

# A Multi-persona Framework for Argument Quality Assessment

Bojun Jin<sup>1</sup>, Jianzhu Bao<sup>1\*</sup>, Yufang Hou<sup>2</sup>, Yang Sun<sup>1,3</sup>, Yice Zhang<sup>1</sup>,  
Huajie Wang<sup>1</sup>, Bin Liang<sup>4</sup>, Ruifeng Xu<sup>1,3,5\*</sup>

<sup>1</sup>Harbin Institute of Technology, Shenzhen, China

<sup>2</sup>IT:U Interdisciplinary Transformation University Austria

<sup>3</sup>Peng Cheng Laboratory, China <sup>4</sup>The Chinese University of Hong Kong

<sup>5</sup>Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies  
24S051022@stu.hit.edu.cn jianzhubao@gmail.com xuruifeng@hit.edu.cn

## Abstract

Argument quality assessment faces inherent challenges due to its subjective nature, where different evaluators may assign varying quality scores for an argument based on personal perspectives. Although existing datasets collect opinions from multiple annotators to model subjectivity, most existing computational methods fail to consider multi-perspective evaluation. To address this issue, we propose MPAQ, a multi-persona framework for argument quality assessment that simulates diverse evaluator perspectives through large language models. It first dynamically generates targeted personas tailored to an input argument, then simulates each persona's reasoning process to evaluate the argument quality from multiple perspectives. To effectively generate fine-grained quality scores, we develop a coarse-to-fine scoring strategy that first generates a coarse-grained integer score and then refines it into a fine-grained decimal score. Experiments on IBM-Rank-30k and IBM-ArgQ-5.3kArgs datasets demonstrate that MPAQ consistently outperforms strong baselines while providing comprehensive multi-perspective rationales.

## 1 Introduction

Argument quality assessment, a fundamental task in computational argumentation, aims to automatically evaluate the persuasiveness and validity of arguments in natural language texts (Wachsmuth et al., 2017; Gienapp et al., 2020; Wachsmuth et al., 2024). It plays a crucial role in multiple domains, such as automatic essay scoring (Persing and Ng, 2013; Nguyen and Litman, 2018; Ding et al., 2023), writing support (Stab and Gurevych, 2014, 2017), and argument generation (Slonim et al., 2021; Lin et al., 2023; Hu et al., 2024). Recent work has predominantly framed argument quality assessment as a ranking task (Gretz et al., 2020; Wang et al., 2023b; Deshpande et al., 2024), where given an

\* Corresponding authors.

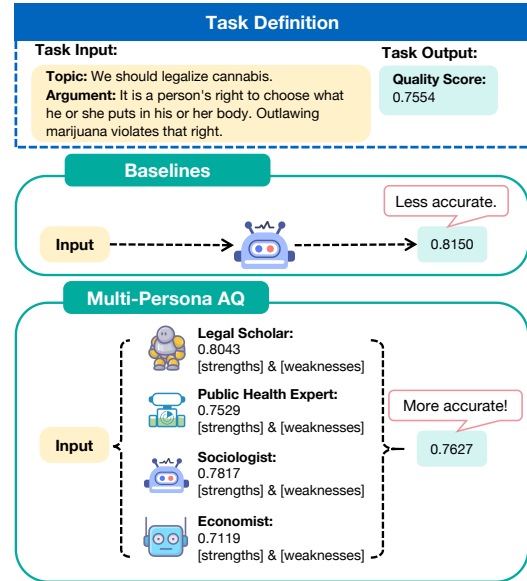


Figure 1: Argument quality assessment takes an argument and its topic as input and outputs a quality score. Unlike baselines that directly predict a single quality score, our framework generates scores from multiple personas, effectively modeling the subjective nature of argument quality assessment.

argument and its associated topic, the goal is to assign a quality score, as illustrated in Figure 1.

However, developing reliable computational models for this task presents significant challenges, primarily due to its inherently subjective nature (Wachsmuth et al., 2017, 2024; Elaraby et al., 2024). One important aspect of subjectivity is that different evaluators may assign varying quality scores to the same argument based on their prior knowledge, domain expertise, and personal beliefs. This subjective nature of argument quality assessment has been well-recognized and carefully considered in data collection efforts. For instance, large-scale datasets such as IBM-Rank-30k (Gretz et al., 2020) and IBM-ArgQ-5.3kArgs (Toledo et al., 2019) employ more than ten crowd annotators per argument to capture diverse perspectives.

tives. These annotators provide binary judgments indicating whether they would recommend the argument to others, which are then reliably aggregated into a consolidated quality score. Despite this careful consideration of subjectivity in data collection efforts, existing computational methods for argument quality assessment often overlook this crucial aspect (Wang et al., 2023b; Bao et al., 2024). Such methods typically employ a single, generic evaluator for argument quality assessment (Figure 1: Baselines), failing to explicitly capture the diverse perspectives necessary for robust evaluation.

This limitation motivates us to rethink the design of argument quality assessment methods: instead of pursuing a single universal standard, we should embrace multiple perspectives in quality prediction, similar to how human annotators approach this task. Inspired by recent advances in leveraging large language models (LLMs) for persona simulation (Tseng et al., 2024), we propose utilizing LLMs’ role-playing capabilities to assess argument quality from multiple personas’ perspectives, thereby obtaining more comprehensive and accurate quality predictions. To this end, we design a multi-persona framework for argument quality assessment (MPAQ, § 3). As illustrated in Figure 1, our framework first dynamically generates targeted personas tailored to the input argument and topic (§ 3, stage 1). Then, MPAQ simulates the reasoning process of each persona, not only assigning a quality score but also analyzing the argument’s strengths and weaknesses from unique perspective (§ 3, stage 2). Finally, the predicted scores from all personas are aggregated into an overall quality score. Furthermore, to address the challenge of generating fine-grained decimal quality scores (e.g., 0.7554) with LLMs, we incorporate a coarse-to-fine scoring strategy: MPAQ first generates a coarse-grained integer score through text generation, which is subsequently refined into a fine-grained decimal score via a regression head.

We conduct extensive experiments on two commonly used datasets (§ 4): IBM-Rank-30k (Gretz et al., 2020) and IBM-ArgQ-5.3kArgs (Toledo et al., 2019). Our experimental results demonstrate that MPAQ consistently outperforms several strong baseline methods, highlighting the advantages of assessing argument quality from multiple personas’ perspectives. Moreover, compared to existing methods, MPAQ offers a significant advantage in its ability to generate comprehensive multi-perspective

rationales for its quality assessments while predicting fine-grained quality scores.<sup>1</sup> Further analysis validates the effectiveness of key design choices in our framework, including the coarse-to-fine scoring strategy, dynamic persona generation, and two-stage architecture design, among others (§ 4.6).

## 2 Related Work

### 2.1 Argument Quality Assessment

The definition of argument quality is a complex problem, and many studies have conducted exploration on this question (Swanson et al., 2015; Wachsmuth et al., 2016; Joshi et al., 2023; Fromm et al., 2023). Building on a comprehensive summary of previous work (Hamblin, 1970; Johnson and Blair, 1977; Aristotle and Kennedy, 1991; Eemeren and Grootendorst, 2003), Wachsmuth et al. (2017) proposed a taxonomy of argumentation quality with three major dimensions: logic, rhetoric, and dialectic. According to their definitions, logic focuses on whether the argument is built on acceptable and relevant premises that are sufficient to support the conclusion, while rhetoric assesses the argument’s ability to persuade the intended audience of the author’s stance, and dialectic examines whether the argument contributes meaningfully and acceptably to resolving the issue for the target audience. Based on this taxonomy, Lauscher et al. (2020) constructed an argument quality corpus and explored interactions between different dimensions. Toledo et al. (2019) presented an argument quality annotation method that can transform binary judgments made by multiple annotators for a given argument into a reliable overall argument quality score. Based on this method, they also constructed an argument quality dataset containing approximately 5.3k arguments annotated with the overall quality scores. Gretz et al. (2020) released a larger argument quality dataset with around 30k arguments, following the same annotation protocol.

Following these data collection efforts, the computational methods for argument quality assessment have evolved significantly. Marro et al. (2022) used argument structure information derived from graph embeddings to enhance the performance of argument quality assessment. Wang et al. (2023b) leveraged contrastive learning to distinguish arguments of different quality more effectively. Bao

<sup>1</sup>The code is available at <https://github.com/bj-jin/MPAQ>.

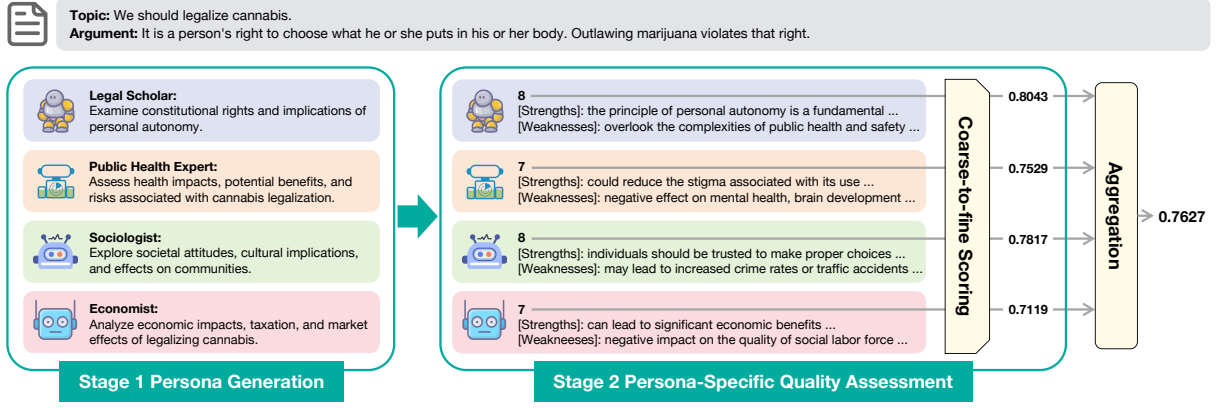


Figure 2: The architecture diagram of MPAQ.

et al. (2024) designed a framework to score the quality of an argument through multiple pairwise comparisons. Deshpande et al. (2024) proposed a framework that enhanced argument quality assessment by contextualizing arguments with relevant knowledge. Rescala et al. (2024) investigated LLMs’ capabilities in recognizing convincing arguments. Elaraby et al. (2024) analyzed the persuasiveness of rationales generated by LLMs in pairwise argument ranking.

