Are Large Language Models Reliable Argument Quality Annotators?*

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein

Bauhaus-Universität Weimar, Germany
{first_name, last_name}@uni-weimar.de

Abstract. Evaluating the quality of arguments is a crucial aspect of any system leveraging argument mining. However, it is a challenge to obtain reliable and consistent annotations regarding argument quality, as this usually requires domain-specific expertise of the annotators. Even among experts, the assessment of argument quality is often inconsistent due to the inherent subjectivity of this task. In this paper, we study the potential of using state-of-the-art large language models (LLMs) as proxies for argument quality annotators. To assess the capability of LLMs in this regard, we analyze the agreement between model, human expert, and human novice annotators based on an established taxonomy of argument quality dimensions. Our findings highlight that LLMs can produce consistent annotations, with a moderately high agreement with human experts across most of the quality dimensions. Moreover, we show that using LLMs as additional annotators can significantly improve the agreement between annotators. These results suggest that LLMs can serve as a valuable tool for automated argument quality assessment, thus streamlining and accelerating the evaluation of large argument datasets.

Keywords: Argumentation quality · Automated argument quality assessment · Large language models · Argument mining

1 Introduction

Computational argumentation is an interdisciplinary research field that combines natural language processing with other disciplines such as artificial intelligence. A central question in computational argumentation is: What makes an argument good or bad? Depending on the goal of the author of a text, argument quality can involve a variety of dimensions. Evaluating the quality of an argument across these diverse dimensions demands a deep understanding of the topic at hand, often coupled with expertise from the argumentation literature. Hence, manual assessment of argument quality is a challenging and time-consuming process.

A promising technology to streamline argument quality assessment are large language models (LLMs) which have demonstrated impressive capabilities in tasks that require a profound understanding of semantic nuances and discourse

structures. LLMs have been effectively employed in tasks such as summarization [32], question answering [16], and relation extraction [31]. Previous research has also investigated the usefulness of LLMs in argument mining tasks such as argument component identification [12], evidence detection [15], and stance classification [4]. Moreover, an emerging trend highlights the adoption of LLMs for data annotation purposes, such as sentiment analysis [8,24], relevance judgement [11], and harm measurement [18]. To the best of our knowledge, no prior work has investigated the potential of LLMs as annotators of argument quality.

In this paper, we analyze the reliability of LLMs as argument quality annotators by comparing automatic quality judgements with human annotations from both experts and novices.[1] We compare these quality ratings not only at an aggregate level, but also examine the individual components that make up argument quality. This includes looking at how well models can judge the relevance and coherence of an argument, the sufficiency of its evidential support, and the effectiveness of its rhetorical appeal. Ultimately, our objective is to understand whether LLMs can serve as a practical and reliable tool that supports and enhances human-led effort in argument quality assessment.

Specifically, we ask the following research questions regarding the potential of employing LLMs as argument quality annotators:

RQ1: Do LLMs provide more consistent evaluations of argument quality compared to human annotators?

RQ2: Do the assessments of argument quality made by LLMs align with those made by either human experts or human novices?

RQ3: Can integrating LLM annotations with human annotations significantly improve the resulting agreement in argument quality ratings?

In the following, Section 2 reviews the related work, Section 3 describes the experimental setup, including the dataset, the annotation procedure, and the employed models, and Section 4 presents the results of these experiments.

2 Related Work

We first review existing literature related to the evaluation and annotation of argument quality. Following that, we explore the works that examined the capabilities of large language models (LLMs) as data annotators as well as the degree of alignment between LLMs and human annotators.

2.1 Evaluating Argument Quality

Collecting argument quality annotations is an intricate task that often requires domain-specific knowledge, a number of annotators, and assured consistency in annotator reliability. Numerous works have studied argumentation quality across

---

[1] Code and data are available at `github.com/webis-de/RATIO-24`.

different domains, employing multiple annotators to classify and evaluate arguments based on various quality criteria. Park and Cardie [22] studied argumentation quality in the domain of web discourse. They employed two annotators to classify 9,000 web-based propositions into four categories based on their level of *verifiability*. Habernal and Gurevych [13] let five crowd-workers annotate a dataset consisting of 16,000 pairs of arguments with a binary *"is more convincing"* label, providing explanations for their decisions. Toledo et al. [27] collected a dataset of 14,000 pairs of arguments, each annotated with relative argument quality scores ranging from 0 to 1. They employed between 15 and 17 annotators for each instance to enhance the reliability of the collected annotations.

In the domain of student essays, Persing and Vincent [23] instructed six human annotators to evaluate 1,000 essays based on the *strength* of argumentation on a scale from 1 to 4. Carlile et al. [3] considered *persuasiveness* as the most important quality dimension of an argumentative essay. They asked two native English speakers to annotate 102 essays with argument components, argument persuasiveness scores, and further attributes such as *specificity, evidence, eloquence, relevance*, and *strength*, that determine the persuasiveness of an argument. Moreover, Marro et al. [19] employed three expert annotators for the annotation of essay components of Stab and Gurevych [25] for three basic argument quality dimensions: *cogency, rhetoric*, and *reasonableness*.

Aiming to create a unified understanding of argument quality properties, Wachsmuth et al. [30] proposed a comprehensive taxonomy of 15 argument quality dimensions derived from the argumentation literature. Three expert annotators were employed to annotate 320 arguments [13]. In Section 3, we use their quality annotations from 1 (low) to 3 (high) as a reference for our experiments.

Despite the multiple attempts and methodologies to evaluate argument quality, the process remains labor-intensive, time-consuming, and requires a significant degree of expertise. To facilitate the task of argument quality annotation, we propose employing LLMs, as they can potentially provide more reliable and consistent annotations while significantly reducing the required manual effort.

