4 31B model while analyzing the different quality dimensions with the Range 1-12 few-shot prompt variant. The highlighted squares represent the set of scores that match the correct score when converted back to the original range.

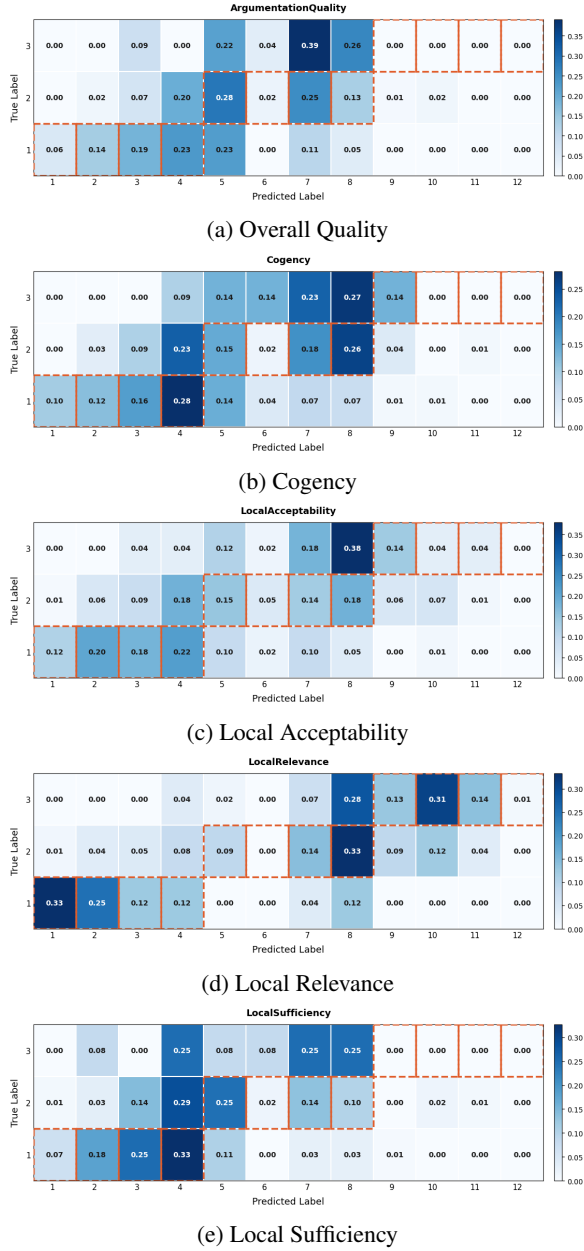
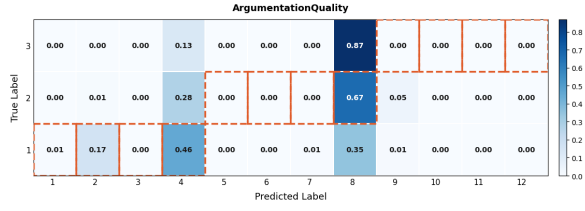


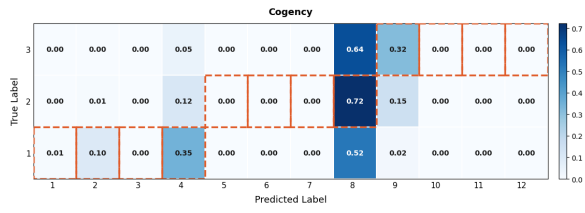
Figure 20: Matrices representing the unconverted scores for the Qwen 3.5 27B model while analyzing the different quality dimensions with the Range 1-12 few-shot prompt variant. The highlighted squares represent the set of scores that match the correct score when converted back to the original range.



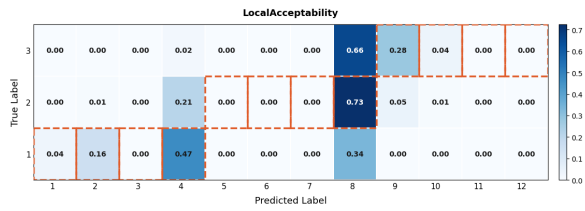
Figure 21: Matrices representing the unconverted scores for the Phi 4 15B model while analyzing the different quality dimensions with the Range 1-12 few-shot prompt variant. The highlighted squares represent the set of scores that match the correct score when converted back to the original range.



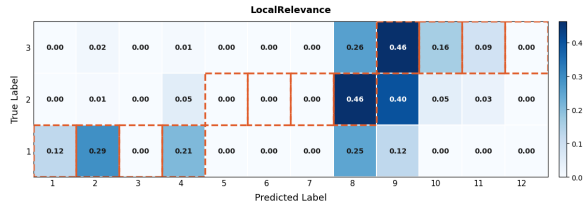
(a) Overall Quality



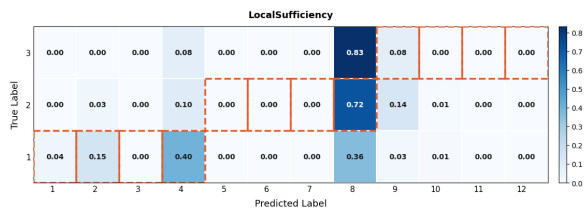
(b) Cogency



(c) Local Acceptability



(d) Local Relevance



(e) Local Sufficiency

Figure 22: Matrices representing the unconverted scores for the Llama 3.3 70B model while analyzing the different quality dimensions with the Range 1-12 few-shot prompt variant. The highlighted squares represent the set of scores that match the correct score when converted back to the original range.