However, existing studies have not effectively modeled the inherent subjectivity of argument quality assessment, as highlighted in previous work (Wachsmuth et al., 2017; Toledo et al., 2019; Heinisch et al., 2023). To tackle this challenge, we propose a multi-persona framework that leverages recent advances in LLMs to simulate diverse personas, thereby improving the robustness of argument quality assessment.

## 2.2 Persona Simulation

Recent studies have increasingly explored the use of LLMs for persona simulation, aiming to enhance various tasks and applications (Tseng et al., 2024). For instance, Kong et al. (2024) proved that assigning LLMs suitable persona (e.g., math teacher) can improve performance on reasoning tasks. Wu et al. (2023) improved the comprehensiveness and accuracy of text summarization evaluation by enabling LLMs to assess content from diverse persona perspectives. In our work, we use LLMs to simulate multiple personas, capturing diverse perspectives to enhance argument quality assessment.

## 3 Method

As shown in Figure 1, the input of argument quality assessment is an argument  $a$  and its topic  $t$ ,

the output is a fine-grained argument quality score  $s^f \in [0.0, 1.0]$ . A higher score indicates a better quality of  $a$ .

We propose a multi-persona framework for argument quality assessment (MPAQ), which consists of two key stages: *persona generation* and *persona-specific quality assessment*. In the first stage, MPAQ dynamically generates multiple relevant personas based on the input argument. In the second stage, it leverages these personas to perform fine-grained quality assessment from multiple perspectives, ultimately aggregating their scores into a final quality score. The architecture of MPAQ is illustrated in Figure 2.

### 3.1 Stage 1: Persona Generation

This stage aims to generate diverse and relevant persona information based on the input argument  $a$  and topic  $t$ . To develop this capability in MPAQ, we leverage advanced LLMs (e.g., GPT-4o-mini (OpenAI, 2023)) to construct training data, following existing studies in data synthesis (Wang et al., 2024; Long et al., 2024). The constructed data is then used to fine-tune a smaller open-source LLM (e.g., Llama-3.1-8B (Dubey et al., 2024)) for persona generation.

### Training Data Construction for Persona Generation.

We use carefully crafted prompt (Table 15 in Appendix C) to instruct an advanced LLM to generate diverse and relevant persona information for each input argument. This includes multiple personas, with each consisting of two components: the persona designation (e.g., legal scholar, economist) and the potential analytical perspective that this persona might adopt when examining the quality of the given argument. Some examples are shown in Figure 2 (Stage 1). By incorporating analytical

perspectives alongside persona designations, we enable a more thorough examination of arguments from well-defined viewpoints that are both relevant to each persona’s expertise and specifically tailored to the argument’s content.

In practice, we generate  $N$  personas<sup>2</sup>  $\{P_i\}_{i=1}^N$  for each training sample in the IBM-Rank-30k and IBM-ArgQ-5.3kArgs datasets. Each persona is represented as  $P_i = (p_i^d, p_i^p)$ , where  $p_i^d$  denotes the persona designation and  $p_i^p$  denotes the analytical perspective.

**Fine-tuning for Persona Generation.** After constructing the training data, we conduct supervised fine-tuning (SFT) on a relatively small open-source LLM via LoRA (Xu et al., 2024) to enable efficient persona generation. The SFT objective is to dynamically generate appropriate personas  $\{P_i\}_{i=1}^N$  given an input argument  $a$  and its topic  $t$ :

$$\{P_i\}_{i=1}^N = \text{LLM}_{\theta^g}(a, t) \quad (1)$$

where  $\text{LLM}_{\theta^g}(\cdot)$  is the fine-tuned LLM with LoRA adapter  $\theta^g$  for persona generation. We use the standard negative log likelihood loss for conditional language modeling during training. The prompt template for SFT is shown in Table 16 (Appendix C).

### 3.2 Stage 2: Persona-specific Quality Assessment

Following persona generation, this stage aims to assess argument quality from each persona’s unique perspective. Specifically, we need to train a persona-specific quality assessment model that can (1) predict a persona-specific quality score, and (2) generate corresponding rationales. To train such a model, appropriate training data is essential. We first construct training data for rationale generation via advanced LLMs, then leverage an indirect supervision strategy to obtain training labels for persona-specific quality score prediction. Finally, we fine-tune an LLM using these training data with a coarse-to-fine scoring strategy.

**Training Data Construction for Rationale Generation.** Given that the training sets of IBM-Rank-30k and IBM-ArgQ-5.3kArgs only provide overall quality scores, we need to construct training data for rationale generation, similar to our approach in the persona generation phase. We again leverage an advanced LLM to generate detailed

analytical rationales from each persona’s perspective, using a carefully designed prompt (Table 17 in Appendix C). For each persona  $P_i = (p_i^d, p_i^p)$  generated in § 3.1<sup>3</sup>, this prompt guides the LLM to analyze argument quality by examining both strengths and weaknesses. This balanced evaluation structure enables a thorough assessment of arguments by capturing both their merits and limitations, leading to more comprehensive and nuanced quality assessment. We denote the strengths and weaknesses analyses from each persona  $P_i$  for a given input argument as  $r_i^s$  and  $r_i^w$ , respectively.

**Indirect Supervision for Persona-specific Quality Score Prediction.** Most large datasets for argument quality assessment only provide an aggregated quality score for each argument (Gretz et al., 2020; Toledo et al., 2019). This creates a key challenge for our framework, as we lack gold labels for persona-specific quality scores. To address this, we adopt an indirect supervision strategy utilizing the aggregated scores as indirect supervision signals for persona-specific quality score prediction during training. This strategy is grounded in our assumption: a persona in our framework represents a collective perspective shared by annotators with similar backgrounds and expertise. While individual annotators may show significant variance in their scoring, the aggregated score from a group of annotators sharing similar persona characteristics should exhibit relatively low deviation from the overall aggregated score. Based on this assumption, we treat the indirect supervision signals as a reasonable approximation of the gold labels for persona-specific quality scores. Also, our experimental results (§ 4.5 and § A.8) validate the effectiveness of this indirect supervision strategy.<sup>4</sup>

Under this indirect supervision strategy, when training our quality assessment model, we use the aggregated quality score  $s^f$  provided in the dataset as labels for persona-specific quality scoring.

**Fine-tuning for Coarse-to-fine Scoring.** Our goal here is to train an LLM that takes an argument  $a$ , a topic  $t$  and a specific persona  $P_i$  as input, then simultaneously predicts a fine-grained persona-specific decimal quality score  $s_i^{f,p}$  and generates corresponding textual rationales  $r_i^s$  and  $r_i^w$ . However, our preliminary analysis (shown as

<sup>3</sup>Here we use the persona information generated by the advanced LLM in Section § 3.1.

<sup>4</sup>We also explored an alternative supervision strategy, which are detailed in § A.8.

<sup>2</sup>We experimentally analyze the impact of the number of personas ( $N$ ) in § 5.



"Llama-Text-Score" in Table 2) reveals that simply training an LLM to generate such fine-grained decimal scores (e.g., 0.7554) with rationales in text format yields poor accuracy. To address this issue, we propose a coarse-to-fine scoring strategy that consists of two steps: (1) the LLM first generates a coarse-grained integer score (0-9)  $s_i^{c,p}$  and corresponding rationales in text format; (2) the coarse-grained score  $s_i^{c,p}$  is then refined into a fine-grained decimal score  $s_i^{f,p}$  through a regression head  $f_{\text{reg}}$ . Specifically, the prompt template we use for this process is shown in Table 1.<sup>5</sup> This coarse-to-fine scoring strategy can be formalized as:

$$s_i^{c,p}, r_i^s, r_i^w = \text{LLM}_{\theta^a}(a, t, P_i) \quad (2)$$

$$s_i^{f,p} = f_{\text{reg}}(\text{Emb}(s_i^{c,p})) \quad (3)$$

where  $\text{LLM}_{\theta^a}(\cdot)$  is the fine-tuned LLM for persona-specific quality assessment,  $\text{Emb}(s_i^{c,p})$  denotes the LLM-encoded token embedding of  $s_i^{c,p}$ .

During training, the coarse-grained score label is derived from the fine-grained score using:

$$s_i^c = \min(\lfloor s_i^f \times 10 \rfloor, 9) \quad (4)$$

where  $\lfloor \cdot \rfloor$  denotes the floor function. This ensures the coarse-grained score stays within the range [0, 9]. The training objective is the sum of two losses: (1) a standard language modeling loss for coarse-grained score and rationale generation; (2) a mean absolute error (MAE) loss for fine-grained score prediction. Note that during training, all persona-specific scores  $s_i^{f,p}$  share the same training label  $s_i^f$  due to our indirect supervision strategy. Despite using the same labels across different personas, we believe the trained model can make differentiated coarse- and fine-grained quality predictions by incorporating persona information and rationales as context during inference. We empirically validate this hypothesis through experiments in § A.10.

In this manner, we fine-tune the same small LLM as in § 3.1 to obtain another LoRA adapter  $\theta^a$  with a regression head that is responsible for producing fine-grained quality scores.

### 3.3 Inference Process

After training through Stage 1 and Stage 2, MPAQ is equipped with two LoRA adapters:  $\theta^g$  for persona generation and  $\theta^a$  for persona-specific quality assessment, both adapting the same base LLM.