2.2 LLMs as Annotators

Recent work has expanded the role of LLMs from language generation and explored the potential of using LLMs as data annotators. Ding et al. [8] assessed the performance of GPT-3 [2] as a data annotator for sentiment analysis, relation extraction, named entity recognition, and aspect sentiment triplet extraction. They compared the efficiency of BERT [7], trained using data annotated by GPT-3, against BERT trained with human-annotated data. Their findings showed a noticeably similar performance level with substantially reduced annotation costs, promising a potentially cost-effective alternative in using GPT-3 for annotation. A study by Gilardi et al. [11] cross-examined the annotations by ChatGPT [20] and those by crowd-workers against expert annotations across four tasks: content relevance assessment, stance detection, topic detection, and general frame detection. They found that ChatGPT not only outperforms crowd-workers in terms of accuracy, but also shows a high degree of consistency in annotations.

The study by Gao et al. [10] explored automatic human-like evaluations of text summarization using ChatGPT compared to human experts. The model was prompted to evaluate the quality of summaries based on *relevance*, *coherence*, *fluency*, and *consistency* of the generated summaries. The authors found that ChatGPT's evaluations were highly correlated with those of human experts.

Zhuo et al. [35] proposed to use LLMs as evaluators of code generation. The authors used the CoNaLa dataset [34] and reported high example-level and corpus-level Kendall-Tau, Pearson, and Spearman correlations with human-rated code usefulness for various programming languages.

In the domain of information retrieval, Faggioli et al. [9] investigated the performance of GPT-3.5 and YouChat for query-passage relevance judgements. Given the high subjectivity of the task, their results showed a reasonable correlation between highly-trained human assessors and fully automated judgements.

Closest to our work is that by Chiang et al. [5], who compared the judgments of GPT-3 on text quality to expert human judgments on a 5-point Likert scale for four quality attributes: *grammaticality, cohesiveness, likability*, and *relevance*. Their findings revealed varying degrees of positive correlations between GPT-3 and human judgments, ranging from weak to strong.

When compared to existing research, our work pioneers the study of argument quality annotations generated by LLMs. In order to provide a thorough evaluation, we use an inter-annotator agreement metric to assess the consistency of annotations from these models, human experts and novices. This comparison allows us to understand the alignment between LLMs and human annotators, and to determine the potential of using LLMs as argument quality annotators.

3 Experimental Design

To investigate the reliability of large language models (LLMs) as annotators of argument quality, we conduct an experiment comparing human annotations with ratings generated automatically by LLMs. We treat LLMs as separate annotators and analyze the agreement both within and across groups of humans and models.

3.1 Expert Annotation

The goals of argumentation are manifold and include persuading audiences, resolving disputes, achieving agreement, completing inquiries, or recommending actions [26]. Due to the variety of these goals, the dimensions of argument quality are equally diverse. Based on a comprehensive survey of argumentation literature, Wachsmuth et al. [30] proposed a fine-granular taxonomy of argument quality dimensions that differentiates logical, rhetorical, and dialectic aspects. An overview of all quality dimensions is provided in Table 1.

In their work, Wachsmuth et al. [30] employed experts to rate the quality of arguments according to their proposed taxonomy. Three experts were selected out of a pool of seven based on their agreement in a pilot annotation study. These three experts comprised two PhDs and one PhD student (two female, one male)

Table 1. Descriptions of argument quality dimensions as per Wachsmuth et al. [29].

Quality Dimension	Description
Cogency	Argument has (locally) acceptable, relevant, and sufficient premises.
Local acceptability	Premises worthy of being believed.
Local relevance	Premises support/attack conclusion.
Local sufficiency	Premises enough to draw conclusion.
Effectiveness	Argument persuades audience.
Credibility	Makes author worthy of credence.
Emotional appeal	Makes audience open to arguments.
Clarity	Avoids deviation from the issue, and uses correct and unambiguous language.
Appropriatness	Language proportional to the issue, supports credibility and emotions.
Arrangement	Argues in the right order.
Reasonableness	Argument is (globally) acceptable, relevant, and sufficient.
Global acceptability	Audience accepts use of argument.
Global relevance	Argument helps arrive at agreement.
Global sufficiency	Enough rebuttal of counterarguments.
Overall quality	Argumentation quality in total.

from three different countries. To construct the Dagstuhl-15512-ArgQuality corpus, the selected experts annotated 320 arguments from the UKPConvArgRank dataset [13]. The resulting corpus contains 15 quality dimensions for each argument, each rated on a 3-point Likert scale (low, medium, high) or as not assessable. Each argument in the corpus belongs to one of 16 different topics and takes a stance for or against the topic. The dataset is balanced and contains 10 supporting and 10 attacking arguments per topic. The annotation guidelines, which define all quality dimensions in more detail, are publicly available online.[2]

3.2 Novice Annotation

To provide an additional point of reference for determining the abilities of LLMs as argument quality annotators, we conducted an annotation study involving humans with no prior experience with computational argumentation. We asked undergraduate students to assess the quality of arguments from the Dagstuhl-15512-ArgQuality corpus using the same taxonomy.

The expert annotation guidelines require that annotators have expertise in computational argumentation. To make this task accessible for novices, we paraphrased the annotation guidelines and the definitions of argument quality dimensions to ensure clarity and comprehension. These simplified definitions for each quality dimension can be found in Appendix. To illustrate, the expert definition of *local acceptability* of an argument is stated as follows:

---

[2] https://zenodo.org/records/3973285

Definition 1 (Local Acceptability (Expert)). *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.*

The above definition requires an annotator to distinguish between premises and arguments. To ease the understanding and reduce the necessary prior knowledge, we simplify the definition of local acceptability as follows:

Definition 2 (Local Acceptability (Novice)). *The reasons are individually believable: they could be true.*

We refer to arguments as "reasons" within the simplified guidelines and combine the stance with the issue into a "conclusion". For example, given the issue *"Is TV better than books?"* and the stance *"No it isn't"*, we state the conclusion as *"TV is not better than books"*.