<sup>5</sup>We also explore the impact of score positions in this template (§ A.6), finding that placing the coarse-grained score before the rationales leads to better performance.

---

You are a/an  $\{p_i^d\}$ . You are required to analyze the strengths and weaknesses of an argument on a controversial topic and assess its quality from your perspectives as a/an  $\{p_i^d\}$ . Your perspectives for analysis are as follows:  $\{p_i^p\}$ .

Now given an argument below, analyze its strengths and weaknesses and assess its argument quality indicated by an integer from 0 to 9 from your perspectives as a/an  $\{p_i^d\}$ .

### Target Argument:

[Topic]:  $\{t\}$

[Target Argument]:  $\{a\}$

### Output:

$\{s_i^{c,p}\}$

[Strengths]:  $\{r_i^s\}$

[Weaknesses]:  $\{r_i^w\}$

---

Table 1: Prompt template for persona-specific argument quality assessment.

During inference, given an input argument  $a$  and its topic  $t$ , MPAQ first employs  $\theta^g$  to generate  $N$  diverse personas. Subsequently,  $\theta^a$  independently assesses the argument quality from each persona’s perspective, producing fine-grained quality scores and corresponding rationales. This process yields  $N$  distinct quality scores, reflecting diverse viewpoints. The final quality score is computed by averaging these  $N$  persona-specific scores. We also explore other score aggregation strategies, which are detailed in § A.9.

## 4 Experiments and Results

### 4.1 Datasets

We evaluate MPAQ on two English datasets, namely IBM-ArgQ-5.3kArgs (Toledo et al., 2019) and IBM-Rank-30k (Gretz et al., 2020). IBM-ArgQ-5.3kArgs contains 5.3k arguments covering 22 topics. Each argument is annotated by 15 to 17 annotators to improve the reliability of the annotations. IBM-Rank-30K is a larger dataset, containing over 30k arguments covering 71 topics, with each sample annotated by 10 annotators. Following the recommendation of Gretz et al. (2020), we use the weighted-average score, which considers the reliability of annotators, as the ground-truth score in our experiments. The annotated argument quality score of both datasets is a decimal from 0.00 to 1.00.

For IBM-Rank-30k, we follow the official data split in Gretz et al. (2020), using the data samples under 49, 7, 15 topics for training, validation, and testing, respectively. For IBM-ArgQ-5.3kArgs, as there is no official data split in Toledo et al. (2019),

Method	Base Model	IBM-ArgQ-5.3kArgs				IBM-Rank-30k			
		Pear.	Spear.	TAU	MAE ↓	Pear.	Spear.	TAU	MAE ↓
CL-DAGN <sup>†</sup>	BERT-base	-	-	-	-	0.5375	0.4949	-	-
CompAQA <sup>†</sup>	BERT-base	0.4563	0.4417	0.3064	0.1580	0.5282	0.4830	0.3390	0.1311
CL-DAGN <sup>†</sup>	RoBERTa-base	-	-	-	-	0.5604	0.5174	-	-
CompAQA <sup>†</sup>	RoBERTa-base	0.4681	0.4585	0.3165	0.1517	0.5642	0.5204	0.3670	0.1299
CompAQA <sup>†</sup>	DeBERTa-base	0.4657	0.4536	0.3127	0.1652	0.5797	0.5373	0.3794	0.1371
CompAQA	Llama-3.1-8B	0.5138	0.5059	0.3530	0.1458	0.6213	0.5791	0.4118	0.1216
Llama-Text-Score		0.4845	0.4777	0.3642	0.1509	0.6040	0.5660	0.4224	0.1275
Llama-Reg-EOS		0.4902	0.4804	0.3336	0.1490	0.6189	0.5716	0.4085	0.1206
MPAQ (Ours)		<b>0.5353*</b>	<b>0.5212*</b>	<b>0.3672*</b>	<b>0.1457</b>	<b>0.6406*</b>	<b>0.6001*</b>	<b>0.4280*</b>	<b>0.1196*</b>

Table 2: Main results for MPAQ on IBM-ArgQ-5.3kArgs and IBM-Rank-30k. “Pear.” and “Spear.” are short for “Pearson” and “Spearman”. For each evaluation metric, the best scores are highlighted in bold. <sup>†</sup> indicates that the results of this method are cited from the the original work. Note that Bao et al. (2024) implemented CompAQA on Llama-3-8B-Instruct. Here, for a fair comparison with our method, we replicate it based on Llama-3.1-8B. \* indicates significantly better ( $p < 0.05$ ) than CompAQA based on Llama-3.1-8B.

we adopt the data split used in Bao et al. (2024), which used the data samples under 14, 4, 4 topics for training, validation, and testing, respectively.

## 4.2 Evaluation Metrics

Following previous work (Gretz et al., 2020; Toledo et al., 2019; Wang et al., 2023b; Bao et al., 2024), we use Pearson and Spearman Correlation Coefficients as our primary evaluation metrics. We also adopt Kendall’s Tau (TAU), Mean Absolute Error (MAE) and Normalized Discounted Cumulative Gain (NDCG) as additional metrics. Due to space constraints, the NDCG results are reported in Table 12 (§ A.11).

## 4.3 Implementation Details

In the main experiments, the persona number  $N$  is set to 4. The advanced LLM we use to construct training data for both persona generation and rationale generation is GPT-4o-mini (gpt-4o-mini-2024-07-18).<sup>6</sup> MPAQ is fine-tuned on Llama-3.1-8B via LoRA.<sup>7</sup> We fix the batch size to 16, lora\_r to 32, lora\_alpha to 64, with all dropout rates set to 0.1. Greedy decoding is used for generation, with default hyper-parameters in the Transformers library.<sup>8</sup> The two stages of MPAQ are trained separately. In stage 1, we train the model for 2 epochs with a learning rate of  $1e-4$  on IBM-ArgQ-5.3kArgs and 3 epochs with a learning rate of  $1e-4$  on IBM-Rank-30k. In stage 2, we train the model for 2

epochs with a learning rate of  $8e-5$  on IBM-ArgQ-5.3kArgs and 2 epochs with a learning rate of  $2e-5$  on IBM-Rank-30k. These hyper-parameters are obtained through hyper-parameter search (See § A.15 for details). The regressor head is a multi-layer perceptron (MLP) with one hidden layer. All results in the main experiments are averaged across three independent runs with different random seeds.

## 4.4 Compared Methods

We compare MPAQ with the following baselines from prior work: (1) **CL-DAGN** utilizes contrastive learning and discourse knowledge to assess argument quality (Wang et al., 2023b); (2) **CompAQA** predicts argument quality scores through pairwise comparisons (Bao et al., 2024), and to the best of our knowledge, it represents the current state of the art. We further implement the following LLM-based baselines using Llama-3.1-8B: (3) **Llama-Text-Score** is trained to directly generates a decimal quality score in text format based on the input argument and topic; (4) **Llama-Reg-EOS** predicts a quality score by feeding the LLM-encoded representation of the input’s EOS (End of Sequence) token into a regression head, similar to Ma et al. (2024).

Furthermore, we conduct in-context learning experiments with GPT-4o-mini and Llama (see § A.14 for details).

## 4.5 Main Results

Table 2 presents the performance comparison between MPAQ and various baselines on IBM-ArgQ-5.3kArgs and IBM-Rank-30k datasets. The results demonstrate that MPAQ achieves consistent im-

<sup>6</sup>We also explore other advanced LLMs in § A.2 and find that our method maintains consistent superior performance.

<sup>7</sup>We also explore other open-source LLMs as base models in § A.3 and observe consistent strong performance.

<sup>8</sup><https://github.com/huggingface/transformers>

Method	IBM-ArgQ-5.3kArgs				IBM-Rank-30k			
	Pear.	Spear.	TAU	MAE ↓	Pear.	Spear.	TAU	MAE ↓
MPAQ (Ours)	<b>0.5353</b>	<b>0.5212</b>	<b>0.3672</b>	<b>0.1457</b>	<b>0.6406</b>	<b>0.6001</b>	<b>0.4280</b>	<b>0.1196</b>
<i>Persona Design</i>								
w/o Multi-persona	0.5078*	0.4974*	0.3473*	0.1463	0.6199*	0.5783*	0.4143*	0.1224*
w/o Dynamic Persona Generation	0.5329	0.5209	0.3658	0.1559*	0.6335*	0.5914*	0.4224*	0.1247*
w/o Analytical Perspectives	0.5254*	0.5128*	0.3583*	0.1471*	0.6393	0.5959*	0.4261	0.1202
w/o Strengths & Weaknesses	0.5223*	0.5098*	0.3556*	0.1662*	0.6281*	0.5875*	0.4190*	0.1201
<i>Scoring Strategy</i>								
w/o Coarse-to-fine Scoring	0.5309*	0.5197	0.3619*	0.1475*	0.6348*	0.5962	0.4259	0.1370*
<i>Architecture Design</i>								
w/o Two-stage Design	0.5145*	0.4980*	0.3483*	0.1911*	0.6238*	0.5794*	0.4146*	0.1754*
w/o Two LoRA Adapters	0.5155*	0.4983*	0.3477*	0.2097*	0.6313*	0.5902*	0.4212*	0.1284*

Table 3: Ablation study of MPAQ examining persona design, scoring strategy, and architecture design. \* indicates significantly worse ( $p < 0.05$ ) than MPAQ (Ours).

provements across all evaluation metrics on both datasets. Several key observations can be made.

First, methods based on Llama-3.1-8B consistently outperform baselines using BERT variants as backbones, mainly due to the significantly larger parameter scale of Llama-3.1-8B. Second, when comparing with methods sharing the same Llama-3.1-8B backbone, MPAQ demonstrates clear advantages across all metrics. Specifically, MPAQ outperforms CompAQA, which relies on pairwise comparisons between arguments, suggesting that simulating multiple perspectives provides a more comprehensive assessment approach than pairwise comparison. The performance gap between MPAQ and Llama-Text-Score baseline highlights the difficulty of directly predicting fine-grained decimal scores through text generation, demonstrating the necessity of our coarse-to-fine scoring strategy. Moreover, compared to the Llama-Reg-EOS baseline which provides a single score without explanation, MPAQ’s multi-persona framework not only achieves better performance but also offers interpretable rationales from diverse perspectives to support its quality assessment.

We further show that these performance advantages remain robust in low-resource settings (§ A.4) and in in-topic setting (§ A.5). We also demonstrate that the significant performance improvements are due to the personas themselves rather than merely optimizing more variance (§ A.7). A qualitative case study of MPAQ is shown in Appendix D.

#### 4.6 Ablation Study

To validate the effectiveness of different components in MPAQ, we conduct ablation studies on

both datasets, as shown in Table 3.

**Impact of Persona Design.** First, removing the multi-persona framework while retaining the coarse-to-fine scoring strategy (“w/o Multi-persona”) leads to substantial performance drops on both datasets, highlighting the importance of incorporating multiple perspectives. “w/o Dynamic Persona Generation”, which removes Stage 1 entirely and uses fixed personas instead, shows performance degradation, validating the effectiveness of our persona generation stage. Additionally, removing analytical perspectives in Stage 1 (“w/o Analytical Perspectives”) or replacing the structured strengths and weaknesses analysis with unrestricted analysis in Stage 2 (“w/o Strengths & Weaknesses”) both lead to decreased performance, indicating that these detailed specifications help guide more accurate quality assessment.

**Impact of Scoring Strategy.** “w/o Coarse-to-fine Scoring”, which replaces the coarse-grained score with a static token before fine-grained prediction, shows performance decreases across both datasets. This demonstrates the effectiveness of our hierarchical scoring strategy.

**Impact of Architecture Design.** Both architectural variants show decreased performance: “w/o Two-stage Design”, which concatenates persona generation and quality assessment into a single inference pipeline, and “w/o Two LoRA Adapters”, which employs a shared adapter for both stages. These results validate our design decision to maintain separate stages with dedicated LoRA adapters, allowing for more focused parameter optimization for each specific task.

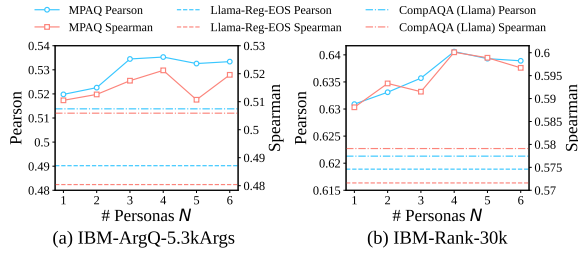


Figure 3: Performance comparison of MPAQ with various persona counts ( $N$ ) on two datasets measured by Pearson and Spearman correlations. The dashed lines represent baselines (Llama-Reg-EOS and CompAQA).

## 5 Discussions

**Impact of the Persona Number  $N$ .** Figure 3 shows the impact of  $N$  on the IBM-ArgQ-5.3kArgs and IBM-Rank-30k datasets. We observe that both Pearson and Spearman correlation coefficients tend to improve as we increase the number of personas, with optimal results achieved at  $N = 4$ . This suggests that incorporating multiple perspectives indeed enhances the quality assessment capability of our framework. Furthermore, across all tested values of  $N$ , MPAQ consistently outperforms both Llama-Reg-EOS and CompAQA baselines (shown as dashed lines), validating the effectiveness of our multi-persona approach regardless of the specific number of personas used. Considering that both training and inference computational costs increase substantially with larger  $N$ , we only experiment up to  $N = 6$  and choose  $N = 4$  as the optimal setting for our main experiments, balancing performance gains and computational efficiency.

**Impact of Removing Personas** To verify that each persona contributes meaningfully to the final quality assessment, we conduct an ablation experiment on IBM-Rank-30k dataset by randomly removing persona-specific scores during the score aggregation process. This experiment aims to demonstrate that our strong performance stems from the collective contribution of multiple perspectives, rather than being driven by a small subset of personas. Specifically, when calculating the final quality score, instead of using all  $N$  personas, we randomly exclude  $N_{minus}$  personas. Figure 4 shows how this progressive removal of personas affects model performance.

The results demonstrate a clear pattern: as more personas are removed from the aggregation process (i.e., increasing  $N_{minus}$ ), both Pearson and Spearman correlation coefficients show a consis-

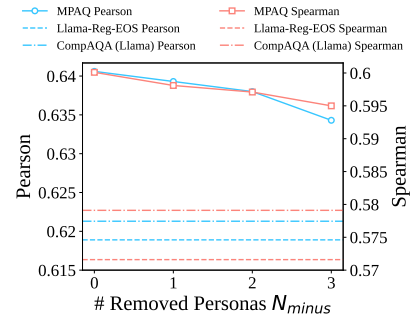


Figure 4: The trends of Pearson and Spearman correlation coefficients on IBM-Rank-30k when removing  $N_{minus}$  personas during score aggregation.

tent decline. This gradual performance degradation indicates that each persona contributes valuable perspective to the final assessment, and the high performance of MPAQ cannot be maintained with just a subset of personas. However, it’s noteworthy that even with several personas removed, MPAQ maintains performance levels above the baseline methods (Llama-Reg-EOS and CompAQA), demonstrating the inherent strength of our multi-persona framework. These findings validate our core hypothesis that argument quality assessment benefits from aggregating multiple diverse perspectives.

### Quality Analysis of Generated Personas and Rationales.

Here we evaluate how well MPAQ has learned to generate personas and rationales. Since we use GPT-4o-mini to construct the training data for both persona and rationale generation, GPT-4o-mini serves as the teacher model of MPAQ, making it an ideal reference point for analyzing the generation quality. To assess whether MPAQ successfully learns the generation capabilities from its teacher model, we randomly sample 50 test samples from the IBM-Rank-30k dataset and perform pairwise comparison evaluation between the outputs of MPAQ and GPT-4o-mini by both human evaluators and GPT-4o (gpt-4o-2024-11-20). See § A.12 for detailed experimental settings.

Figure 5 (a) and (b) present the evaluation results for persona generation and rationale generation, respectively. For persona generation, Figure 5 (a) shows comparable performance between MPAQ and its teacher model GPT-4o-mini, indicating that MPAQ has successfully learned the persona generation capability. The comparison of rationale generation (Figure 5 (b)) reveals more promising results, with MPAQ demonstrating slightly better performance than GPT-4o-mini. Moreover, a detailed



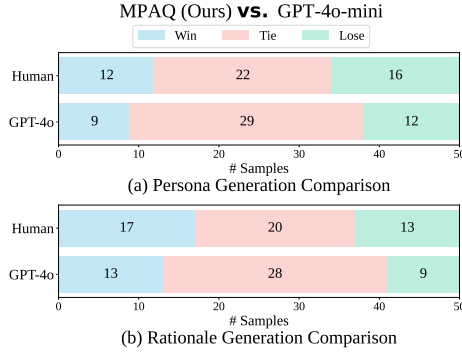


Figure 5: Generation quality comparison of MPAQ and GPT-4o-mini evaluated by human and GPT-4o.

analysis of the distribution and diversity of personas generated by MPAQ can be found in § A.1.

While Figure 5 presents the relative generation quality between MPAQ and the teacher model (GPT-4o-mini), it does not reveal the absolute generation quality. To address this limitation, we conduct a comprehensive human evaluation to directly assess the generation quality of GPT-4o-mini. We employ the same 50 data instances used in the pairwise human evaluation described in § A.12. Two annotators, both graduate students specializing in NLP, independently evaluate each persona and rationale generated by GPT-4o-mini on a scale of 1 to 5. The final scores are calculated by averaging the ratings from both annotators. The evaluation results demonstrate that the personas and rationales generated by GPT-4o-mini achieve average scores of 4.78 and 4.66, respectively. These high scores indicate that GPT-4o-mini produces reliable personas and rationales that can effectively analyze and evaluate argument quality from diverse perspectives. Additional details are provided in § A.13.

**Performance Analysis of Different Quality Score Intervals.** To better understand how our model handles arguments of different quality ranges, we divide the IBM-Rank-30k test set according to ground truth quality scores into four intervals and compare the Pearson and Spearman correlation coefficients of different methods within each interval. As shown in Figure 6, MPAQ consistently outperforms both Llama-Reg-EOS and CompAQA(Llama) across all quality intervals, indicating that our framework provides robust benefits regardless of the argument’s quality level. Notably, all three models show relatively weaker performance in evaluating arguments with lower quality scores ([0.00, 0.25) and [0.25, 0.50)). This perfor-

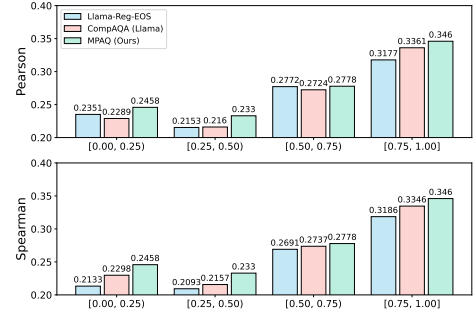


Figure 6: Performance analysis on IBM-Rank-30k test set across different quality score intervals, measured by Pearson (top) and Spearman (bottom) correlations.

mance pattern can be attributed to the imbalanced distribution of the training data (Table 13). Despite this data limitation, MPAQ’s performance advantage over baseline methods is more pronounced when assessing lower-quality arguments, highlighting its superior robustness.