Each novice annotator was presented with an argument, a conclusion, and the simplified definitions of the quality dimensions. Identical to the annotation procedure for expert annotations, the annotators were tasked to rate each quality dimension of the argument on a 3-point Likert scale or as not assessable.

In total, we acquired 108 students to annotate the quality of the 320 arguments from the dataset. We assigned 10 arguments to each student to annotate in order to obtain at least three annotations per argument and quality dimension. Since not all students finished their annotations and some students annotated a wrong set of arguments, we obtained a minimum of three annotations per argument and quality dimension only for 248 arguments. We treat the missing annotations of the 72 arguments as non-evaluable. For the 163 arguments for which we collected more than 3 annotations, we select three annotations that maximize the inter-annotator agreement measured by Krippendorff's α.

3.3 Models

Due to the complexity of the task, we focus on state-of-the-art LLMs for the automatic evaluation of argumentation quality. Building upon previous research regarding LLMs as annotators (cf. Section 2.2), one of the most commonly used models is GPT-3 [2]. Specifically, we use the `gpt-3.5-turbo-0613` accessible via OpenAI's API.[3] Despite the availability of the newer GPT-4 model [21], we do not employ it in our study due to the significantly higher associated costs.

In addition, we use Google's recently released PaLM 2 model [1], the successor to the original PaLM model [6]. The authors report comparable results to GPT-4 in semantic reasoning tasks, which makes it interesting for the evaluation of argument quality. For PaLM 2, we use the `text-bison@001` version of the model.

Both PaLM 2 and GPT-3 are closed-source language models. We initially intended to incorporate Meta's Llama 2 model [28] in our experiments, in order to evaluate the performance of open-source LLMs on our task. However, in

---

[3] https://platform.openai.com/

```
### Instruction:
Please answer the following questions for the given comment from
an online debate forum on a given issue.

### Issue:
Is TV better than books?

### Stance:
No, it isn't.

### Argument:
Books will be always great whatever the new technological
developments emerges books has its fixed place in every humans
heart.

### Quality dimension definition:
Clarity: The style of an argumentation should be seen as clear if
it uses grammatically correct and widely unambiguous language as
well as if it avoids unnecessary complexity and deviation from the
discussed issue. The used language should make it easy for you to
understand without doubts what the author argues for and how.

### Question:
How would you rate the clarity of the style of the author's
argumentation? Choose one of the options below [and explain your
reasoning]:
3 - High
2 - Medium
1 - Low
? - Cannot judge
```

Fig. 1. An expert prompt that contains instructions and an example issue, stance, and argument from the Dagstuhl-15512 ArgQuality corpus. This particular prompt example asks the model to rate the *clarity* of the argument. The reasoning variant of this prompt is colored in gray.

pilot experiments, Llama 2 with 7 billion parameters did not follow the instructions and therefore did not produce quality scores. Even though the 13 billion parameter version of Llama 2 did generate quality scores, they were seemingly random, with agreement across the multiple runs close to zero. Due to hardware limitations, we did not test the largest Llama 2 model with 70 billion parameters.

PaLM 2 and GPT-3 allow to specify a set of parameters such as temperature to control the diversity of the output, where lowering the temperature reduces the 'randomness' of the output. For our experiments, we choose a reasonably low temperature of 0.3. Other parameters that we keep constant across models include $p = 1.0$ of the nucleus sampling [14], most probable tokens $k = 40$, and a maximum of 256 newly generated tokens.

3.4 Prompting

Two different groups of human annotators, the expert annotators of Wachsmuth et al. [30] and the novice annotators recruited for this work, had access to different knowledge sources in their annotation guidelines. To determine whether the impact of this difference is similar between humans and LLMs, we created prompts that reflect the knowledge from the annotation guidelines of experts and novices. We refer to these prompt types as *expert* and *novice* prompts.

Besides instructions, an expert prompt consists of an issue, a stance, and an argument from the Dagstuhl-15512-ArgQuality corpus. The expert prompt also contains the name and original definition of the quality dimension from Wachsmuth et al. [30] as well as the annotation scheme (3-point Likert scale or "not assessable"). An example of an expert prompt is shown in Figure 1.

In contrast to the expert prompt type, novice prompts contain a conclusion (as described in Section 3.2) instead of an issue and stance. In the novice prompt, the definition of the quality dimension to be assessed is replaced by the simplified definition. However, the textual argument, which is renamed to "reasons", and the annotation scheme remain identical to the expert prompt.

Recently, it has been shown that explanation-augmented prompts can elicit reasoning capabilities in LLMs and improve their performance across various tasks [33,17]. In pilot experiments, we found that GPT-3 produces more consistent annotations if we prompt the model to provide an explanation for the chosen score. We therefore test *reasoning* prompt variants in which we ask the model to provide an explanation for the generated quality rating.

To take the randomness of the output of LLMs for the same prompt into account, each prompt variant is repeated (at least) three times for each argument and quality dimension. Each prompt repetition is considered as a separate annotator in order to calculate the agreement between the annotation and to draw conclusions about the consistency of the quality annotations of LLMs.