## 6 Conclusion

In this paper, we propose a multi-persona framework for argument quality assessment (MPAQ). It simulates diverse evaluator perspectives through LLMs to model the inherent subjectivity in argument quality assessment. Extensive experiments on two big datasets (IBM-Rank-30k and IBM-ArgQ-5.3kArgs) demonstrate MPAQ’s superior performance compared to several strong baselines.

## Limitations

While our method demonstrates superior performance, it incurs additional computational overhead due to the increased data volume from our two-stage process and multi-perspective evaluation framework. Nevertheless, we consider this computational trade-off worthwhile given the significant performance improvements achieved.

Furthermore, since our method requires a reasonable amount of training data, we evaluate our model on the two largest argument quality assessment datasets. Adapting our framework to medium-sized argument quality datasets on different quality dimensions is left for future work.

Finally, due to the absence of annotator profile information in argument quality assessment datasets, we could not directly evaluate the generated personas and their corresponding argument quality scores. Future work should compare the generated persona information with real-world data.

## Ethics Statement

In all our experiments, we utilize existing datasets that have been widely used in previous scientific publications. These datasets are publicly available under the Creative Commons BY-SA 4.0 License and do not contain any personally identifiable information of offensive content. While our study do not uncover harmful outputs from ChatGPT, prior research has reported instances of unexpected generations. We advise users to carefully assess the ethical considerations of generated content, exercise caution when applying the system in practical scenarios, and encourage further research into ChatGPT's responses to specific prompts that might lead to the dissemination of problematic information. The use of the Llama models complies with the Meta Llama 3 License<sup>9</sup>. All other artifacts are used under their respective licenses, following their intended purposes.

## Acknowledgements

This work was partially supported by the National Natural Science Foundation of China 62176076, Natural Science Foundation of Guang Dong 2023A1515012922, the Shenzhen Foundational Research Funding JCYJ20220818102415032, the Major Key Project of PCL2023A09, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005 and CIPSC-SMP-ZHIPU Large Model Cross-Disciplinary Fund ZPCG20241119405.

## References

- Aristotle and George A. Kennedy. 1991. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press (OUP) USA.
- Jianzhu Bao, Bojun Jin, Yang Sun, Yice Zhang, Yuhang He, and Ruifeng Xu. 2024. A comparison-based framework for argument quality assessment. *Electronics*, 13(20):4088.
- Darshan Deshpande, Zhivar Sourati, Filip Ilievski, and Fred Morstatter. 2024. [Contextualizing argument quality assessment with relevant knowledge](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 316–326. Association for Computational Linguistics.
- Yuning Ding, Marie Bexte, and Andrea Horbach. 2023. [Score it all together: A multi-task learning study on automatic scoring of argumentative essays](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13052–13063. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Han-nah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. [The Llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Frans H. Van Eemeren and Rob Grootendorst. 2003. *A Systematic Theory of Argumentation: The Pragmatic-Dialectical Approach*. Cambridge University Press.
- Mohamed Elaraby, Diane J. Litman, Xiang Li, and Ahmed Magooda. 2024. [Persuasiveness of generated free-text rationales in subjective decisions: A case study on pairwise argument ranking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14311–14329. Association for Computational Linguistics.
- Charles-Olivier Favreau, Amal Zouaq, and Sameer Bhatnagar. 2022. [Learning to rank with BERT for argument quality evaluation](#). In *Proceedings of the Thirty-Fifth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2022, Hutchinson Island, Jensen Beach, Florida, USA, May 15-18, 2022*.
- Michael Fromm, Max Berrendorf, Evgeniy Faerman, and Thomas Seidl. 2023. [Cross-domain argument](#)

<sup>9</sup><https://www.llama.com/llama3/license>

- quality estimation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13435–13448. Association for Computational Linguistics.
- Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. [Efficient pairwise annotation of argument quality](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5772–5781. Association for Computational Linguistics.
- Shai Gretz, Roni Friedman, Edo Cohen-Karlik, As-saf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. [A large-scale dataset for argument quality ranking: Construction and analysis](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7805–7813. AAAI Press.
- Charles L. Hamblin. 1970. Fallacies. *Tijdschrift Voor Filosofie*, 33(1):183–188.
- Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. [Architectural sweet spots for modeling human label variation by the example of argument quality: It’s best to relate perspectives!](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11138–11154. Association for Computational Linguistics.
- Zhe Hu, Hou Pong Chan, and Yu Yin. 2024. [AMERICANO: argument generation with discourse-driven decomposition and agent interaction](#). In *Proceedings of the 17th International Natural Language Generation Conference, INLG 2024, Tokyo, Japan, September 23 - 27, 2024*, pages 82–102. Association for Computational Linguistics.
- Ralph Henry Johnson and J. Anthony Blair. 1977. *Logical Self-Defense*. McGraw-Hill, Toronto, Canada.
- Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. [Arganalysis35k : A large-scale dataset for argument quality analysis](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13916–13931. Association for Computational Linguistics.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. [Better zero-shot reasoning with role-play prompting](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4099–4113. Association for Computational Linguistics.
- Anne Lauscher, Lily Ng, Courtney Napoles, and Joel R. Tetreault. 2020. [Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4563–4574. International Committee on Computational Linguistics.
- Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. [Distilling ChatGPT for explainable automated student answer assessment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6007–6026. Association for Computational Linguistics.
- Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. [Argue with me tersely: Towards sentence-level counter-argument generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16705–16720. Association for Computational Linguistics.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. [On LLMs-driven synthetic data generation, curation, and evaluation: A survey](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11065–11082. Association for Computational Linguistics.
- Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. [Fine-tuning LLaMA for multi-stage text retrieval](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2421–2425. ACM.
- Santiago Marro, Elena Cabrio, and Serena Villata. 2022. [Graph embeddings for argumentation quality assessment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4154–4164. Association for Computational Linguistics.
- Huy V. Nguyen and Diane J. Litman. 2018. [Argument mining for improving the automated scoring of persuasive essays](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5892–5899. AAAI Press.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Isaac Persing and Vincent Ng. 2013. [Modeling thesis clarity in student essays](#). In *Proceedings of the 51st*



- Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 260–269. The Association for Computer Linguistics.
- Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. [Can language models recognize convincing arguments?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 8826–8837. Association for Computational Linguistics.
- Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Rutu Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. [An autonomous debating system](#). *Nat.*, 591(7850):379–384.
- Christian Stab and Iryna Gurevych. 2014. [Annotating argument components and relations in persuasive essays](#). In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1501–1510. Association for Computational Linguistics.
- Christian Stab and Iryna Gurevych. 2017. [Recognizing insufficiently supported arguments in argumentative essays](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 980–990. Association for Computational Linguistics.
- Reid Swanson, Brian Ecker, and Marilyn A. Walker. 2015. [Argument mining: Extracting arguments from online dialogue](#). In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 217–226. The Association for Computer Linguistics.
- Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. [Automatic argument quality assessment - new datasets and methods](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5624–5634. Association for Computational Linguistics.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two tales of persona in llms: A survey of role-playing and personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 16612–16631. Association for Computational Linguistics.
- Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. [Using argument mining to assess the argumentation quality of essays](#). In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1680–1691. Association for Computational Linguistics.
- Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. [Argument quality assessment in the age of instruction-following large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1519–1538. ELRA and ICCL.
- Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. [Computational argumentation quality assessment in natural language](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 176–187. Association for Computational Linguistics.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. [Is ChatGPT a good NLG evaluator? a preliminary study](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.
- Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. 2024. [A survey on data synthesis and augmentation for large language models](#). *CoRR*, abs/2410.12896.
- Yiran Wang, Xuanang Chen, Ben He, and Le Sun. 2023b. [Contextual interaction for argument post quality assessment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10420–10432. Association for Computational Linguistics.
- Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. [Large language models are diverse role-players for summarization evaluation](#). In



Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. 2024. **QA-LoRA: Quantization-aware low-rank adaptation of large language models**. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

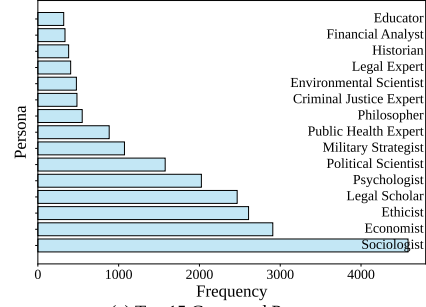
## A Additional Experiments

### A.1 Personas Distribution and Diversity Analysis

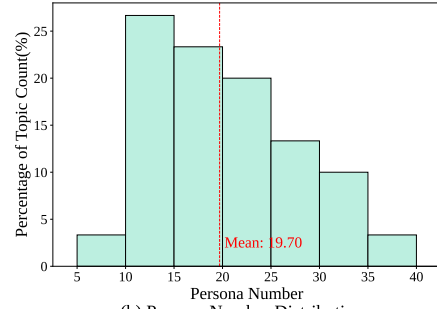
We analyze the personas generated by MPAQ during Stage 1 for both their distribution and diversity. This analysis is conducted on the test set of IBM-Rank-30k. Figure 7 (a) presents the top 15 most commonly generated personas. The results reveal that sociologists, economists, ethicists, and legal scholars emerge as the most frequently generated personas, reflecting the key dimensions that need to be considered for the test set of this dataset. Figure 7 (b) further illustrates the percentage distribution of unique personas per topic, with an average of 19.70 personas per topic. Notably, more than 95% of topics are associated with 10 or more distinct personas, demonstrating the MPAQ’s ability to generate diverse and relevant persons for argument quality assessment.

### A.2 Impact of Different Advanced LLMs for Training Data Construction

Since we use an advanced LLM to construct the training data for both persona and rationale generation, this advanced LLM can be viewed as a teacher model for training MPAQ. Here, we explore the performance of MPAQ on IBM-Rank-30k dataset using different teacher models for training data construction. Besides GPT-4o-mini used in our main experiments, we also experiment with GPT-3.5-Turbo and Llama-3.1-70B-Instruct as teacher models. The results are listed in Table 4. Among the teacher models evaluated, GPT-4o-mini and Llama-3.1-70B-Instruct demonstrate comparable performance across all evaluation metrics, while GPT-3.5-Turbo shows relatively weaker performance. Notably, MPAQ consistently outperforms the baseline methods regardless of the teacher model used, demonstrating the effectiveness and robustness of our approach.



(a) Top 15 Generated Personas



(b) Persona Number Distribution

Figure 7: Personas distribution and diversity analysis: (a) The top 15 most commonly generated personas; (b) Distribution of the number of unique personas generated per topic (mean = 19.70).

### A.3 Impact of Different Base Models

To evaluate the generalizability of our method across different base models, we conduct experiments on IBM-Rank-30k dataset by replacing the base model, Llama-3.1-8B, used in our main experiments with several alternatives: Llama-3-8B, Llama-2-7b, and Mistral-7B-v0.3. For a fair comparison, we implement both baseline methods (Llama-Reg-EOS and CompAQA) using the same base models. The results are presented in Table 5.

The experimental results demonstrate that, overall, MPAQ outperforms both baseline methods regardless of the base model used. Specifically, Llama-3-8B achieves comparable performance to Llama-3.1-8B, while Llama-2-7b and Mistral-7B-v0.3 show slightly weaker performance. These results demonstrate the robustness and generalizability of our method across different base models.

### A.4 Performance Analysis in Low-resource Settings

Since collecting high-quality annotations for argument quality assessment is costly, it is crucial to evaluate model performance under low-resource conditions. We randomly select 10%, 20%, 30%, and 40% of the topics from IBM-Rank-30k’s training set and use their associated samples as training

Method	Teacher Model	IBM-Rank-30k			
		Pear.	Spear.	TAU	MAE ↓
CompAQA (Llama)	-	0.6213	0.5791	0.4118	0.1216
Llama-Reg-EOS	-	0.6189	0.5716	0.4085	0.1206
MPAQ (Ours)	GPT-4o-mini	0.6406	0.6001	0.4280	<b>0.1196</b>
	GPT-3.5-Turbo	0.6408	0.5963	0.4259	0.1261
	Llama-3.1-70B-Instruct	<b>0.6427</b>	<b>0.6005</b>	<b>0.4297</b>	0.1206

Table 4: Experimental results of using different advanced LLMs for training data construction.

Base Model	Method	IBM-Rank-30k			
		Pear.	Spear.	TAU	MAE ↓
Llama-3.1-8B	Llama-Reg-EOS	0.6189	0.5716	0.4085	0.1206
	CompAQA (Llama)	0.6213	0.5791	0.4118	0.1216
	MPAQ (Ours)	<b>0.6406</b>	<b>0.6001</b>	<b>0.4280</b>	<b>0.1196</b>
Llama-3-8B	Llama-Reg-EOS	0.6128	0.5663	0.4047	0.1228
	CompAQA (Llama)	0.6216	0.5815	0.4138	0.1221
	MPAQ (Ours)	<b>0.6411</b>	<b>0.6017</b>	<b>0.4304</b>	<b>0.1202</b>
Llama-2-7b	Llama-Reg-EOS	0.6193	0.5706	0.4076	<b>0.1211</b>
	CompAQA (Llama)	0.6185	0.5736	0.4079	0.1217
	MPAQ (Ours)	<b>0.6366</b>	<b>0.5940</b>	<b>0.4238</b>	0.1358
Mistral-7B-v0.3	Llama-Reg-EOS	0.6072	0.5605	0.3993	0.1294
	CompAQA (Llama)	0.6088	0.5606	0.3992	0.1304
	MPAQ (Ours)	<b>0.6343</b>	<b>0.5864</b>	<b>0.4187</b>	<b>0.1205</b>

Table 5: Experimental results of using different base models.

data for low-resource experiments. We evaluate the model on the test set and compare its performance with Llama-Reg-EOS and CompAQA trained under the same training data conditions. The results are presented in Table 6.

The results demonstrate our model’s strong performance under low-resource conditions. MPAQ consistently outperforms both baseline methods across almost all evaluation metrics at different training data sizes. As training data increases, the performance of MPAQ steadily improves while maintaining its advantage over the baselines.

### A.5 Cross-topic vs. In-topic Experiment

In our main experiments, we follow the official data split of IBM-Rank-30k, which represents a cross-topic setting where the topics in validation and test sets have no overlap with those in the training set. To further evaluate MPAQ’s capabilities, we also conduct experiments in an in-topic setting by randomly splitting the entire dataset, allowing arguments from the same topics to appear in both training and testing.

As shown in Table 10, while all methods demonstrate relatively stronger performance in the in-topic setting, MPAQ consistently outperforms the

baselines across most metrics in both settings.

### A.6 Impact of Score Positions

We investigate whether the position of the coarse-grained quality score relative to the rationales affects model performance. Following previous work (Li et al., 2023; Wang et al., 2023a), our main experiment places the score before the rationales (Table 1). We also experiment with positioning it after the rationales. The prompt template is shown in Table 18.

Results in Table 7 show that placing the score after the rationales leads to slightly lower performance across most metrics, with marginal improvements only in MAE. These findings validate our choice of score-first placement in the main experiment.

### A.7 Comparison with Inferring for Multiple Times

To demonstrate the effectiveness of personas themselves rather than merely optimizing more variance, we conduct an additional experiment. Specifically, during training, we do not consider any persona-related designs, allowing the LLM to generate rationales and corresponding coarse-grained scores

Training Data	Method	IBM-Rank-30k			
		Pear.	Spear.	TAU	MAE ↓
10%	Llama-Reg-EOS	0.5823	0.5499	0.3909	0.2046
	CompAQA (Llama)	0.5627	0.5298	0.3729	0.1359
	MPAQ (Ours)	<b>0.6033</b>	<b>0.5562</b>	<b>0.3948</b>	<b>0.1232</b>
20%	Llama-Reg-EOS	0.5974	0.5533	0.3947	0.1441
	CompAQA (Llama)	0.5955	0.5554	0.3933	<b>0.1240</b>
	MPAQ (Ours)	<b>0.6215</b>	<b>0.5733</b>	<b>0.4078</b>	0.1285
30%	Llama-Reg-EOS	0.5908	0.5519	0.3921	0.1492
	CompAQA (Llama)	0.5887	0.5552	0.3941	<b>0.1276</b>
	MPAQ (Ours)	<b>0.6276</b>	<b>0.5838</b>	<b>0.4156</b>	0.1347
40%	Llama-Reg-EOS	0.6032	0.5631	0.4003	0.1233
	CompAQA (Llama)	0.5999	0.5606	0.3980	0.1247
	MPAQ (Ours)	<b>0.6298</b>	<b>0.5906</b>	<b>0.4210</b>	<b>0.1202</b>
Full	Llama-Reg-EOS	0.6189	0.5716	0.4085	0.1206
	CompAQA (Llama)	0.6213	0.5791	0.4118	0.1216
	MPAQ (Ours)	<b>0.6406</b>	<b>0.6001</b>	<b>0.4280</b>	<b>0.1196</b>

Table 6: Experimental results in low-resource settings.

Method	IBM-Rank-30k			
	Pear.	Spear.	TAU	MAE ↓
Score before Rationales	<b>0.6406</b>	<b>0.6001</b>	<b>0.4280</b>	0.1196
Score after Rationales	0.6359	0.5971	0.4273	<b>0.1187</b>

Table 7: Comparison of score positions.

based on the input argument, and then output fine-grained scores using our proposed "Coarse-to-fine Scoring Strategy." During inference, we sample rationales using 4 different seeds (consistent with the number of personas in our main experiment) under the top\_p decoding strategy and predict the corresponding fine-grained scores, then we average the 4 scores obtained from the 4 inferences as the final score. The results are shown in Table 8.

As shown in the table, there remains a significant performance gap between the multiple inference method and our proposed approach, demonstrating the effectiveness of the personas themselves.

### A.8 Alternative Supervision Strategy for Quality Score Prediction

We explore an alternative supervision approach that applies the ground truth score at the aggregated level rather than the persona level. Unlike our indirect supervision approach where each persona’s prediction is supervised by the ground truth aggregated score, this strategy only computes loss after averaging predictions from all personas. Specifically, during training, we first aggregate the predicted scores from all personas to obtain an average score, then calculate the MAE loss between this

aggregated predicted score and the ground truth score.

Results in Table 9 show that this alternative supervision approach underperforms our persona-level indirect supervision approach across all metrics. This validates our design choice of indirect supervision.

### A.9 Score Aggregation Strategies

We explore both statistical-based and confidence-based strategies for aggregating persona-specific scores. For statistical-based approaches, we evaluate three methods to aggregate persona-specific scores: arithmetic mean (our approach in the main experiment), median, and geometric mean. We also explore confidence-based aggregation where each persona’s score is weighted by its generation confidence. We derive three types of weights from model’s output probabilities: (1) Stage 1 weight by averaging the token generation probabilities in Stage 1, (2) Stage 2 weight from token probabilities in Stage 2 output, and (3) Combined weight that average weights from both stages.

As shown in Table 11, all strategies achieve comparable performance. Given these minimal differences, we adopt the arithmetic mean in our main

Method	IBM-Rank-30k			
	Pear.	Spear.	TAU	MAE ↓
MPAQ (Ours)	<b>0.6406</b>	<b>0.6001</b>	<b>0.4280</b>	<b>0.1196</b>
Multiple Inferences	0.6291	0.5828	0.4155	0.1284

Table 8: Comparison between multi-persona framework and merely inferring for multiple times.

Method	IBM-Rank-30k			
	Pear.	Spear.	TAU	MAE ↓
Persona Level (Indirect Supervision)	<b>0.6406</b>	<b>0.6001</b>	<b>0.4280</b>	<b>0.1196</b>
Aggregated Level	0.6296	0.5829	0.4159	0.1202

Table 9: Experiment results of an alternative supervision strategy for quality score prediction.