4 Results

To understand the strengths and weaknesses of large language models (LLMs) as argument quality assessors and to answer our research questions, we use the prompting approaches described in Section 3.4 to generate LLM annotations for arguments from the Dagstuhl-15512-ArgQuality corpus. The dataset contains 320 statements, 16 of which were originally judged as non-argumentative by expert human annotators and therefore are excluded from the analysis.

First, to identify biases in human and LLM argument quality annotations, we analyze the distribution of assigned labels across all quality dimensions. This distribution is visualized in Figure 2. Human annotations show an almost balanced distribution between low, medium and high quality ratings. However, it is noteworthy that human novices show a tendency to assign high ratings more frequently than experts. As for models, GPT-3 with expert prompts displays a much more skewed distribution, showing a strong bias towards medium ratings,
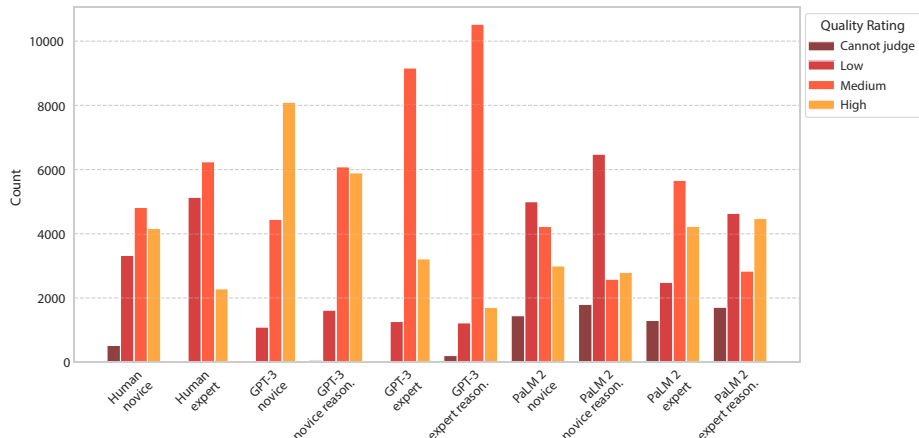
Fig. 2. Distribution of the assigned quality ratings across all quality dimensions compared between human annotators and LLMs.

deviating from the trend observed in human assessors. On the contrary, when GPT-3 is prompted with novice-level guidelines, it tends to assign high-quality ratings more frequently. Notably, annotations generated by PaLM 2 have a similar distribution to that of human annotators which seems promising for the subsequent analysis of agreement with human assessments.

Overall, it can be stated that not only the choice of model, but also the prompt type has a major influence on the generated argument quality ratings. Even slight prompt modifications, such as asking to justify the score, can result in a notable change in the assigned quality scores, which is especially prominent for PaLM 2 with expert prompts in our case. Another interesting observation is that GPT-3 almost always provides a rating for a given dimension: only in 214 out of 21,120 cases ($\approx 1\%$) this model did not generate a score. The instances where PaLM 2 did not assess argument quality sum up to 4,972 ($\approx 23\%$) and mostly stem from content policies, particularly in cases where arguments revolve around graphic topics such as pornography or contain offensive statements.

4.1 Consistency of Argument Quality Annotations

We address our first research question concerning the consistency of argument quality assessments by comparing the agreement levels within LLM groups with those of human assessors. To quantify the agreement within each group of annotators, we use Krippendorff's α. To ensure a fair comparison with human annotators, we evaluate the agreement between three LLM annotation runs.

Table 2 shows Krippendorff's α for human experts, human novices, and all LLM prompt variants across individual quality dimensions as well as overall agreement. Human annotators exhibit generally low agreement, with a maximum of 0.43 on the *local acceptability* dimension for novices and 0.45 on *reasonableness* for experts. This low level of agreement between humans emphasizes

Table 2. Inter-annotator agreement per argument quality dimension within each group of human annotators and LLMs, reported as Krippendorf's α. The dimension with the highest agreement within each group is marked in bold.

Quality Dimension	Human		GPT-3				PaLM 2			
					Reasoning				Reasoning	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Cogency	0.38	0.38	0.72	0.73	0.77	0.72	**0.99**	0.98	0.73	0.74
Local Acceptability	**0.43**	0.33	0.64	0.69	0.70	0.75	0.98	0.97	0.60	0.71
Local Relevance	0.36	0.41	0.70	0.59	0.76	0.61	0.98	0.98	0.78	0.68
Local Sufficiency	0.35	0.27	0.74	0.69	0.79	0.72	0.98	0.97	0.63	0.63
Effectiveness	0.41	0.33	0.72	0.70	0.77	0.74	0.98	**0.99**	0.78	0.79
Credibility	0.36	0.23	**0.79**	**0.79**	0.81	0.78	**0.99**	0.97	0.72	0.67
Emotional Appeal	0.35	0.21	0.73	0.56	0.74	0.70	0.98	0.97	0.64	0.72
Clarity	0.27	0.25	0.72	0.69	0.71	0.69	**0.99**	**0.99**	**0.82**	0.80
Appropriateness	0.39	0.17	0.66	0.50	0.68	0.58	**0.99**	**0.99**	0.75	0.81
Arrangement	0.39	0.26	0.68	0.66	0.71	0.69	**0.99**	**0.99**	0.69	0.65
Reasonableness	0.35	**0.45**	0.73	0.78	0.81	0.74	0.97	0.97	0.70	0.76
Global Acceptability	0.37	0.39	0.72	0.77	0.77	0.74	0.98	0.97	0.66	0.70
Global Relevance	0.38	0.26	0.69	0.71	0.81	0.70	**0.99**	0.98	0.74	**0.85**
Global Sufficiency	0.27	0.17	0.72	0.69	0.72	0.75	0.98	0.96	0.62	0.47
Overall Quality	0.41	0.44	0.77	0.77	**0.82**	**0.81**	0.98	0.97	0.77	0.78
Overall α	0.37	0.40	0.76	0.73	0.78	0.74	0.99	0.98	0.76	0.78

the subjectivity and complexity of assessing argument quality in a fine-grained taxonomy. For most of the quality dimensions, novice annotators show slightly higher agreement than those of experts, which could be due to the clearer definitions of the quality dimensions or perhaps due to the optimization of agreement for arguments that received more than three annotations (cf. Section 3.2).