Setting	Method	IBM-Rank-30k			
		Pear.	Spear.	TAU	MAE ↓
Cross-topic	Llama-Reg-EOS	0.6189	0.5716	0.4085	0.1206
	CompAQA (Llama)	0.6213	0.5791	0.4118	0.1216
	MPAQ (Ours)	<b>0.6406</b>	<b>0.6001</b>	<b>0.4280</b>	<b>0.1196</b>
In-topic	Llama-Reg-EOS	0.6470	0.5983	0.4293	<b>0.1161</b>
	CompAQA (Llama)	0.6380	0.5850	0.4178	0.1179
	MPAQ (Ours)	<b>0.6644</b>	<b>0.6136</b>	<b>0.4406</b>	0.1180

Table 10: Experimental results in cross-topic and in-topic settings.

experiments due to its simplicity.

#### A.10 Analysis of Predicted Score Differentiation Across Personas

To validate that our model predicts meaningfully differentiated scores across personas despite using the same supervision signal during training, we conduct a validation experiment on the IBM-Rank-30k dataset. Based on our main experimental setting ( $N = 4$ ), we design a setup where two personas focus solely on analyzing strengths while predicting scores, and the other two focus exclusively on weaknesses. We then compute the maximum predicted score among strength-focused personas and the minimum predicted score among weakness-focused personas, and calculate the difference between these two values to quantify the scoring differentiation between personas.

Figure 8 presents the distribution of these score differences. The majority of the score differences fall within the range of 0.025 to 0.05, with some cases showing larger differences exceeding 0.05. These results demonstrate that, despite using the same supervision signal during training, our model can generate meaningfully differentiated scores by incorporating persona information and rationales.

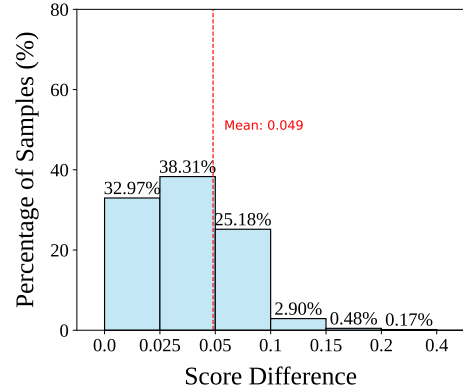


Figure 8: The distribution of predicted score differentiation across personas.

#### A.11 Additional Main Results

Following the previous work (Favreau et al., 2022; Wang et al., 2023b; Bao et al., 2024), we also report the performance of our method on Normalized Discounted Cumulative Gain (NDCG) metric. The results are shown in Table 12.

MPAQ achieves better performance on NDCG@50 and NDCG@200 across both datasets. However, for top-k rankings where  $k \leq 15$ , RoBERTa-based models show stronger performance, particularly on IBM-Rank-30k. This suggests that while MPAQ excels at overall ranking



Method	IBM-Rank-30k			
	Pear.	Spear.	TAU	MAE ↓
<i>Statistical-based</i>				
Arithmetic Mean (Ours)	0.6406	0.6001	0.4280	<b>0.1196</b>
Median	0.6400	0.5995	<b>0.4283</b>	0.1201
Geometric Mean	0.6407	0.6000	0.4282	0.1199
<i>Confidence-based</i>				
Stage 1 weight	0.6405	<b>0.6003</b>	0.4281	0.1200
Stage 2 weight	0.6407	0.6001	0.4279	0.1198
Stage 1+2 weight	<b>0.6408</b>	0.6002	0.4280	0.1197

Table 11: Performance comparison of different score aggregation strategies.

quality, smaller models may be more effective at identifying the highest-quality arguments. Given that NDCG@5, NDCG@10, and NDCG@15 only evaluate performance on a very small subset of samples, we consider NDCG@50 and NDCG@200 to be more representative of overall model capabilities.

#### A.12 Experimental Settings for Human and GPT Evaluations

To evaluate the generation quality of MPAQ compared to its teacher model GPT-4o-mini, we conduct a comprehensive evaluation using both human annotators and GPT-4o (gpt-4o-2024-11-20). Our evaluation focuses on both persona generation and rationale generation capabilities.

**Evaluation Data Preparation** We randomly sample 50 test examples from IBM-Rank-30k dataset. For each example, we obtain two sets of generations: one from MPAQ and one from GPT-4o-mini (gpt-4o-mini-2024-07-18) using identical prompts from Tables 15 and 17. These generations are paired to create 50 comparison pairs, with the order of MPAQ and GPT-4o-mini outputs randomized within each pair.

**Evaluation Process** We employ two independent evaluation approaches:

- **Human Evaluation:** Two annotators, graduate students in the field of artificial intelligence, independently assess each comparison pair to determine which system produces better personas and rationales. Any disagreements between annotators are resolved through discussion sessions. The interface is shown in Figure 9.
- **GPT-4o Evaluation:** We use GPT-4o (gpt-4o-2024-11-20) with a carefully designed evaluation prompt (Table 19) to judge which system

generates better personas and rationales for each comparison pair.

#### A.13 Generation Quality of Teacher Model

Since we use the personas and rationales generated by GPT-4o-mini, the generation quality of GPT-4o-mini is significant for training MPAQ. Therefore, we conduct a human evaluation of the personas and rationales generated by GPT-4o-mini. Specifically, we use the same 50 data instances that are used for the pairwise human evaluation in § A.12. Each persona and rationale generated by the teacher model, GPT-4o-mini, is scored on a scale of 1 to 5 by two annotators (both are graduate students in the field of NLP), and the final score is obtained by averaging their ratings. The scoring criteria are as follows:

- **Score 1:** None of the generated personas/rationales are reasonable and relevant to the argument/argument and persona.
- **Score 2:** Only one of the generated persona/rationale is reasonable and relevant to the argument/argument and persona, none of the others are reasonable and relevant.
- **Score 3:** Two of the generated personas/rationales are reasonable and relevant to the argument/argument and persona, none of the others are reasonable and relevant.
- **Score 4:** Three of the generated personas/rationales are reasonable and relevant to the argument/argument and persona, the other one is unreasonable or irrelevant.
- **Score 5:** All of the generated personas/rationales are reasonable and relevant to the argument/argument and persona.

Method	Base Model	IBM-ArgQ-5.3kArgs					IBM-Rank-30k				
		N.@5	N.@10	N.@15	N.@50	N.@200	N.@5	N.@10	N.@15	N.@50	N.@200
CL-DAGN <sup>†</sup>	BERT-base	-	-	-	-	-	0.9330	0.9372	0.9388	-	-
CompAQA <sup>†</sup>	BERT-base	-	-	0.8097	-	-	-	-	0.9635	-	-
CL-DAGN <sup>†</sup>	RoBERTa-base	-	-	-	-	-	<b>0.9799</b>	<b>0.9769</b>	<b>0.9648</b>	-	-
CompAQA <sup>†</sup>	RoBERTa-base	-	-	0.7630	-	-	-	-	0.9543	-	-
CompAQA <sup>†</sup>	DeBERTa-base	-	-	0.7352	-	-	-	-	0.9500	-	-
CompAQA	Llama-3.1-8B	0.6903	0.6921	0.6980	0.7324	0.8219	0.9418	0.9479	0.9501	0.9419	0.9441
Llama-Text-Score		0.5052	0.5898	0.6250	0.7165	0.8027	0.9365	0.9376	0.9383	0.9398	0.9364
Llama-Reg-EOS		<b>0.7670</b>	<b>0.7894</b>	0.7733	0.7633	0.8184	0.9376	0.9325	0.9322	0.9454	0.9360
MPAQ (Ours)		0.7447	0.7663	<b>0.7868</b>	<b>0.7979</b>	<b>0.8439</b>	0.9354	0.9434	0.9502	<b>0.9486</b>	<b>0.9492</b>

Table 12: Additional main results for MPAQ on IBM-ArgQ-5.3kArgs and IBM-Rank-30k on other evaluation metrics.

If one persona is duplicated with another persona, it is also considered unreasonable.

The results show that the personas and rationales generated by GPT-4o-mini receive average scores of 4.78 and 4.66, respectively. This indicates that the personas and rationales generated by GPT-4o-mini have high reliability and can effectively analyze and evaluate the quality of arguments from diverse perspectives.

#### A.14 In-context Learning Experiments

We conduct in-context learning (ICL) experiments using three different LLMs: GPT-4o-mini, Llama-3.1-70B-Instruct, and Llama-3.1-8B-Instruct. For each model, we evaluate its performance with varying numbers of demonstrations (0-shot, 1-shot, 2-shot, and 4-shot) on both datasets. The results are shown in Table 14.

We observe that increasing the number of demonstrations does not lead to clear improvements in the models’ ICL performance on argument quality assessment. We hypothesize this is due to the subjective nature of the task, making it difficult for models to learn effective evaluation abilities from a limited number of demonstrations. Notably, these ICL results substantially underperform our fine-tuned approach, highlighting the importance of task-specific training for argument quality assessment.

#### A.15 More Implementation Details for the Main Experiments

During training, we conduct a random hyper-parameter search. For 2 stages on both IBM-Rank-30k and IBM-ArgQ-5.3kArgs datasets, we set the learning rate search range to  $[2e-5, 4e-5, 8e-5, 1e-4, 2e-4]$  and the epoch search range to

$[1, 2, 3]$ . Ultimately, for Stage 1, we select the hyper-parameters that minimize the auto-regressive loss on the validation set. For Stage 2, we choose the hyper-parameters that maximize the Pearson correlation coefficient on the validation set.