In contrast, LLM agreement between annotation repetitions is substantially higher. Interestingly, the PaLM 2 model shows near-perfect agreement for both expert and novice prompts, but shows a notable drop when asked to explain its reasoning. In contrast to PaLM 2, the GPT-3 model exhibits a slight improvement in agreement when asked to provide an explanation. Such disparities might be due to the differences in the underlying architectures and training methodologies of the two models, which requires further exploration beyond the work at hand. Overall, both models show a high degree of agreement across different runs, with varying impact of reasoning prompts on the agreement depending on the employed model.

RQ1: Do LLMs provide more consistent evaluations of argument quality compared to human annotators? The observed low agreement among human annotators underscores that evaluating argument quality is indeed a subjective and challenging task. In contrast, the significantly higher agreement among different LLM runs highlights the potential of these models for providing more consistent argument quality annotations.

Table 3. Number of arguments with perfect agreement for each argument dimension within each group of human annotators (expert, novice).

Quality Dimension	Expert	Novice
Cogency	122	105
Local Acceptability	82	115
Local Relevance	99	100
Local Sufficiency	113	89
Effectiveness	128	118
Credibility	115	86
Emotional Appeal	130	90
Clarity	89	92
Appropriateness	53	102
Arrangement	81	102
Reasonableness	126	119
Global Acceptability	96	102
Global Relevance	66	96
Global Sufficiency	136	81
Overall Quality	134	130

4.2 Agreement between Humans and LLMs

We discovered that LLMs generate annotations more consistently than humans. However, to assert that LLMs can reliably evaluate the quality of arguments, we need to test how the automatic annotations align with the human annotations. Given the low agreement among human annotators, we created subsets of arguments for each quality dimension, where either all expert annotators or all novice annotators unanimously agreed on a score. Table 3 presents the statistics of the resulting subsets with perfect agreement, which we employ for further inter-annotator agreement analysis.

Figure 3 shows the agreement for each quality dimension, as measured by Krippendorff's α, between human annotations and automatically generated quality ratings by LLMs with different prompts. Overall, we observe moderate agreement across most quality dimensions, with the annotations by PaLM 2 reaching a maximum of 0.71 for *appropriateness* and *global relevance*. Regardless of the prompt type, PaLM 2 annotations generally achieve higher agreement with human annotations compared to GPT-3. In the case of *local* and *global sufficiency*, there are even systematic disagreements between the GPT-3 assessments and those of human experts. Similarly, disagreement is observed between PaLM 2 annotations and human novices for the *global sufficiency* dimension.

Overall, there is a large variance in agreement between model and human judgments across different quality dimensions. For example, while the agreement on *credibility* and *appropriateness* are in the range of [0.08, 0.62] and [0.06, 0.71] respectively, the agreement on *local* and *global sufficiency* fluctuates even more.

In terms of prompt variants, we can see that GPT-3 with expert prompts shows a higher agreement with human expert annotations than with human novice annotations, and a similar trend is observed for GPT-3 with novice prompts and human novices. On the other hand, PaLM 2 with either of the
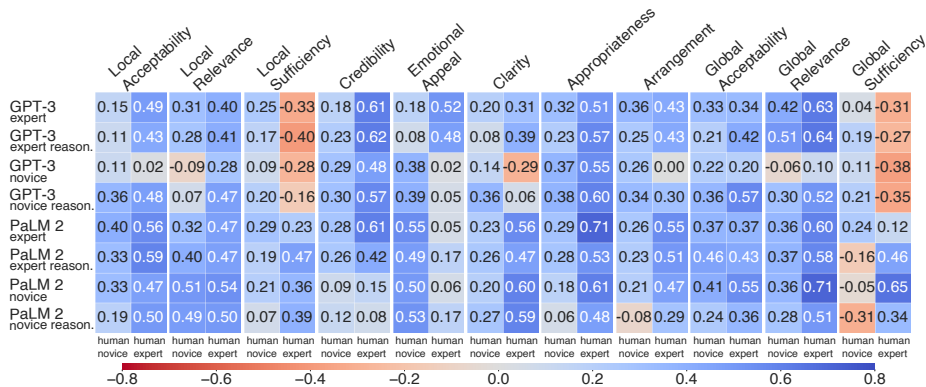
Fig. 3. Inter-annotator agreement (Krippendorff's α) between human and LLM annotations for each fine-grained argument quality dimension.

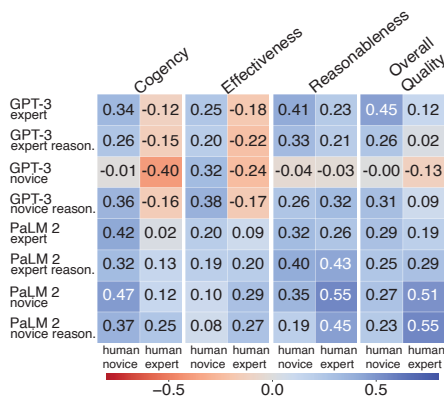|	Local Acceptability		Local Relevance		Local Sufficiency		Credibility		Emotional Appeal		Clarity		Appropriateness		Arrangement		Global Acceptability		Global Relevance		Global Sufficiency	
	novice	expert	novice	expert	novice	expert	novice	expert	novice	expert	novice	expert	novice	expert	novice	expert	novice	expert	novice	expert	novice	expert
GPT-3 expert	0.15	0.49	0.31	0.40	0.25	-0.33	0.18	0.61	0.18	0.52	0.20	0.31	0.32	0.51	0.36	0.43	0.33	0.34	0.42	0.63	0.04	-0.31
GPT-3 expert reason.	0.11	0.43	0.28	0.41	0.17	-0.40	0.23	0.62	0.08	0.48	0.08	0.39	0.23	0.57	0.25	0.43	0.21	0.42	0.51	0.64	0.19	-0.27
GPT-3 novice	0.11	0.02	-0.09	0.28	0.09	-0.28	0.29	0.48	0.38	0.02	0.14	-0.29	0.37	0.55	0.26	0.00	0.22	0.20	-0.06	0.10	0.11	-0.38
GPT-3 novice reason.	0.36	0.48	0.07	0.47	0.20	-0.16	0.30	0.57	0.39	0.05	0.36	0.06	0.38	0.60	0.34	0.30	0.36	0.57	0.30	0.52	0.21	-0.35
PaLM 2 expert	0.40	0.56	0.32	0.47	0.29	0.23	0.28	0.61	0.55	0.05	0.23	0.56	0.29	0.71	0.26	0.55	0.37	0.37	0.36	0.60	0.24	0.12
PaLM 2 expert reason.	0.33	0.59	0.40	0.47	0.19	0.47	0.26	0.42	0.49	0.17	0.26	0.47	0.28	0.53	0.23	0.51	0.46	0.43	0.37	0.58	-0.16	0.46
PaLM 2 novice	0.33	0.47	0.51	0.54	0.21	0.36	0.09	0.15	0.50	0.06	0.20	0.60	0.18	0.61	0.21	0.47	0.41	0.55	0.36	0.71	-0.05	0.65
PaLM 2 novice reason.	0.19	0.50	0.49	0.50	0.07	0.39	0.12	0.08	0.53	0.17	0.27	0.59	0.06	0.48	-0.08	0.29	0.24	0.36	0.28	0.51	-0.31	0.34

Fig. 4. Inter-annotator agreement (Krippendorff's α) between human and LLM annotations for each coarse-grained argument quality dimension.

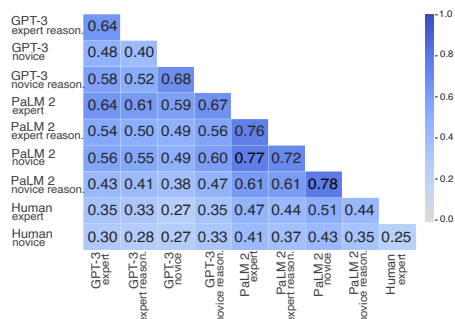|	Cogency		Effectiveness		Reasonableness		Overall Quality	
	human novice	human expert	human novice	human expert	human novice	human expert	human novice	human expert
GPT-3 expert	0.34	-0.12	0.25	-0.18	0.41	0.23	0.45	0.12
GPT-3 expert reason.	0.26	-0.15	0.20	-0.22	0.33	0.21	0.26	0.02
GPT-3 novice	-0.01	-0.40	0.32	-0.24	-0.04	-0.03	-0.00	-0.13
GPT-3 novice reason.	0.36	-0.16	0.38	-0.17	0.26	0.32	0.31	0.09
PaLM 2 expert	0.42	0.02	0.20	0.09	0.32	0.26	0.29	0.19
PaLM 2 expert reason.	0.32	0.13	0.19	0.20	0.40	0.43	0.25	0.29
PaLM 2 novice	0.47	0.12	0.10	0.29	0.35	0.55	0.27	0.51
PaLM 2 novice reason.	0.37	0.25	0.08	0.27	0.19	0.45	0.23	0.55

Fig. 5. Overall inter-annotator agreement (Krippendorff's α) between each combination of human expert, novice, and LLM-generated annotations.

	GPT-3 expert	GPT-3 expert reason.	GPT-3 novice	GPT-3 novice reason.	PaLM 2 expert	PaLM 2 expert reason.	PaLM 2 novice	PaLM 2 novice reason.	Human expert
GPT-3 expert reason.	0.64								
GPT-3 novice	0.48	0.40							
GPT-3 novice reason.	0.58	0.52	0.68						
PaLM 2 expert	0.64	0.61	0.59	0.67					
PaLM 2 expert reason.	0.54	0.50	0.49	0.56	0.76				
PaLM 2 novice	0.56	0.55	0.49	0.60	0.77	0.72			
PaLM 2 novice reason.	0.43	0.41	0.38	0.47	0.61	0.61	0.78		
Human expert	0.35	0.33	0.27	0.35	0.47	0.44	0.51	0.44	
Human novice	0.30	0.28	0.27	0.33	0.41	0.37	0.43	0.35	0.25

prompt types tends to show higher agreement with human experts. Similar findings can be inferred from the agreement between LLMs and human novices and experts on the coarse-grained quality dimensions that are visualized in Figure 4.

RQ2: Do the assessments of argument quality made by LLMs align with those made by either human experts or human novices? We found that LLMs agree most with human argument quality ratings on fine-grained quality dimensions such as *credibility, emotional appeal, appropriateness*, and *global relevance* or on coarse-grained dimensions such as *reasonableness* and *overall quality*. Overall, we found varying degrees of agreement between LLMs and human annotators, with PaLM 2 annotations tending to generally align more with those of humans.

Table 4. Change in overall Krippendorf's α after adding LLM annotations to human expert or novice annotations. Significant changes ($p < 0.05$) between the agreement of the original annotations and the modified annotations set are marked with *.

Annotations	Expert		Novice		Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2		GPT-3	PaLM 2	GPT-3	PaLM 2
Human experts	0.40	0.40	0.40	0.40	Human novices	0.37	0.37	0.37	0.37
+1 annotation	0.32*	0.37*	0.30*	0.37*	+1 annotation	0.27*	0.29*	0.27*	0.27*
+2 annotations	0.31*	0.40	0.32*	0.40	+2 annotations	0.26*	0.33	0.29*	0.30
+3 annotations	0.33	0.44*	0.35	0.44*	+3 annotations	0.28*	0.37*	0.33	0.35*
+4 annotations	0.35	0.47*	0.39	0.47*	+4 annotations	0.30*	0.41*	0.37	0.39*
+5 annotations	0.37	0.50*	0.42*	0.50*	+5 annotations	0.32*	0.45*	0.40*	0.43*

4.3 LLMs as Additional Annotators

LLMs can be employed either as independent automatic argument quality raters or as a source of additional annotations to validate a set of human annotations. For the second scenario, we analyze the overall agreement between different combinations of human (expert or novice) and LLM annotators.

Figure 5 illustrates the overall Krippendorff's α agreement for each combination of annotator group. We can see that there is low to medium agreement for each combination of annotators, with the lowest value being 0.25 between human novices and human experts and the highest value being 0.77 between PaLM with novice and expert prompts. Regardless of the prompt type, the agreement between PaLM 2 and GPT-3 is moderate, ranging from 0.38 to 0.67. This suggests the potential efficacy of employing diverse models as supplementary annotators.

We further investigate whether the agreement changes if we incrementally integrate automatically generated annotations to the original set of human annotations. The results reported in Table 4 show that adding PaLM 2 annotations can significantly improve the agreement of human experts as well as human novices. A significant increase is already visible after adding three annotations to the annotations of human experts and four to the annotations of human novices. However, the introduction of GPT-3 annotations leads to a significant decrease in agreement. This can be attributed to the relatively low level of agreement between GPT-3 and human annotators (cf. Figure 5).

RQ3: Can integrating LLM annotations with human annotations significantly improve the resulting agreement in argument quality ratings? The analysis indicates that the impact on agreement levels when incorporating generated quality assessments with human annotations varies based on the employed LLM. When using a powerful model such as PaLM 2, the agreement of human annotations can be significantly increased by adding three or more generated annotations. These results underscore LLMs as valuable contributors to the annotator ensemble.

5 Conclusion

In this paper, we investigated the effectiveness of LLMs, specifically GPT-3 and PaLM 2, in evaluating argument quality. We utilized four distinct prompt types to solicit quality ratings from these models and compared their assessments with those made by human novices and experts. The results reveal that LLMs exhibit greater consistency in evaluating argument quality compared to both novice and expert human annotators, showcasing their potential reliability. Based on our empirical analysis, we can recommend two modes of application for LLMs as annotators of argument quality: (1) a fully automatic annotation procedure with LLMs as automatic quality raters, for which we found moderately high agreement between PaLM 2 and human expert quality ratings, or (2) a semi-automatic procedure using LLMs as additional quality annotators, resulting in a significant enhancement in agreement when combined with human annotations. In both modi, LLMs can serve as a valuable tool for streamlining the argument quality annotation process on a large scale.

To further minimize annotation expenses, we intend to expand these experiments to various open-source large language models. In addition to the investigated zero-shot prompting technique, enhancing agreement with human annotations could involve utilizing few-shot prompting technique or fine-tuning LLMs based on human judgments of argument quality. We see great potential in LLMs as argument quality raters, which, if further optimized to agree more closely with human assessments, can reduce manual effort and expenses, establishing them as valuable tools in argument mining.

6 Limitations

The experiments in this paper are based on the hypothesis that multiple runs of the same model, prompt, and hyperparameters simulate different annotators as a result of nucleus sampling. This hypothesis has not yet been proven, and its validity cannot be inferred from the analysis. The higher inter-model agreement indicates a lower variance in the automatically generated annotations, which might argue against this hypothesis. Therefore, the agreement between model and human annotations has been calculated using examples with perfect agreement only in order to exclude effects of this variance. Further experiments are needed to determine how to replicate the behavior of different annotators.

Although we deeply investigate LLMs as quality assessors for arguments, the generalizability of our results beyond argumentation is not yet clear. However, due to the complexity and subjectivity of argument quality assessment, as seen from low human inter-annotator agreement, we argue that this task might be a worst-case scenario for LLMs, and we would expect comparable or even better results in less subjective task domains. However, more experiments are needed to confirm or reject this hypothesis.

References

1. Anil, R., Dai, A.M., Firat, O., Johnson, M., et al.: PaLM 2 Technical Report. CoRR **abs/2305.10403** (2023)
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., et al.: Language Models are Few-Shot Learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)
3. Carlile, W., Gurrapadi, N., Ke, Z., Ng, V.: Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. pp. 621–631. Association for Computational Linguistics (2018)
4. Chen, G., Cheng, L., Tuan, L.A., Bing, L.: Exploring the Potential of Large Language Models in Computational Argumentation. CoRR **abs/2311.09022** (2023)
5. Chiang, D.C., Lee, H.: Can Large Language Models Be an Alternative to Human Evaluations? In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 15607–15631. Association for Computational Linguistics (2023)
6. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., et al.: PaLM: Scaling Language Modeling with Pathways. CoRR **abs/2204.02311** (2022)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019)
8. Ding, B., Qin, C., Liu, L., Chia, Y.K., Li, B., Joty, S., Bing, L.: Is GPT-3 a Good Data Annotator? In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 11173–11195. Association for Computational Linguistics (2023)
9. Faggioli, G., Dietz, L., Clarke, C.L.A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., Wachsmuth, H.: Perspectives on Large Language Models for Relevance Judgment. In: Yoshioka, M., Kiseleva, J., Aliannejadi, M. (eds.) Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023. pp. 39–50. ACM (2023)
10. Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., Wan, X.: Human-like Summarization Evaluation with ChatGPT. CoRR **abs/2304.02554** (2023)

11. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. CoRR **abs/2303.15056** (2023)
12. Guo, J., Cheng, L., Zhang, W., Kok, S., Li, X., Bing, L.: AQE: Argument Quadruplet Extraction via a Quad-Tagging Augmented Generative Approach. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 932–946. Association for Computational Linguistics (2023)
13. Habernal, I., Gurevych, I.: 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics (2016)
14. Holtzman, A., Buys, J., Forbes, M., Choi, Y.: The Curious Case of Neural Text Degeneration. CoRR **abs/1904.09751** (2019)
15. Huo, S., Arabzadeh, N., Clarke, C.L.A.: Retrieving Supporting Evidence for Generative Question Answering. In: Ai, Q., Liu, Y., Moffat, A., Huang, X., Sakai, T., Zobel, J. (eds.) Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023, Beijing, China, November 26-28, 2023. pp. 11–20. ACM (2023)
16. Kamalloo, E., Dziri, N., Clarke, C.L.A., Rafiei, D.: Evaluating Open-Domain Question Answering in the Era of Large Language Models. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 5591–5606. Association for Computational Linguistics (2023)
17. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large Language Models are Zero-Shot Reasoners. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
18. Magooda, A., Helyar, A., Jackson, K., Sullivan, D., et al.: A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications. CoRR **abs/2310.17750** (2023)
19. Marro, S., Cabrio, E., Villata, S.: Argumentation Quality Assessment: an Argument Mining Approach. In: ECA 2022-European conference on argumentation (2022)
20. OpenAI: ChatGPT (2022)
21. OpenAI: GPT-4 Technical Report. CoRR **abs/2303.08774** (2023)
22. Park, J., Cardie, C.: Identifying Appropriate Support for Propositions in Online User Comments. In: Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA. pp. 29–38. The Association for Computer Linguistics (2014)
23. Persing, I., Ng, V.: Modeling Argument Strength in Student Essays. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. pp. 543–552. The Association for Computer Linguistics (2015)
24. Rønningstad, E., Velldal, E., Øvrelid, L.: A GPT among Annotators: LLM-based Entity-Level Sentiment Annotation. In: Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII). pp. 133–139 (2024)

25. Stab, C., Gurevych, I.: Parsing Argumentation Structures in Persuasive Essays. Comput. Linguistics **43**(3), 619–659 (2017)
26. Tindale, C.W.: Fallacies and Argument Appraisal. Cambridge University Press, 1 edn. (Jan 2007)
27. Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., et al.: Automatic Argument Quality Assessment - New Datasets and Methods. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 5624–5634. Association for Computational Linguistics (2019)
28. Touvron, H., Martin, L., Stone, K., Albert, P., et al.: Llama 2: Open Foundation and Fine-Tuned Chat Models. CoRR **abs/2307.09288** (2023)
29. Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., Stein, B.: Argumentation Quality Assessment: Theory vs. Practice. In: Barzilay, R., Kan, M.Y. (eds.) 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). pp. 250–255. Association for Computational Linguistics (Aug 2017)
30. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: 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. pp. 176–187 (2017)
31. Wadhwa, S., Amir, S., Wallace, B.C.: Revisiting Relation Extraction in the era of Large Language Models. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 15566–15589. Association for Computational Linguistics (2023)
32. Wang, Y., Zhang, Z., Wang, R.: Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method. arXiv preprint arXiv:2305.13412 (2023)
33. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
34. Yin, P., Deng, B., Chen, E., Vasilescu, B., Neubig, G.: Learning to mine aligned code and natural language pairs from stack overflow. In: Zaidman, A., Kamei, Y., Hill, E. (eds.) Proceedings of the 15th International Conference on Mining Software Repositories, MSR 2018, Gothenburg, Sweden, May 28-29, 2018. pp. 476–486. ACM (2018)
35. Zhuo, T.Y.: Large Language Models Are State-of-the-Art Evaluators of Code Generation. CoRR **abs/2304.14317** (2023)

Appendix

Table 5. The set of simplified definitions of argument quality dimensions.

Quality Dimension	Definition
Local acceptability	The reasons are individually believable: they could be true.
Local relevance	The reasons (assuming they are true) are relevant to the conclusion: they tell why one could accept the conclusion.
Local sufficiency	The reasons (assuming they are true) are sufficient to draw the conclusion: no other reason is necessary to arrive at the conclusion.
Credibility	The reasons make the author seem trustworthy and knowledgeable.
Emotional appeal	The reasons can make people feel emotions that make them willing to agree with the author.
Clarity	The author uses clear, grammatically correct and unambiguous language. The author sticks to the main topic and does not make things overly complicated.
Appropriatness	The author uses an appropriate style for the reasons and this style fits to the topic's importance.
Arrangement	The reasons are properly organized: coherent, easy to follow, well-structured.
Global acceptability	The reasons and conclusion combined are believable: everything taken together could be true.
Global relevance	The reasons (assuming they are true) are relevant for resolving a discussion around the conclusion's topic.
Global sufficiency	The reasons (assuming they are true) are sufficient in that they consider any counter-arguments that may arise.