In this work, GPT series models are accessed through the official API provided by OpenAI.<sup>10</sup> Other open-source models are obtained from Huggingface.<sup>11</sup>

Pearson, Spearman correlation coefficients and Kendall’s Tau evaluation metrics are calculated using the python SciPy library, while the other metrics are evaluated using the python scikit-learn library.

## B Datasets Information

Statistics of sample quantity in different quality score intervals are shown in Table 13.

Quality interval	IBM-Rank-30k		
	Train	Validation	Test
0.00-0.25	235	53	112
0.25-0.50	1636	231	551
0.50-0.75	5306	774	1629
0.75-1.00	13797	2150	4023
Full	20974	3208	6315

Table 13: Statistics of sample quantity in different quality score intervals of IBM-Rank-30k dataset.

## C Prompts

In this Appendix section, we demonstrate the prompt templates we use in our experiments.

<sup>10</sup><https://openai.com/>

<sup>11</sup><https://huggingface.co/>

Model	Demo Count	IBM-ArgQ-5.3kArgs				IBM-Rank-30k			
		Pear.	Spear.	TAU	MAE ↓	Pear.	Spear.	TAU	MAE ↓
GPT-4o-mini	0-Shot	0.3625	0.3696	0.2749	0.1921	0.4049	0.3583	0.2613	0.2197
	1-Shot	0.3719	0.3786	0.2659	0.2040	0.3952	0.3687	0.2616	0.2327
	2-Shot	0.4054	0.4067	0.2827	0.2004	0.3911	0.3627	0.2547	0.2308
	4-Shot	0.4009	0.3992	0.2820	0.1991	0.4159	0.3861	0.2710	0.2063
Llama-3.1-70B-Instruct	0-Shot	0.4257	0.4069	0.2986	0.1782	0.4589	0.4365	0.3261	0.3362
	1-Shot	0.4046	0.3659	0.2590	0.1818	0.4574	0.4146	0.2949	0.2114
	2-Shot	0.4152	0.3973	0.2825	0.1813	0.4601	0.3969	0.2832	0.1878
	4-Shot	0.4060	0.3772	0.2705	0.1857	0.4591	0.3872	0.2779	0.1709
Llama-3.1-8B-Instruct	0-Shot	0.3976	0.3930	0.2915	0.1825	0.3156	0.3213	0.2479	0.3323
	1-Shot	0.3922	0.3762	0.2689	0.1783	0.3833	0.3658	0.2607	0.2370
	2-Shot	0.3813	0.3725	0.2643	0.1810	0.3946	0.3679	0.2630	0.2311
	4-Shot	0.3986	0.3702	0.2667	0.1821	0.3888	0.3594	0.2574	0.2099

Table 14: ICL results on IBM-ArgQ-5.3kArgs and IBM-Rank-30k.

#### Prompt Template of Training Data Construction for Persona Generation

Given an argument below, it is now necessary to analyze its strengths and weaknesses from different perspectives. Please provide  $\{N\}$  distinct personas suitable for evaluating its strengths and weaknesses from various viewpoints, along with their respective perspectives described in some words or phrases.

Ensure that the  $N$  personas have minimal overlap and provide the broadest possible coverage.

### Target Argument:

[Topic]:  $\{t\}$

[Target Argument]:  $\{a\}$

Please format your answers as follows, with each line being a JSON object:

$\{"persona\_id": 1, "persona": the\_name\_of\_persona1, "perspectives": the\_perspectives\_of\_persona1\}$

...

$\{"persona\_id": N, "persona": the\_name\_of\_personaN, "perspectives": the\_perspectives\_of\_personaN\}$

Table 15: Prompt template of training data construction for persona generation.

#### Prompt Template of Supervised Fine-tuning for Persona Generation

Given an argument below, it is now necessary to analyze its strengths and weaknesses from different perspectives. Please provide  $\{N\}$  distinct personas suitable for evaluating its strengths and weaknesses from various viewpoints, along with their respective perspectives described in some words or phrases.

Ensure that the  $N$  personas have minimal overlap and provide the broadest possible coverage.

### Target Argument:

[Topic]:  $\{t\}$

[Target Argument]:  $\{a\}$

### Output:

$\{p_1^d: p_1^p\}$

...

$\{p_N^d: p_N^p\}$

Table 16: Prompt template of supervised fine-tuning for persona generation.

## D Case Study

In this section, we present an example to illustrate how MPAQ generates diverse personas and their corresponding assessments. Table 20 shows an example argument with its associated personas and rationales. We can observe that MPAQ generates relevant personas who analyze the argument from

different perspectives aligned with their expertise. The generated rationales demonstrate how each persona contributes unique insights while maintaining a balanced assessment of both strengths and weaknesses. This example highlights MPAQ’s ability to provide comprehensive, multi-perspective evaluations of argument quality.

---

**Prompt Template of Training Data Construction for Rationale Generation**

---

You are a/an  $\{p_i^d\}$ . You are required to analyze the strengths and weaknesses of an argument on a controversial topic from the perspectives of a/an  $\{p_i^d\}$ . Your perspectives for analysis are as follows:  $\{p_i^p\}$ .

Now, given an argument below, analyze its argumentative strengths and weaknesses from your perspective as a/an  $\{p_i^d\}$ . Please only generate strengths/weaknesses that are consistent with your persona and perspectives, and make sure the generated strengths/weaknesses are reasonable and realistic.

### Target Argument:

[Topic]:  $\{t\}$

[Target Argument]:  $\{a\}$

### Answer:

Give your answer in json format like this: {"Strengths": [strengths], "Weaknesses": [weaknesses]}

---

Table 17: Prompt template of training data construction for rationale generation.

---

**Prompt Template for Placing Coarse-grained Score after Rationale**

---

You are a/an  $\{p_i^d\}$ . You are required to analyze the strengths and weaknesses of an argument on a controversial topic and assess its quality from your perspectives as a/an  $\{p_i^d\}$ . Your perspectives for analysis are as follows:  $\{p_i^p\}$

Now given an argument below, analyze its strengths and weaknesses and assess its argument quality indicated by an integer from 0 to 9 from your perspectives as a/an  $\{p_i^d\}$ .

### Target Argument:

[Topic]:  $\{t\}$

[Target Argument]:  $\{a\}$

### Output:

[Strengths]:

$\{r_i^s\}$

[Weaknesses]:

$\{r_i^w\}$

$\{s_i^{c,p}\}$

---

Table 18: Prompt template for placing coarse-grained score after rationale.



---

**Prompt Template for GPT-4o Evaluation**

---

Multi-persona argument quality assessment framework assesses the quality of an argument from multiple perspectives. It first generates  $\{N\}$  personas and corresponding perspectives for assessing quality of input argument, then analyzes its argumentative strengths and weaknesses (i.e., analytical rationales) from each of the  $\{N\}$  personas and perspectives. Now, there are personas/perspectives and corresponding analytical rationales generated from two different models (i.e., Model 1 and Model 2), please evaluate which model generates better personas and which model generates better rationales separately.  
0 indicates Model 1 is better than Model 2, 1 indicates Model 2 is better than Model 1, 2 indicates Model 1 and Model 2 are equally good.

### Target Argument:

[Topic]:  $\{t\}$

[Target Argument]:  $\{a\}$

### Model 1:

[Persona 1]:  $\{p_1^d\}$

[Perspectives 1]:  $\{p_1^p\}$

[Rationale 1]: [Strengths]:  $\{r_1^s\}$  | [Weaknesses]:  $\{r_1^w\}$

...

### Model 2:

[Persona 1]:  $\{p_1^d\}$

[Perspectives 1]:  $\{p_1^p\}$

[Rationale 1]: [Strengths]:  $\{r_1^s\}$  | [Weaknesses]:  $\{r_1^w\}$

...

### Answer:

Give your answer in json format like this: {"Persona": the\_model\_generating\_better\_personas, "Rationale": the\_model\_generating\_better\_rationales}

---

Table 19: Prompt template for GPT-4o evaluation.

---

**Topic:** We should legalize cannabis.

**Argument:** It is a person's right to choose what he or she puts in his or her body. Outlawing marijuana violates that right.

**Ground-truth Quality Score:** 0.7554

---

**Stage 1: Persona Generation**

{"Legal Scholar": "Examine constitutional rights, and implications of personal autonomy."}

{"Public Health Expert": "Assess health impacts, potential benefits, and risks associated with cannabis legalization."}

{"Sociologist": "Explore societal attitudes, cultural implications, and effects on communities."}

{"Economist": "Analyze economic impacts, taxation, and market effects of legalizing cannabis."}

---

**Stage 2: Argument Quality Assessment**

**Legal Scholar:**

[Coarse-grained Predicted Score]: 8

[Fine-grained Predicted Score]: 0.8043

[Strengths]: ['The argument appeals to the principle of personal autonomy, which is a fundamental aspect of constitutional rights, particularly in the context of bodily integrity and self-determination.', 'The argument can be supported by the growing body of evidence that suggests cannabis has medicinal benefits, which further reinforces the notion of personal autonomy in healthcare decisions.']

[Weaknesses]: ['The argument may overlook the complexities of public health and safety concerns, such as the potential for increased substance abuse and its impact on society, which legal scholars must consider in evaluating the implications of personal autonomy.', 'The argument does not address the government's compelling interest in regulating substances for public health and safety, which can override personal choice in certain circumstances. Courts often allow restrictions on personal freedoms when there is a significant public interest at stake.']

**Public Health Expert:**

[Coarse-grained Predicted Score]: 7

[Fine-grained Predicted Score]: 0.7529

[Strengths]: ['The argument emphasizes personal autonomy and individual rights, which are important principles in public health ethics. Respecting a person's choice may encourage a more engaged and informed public regarding substance use.', 'Legalizing cannabis could potentially reduce the stigma associated with its use, encouraging open discussions about mental health and substance use, which can lead to better public health outcomes.']

[Weaknesses]: ['The argument does not address the potential health impacts of cannabis use, such as addiction, mental health issues, or adverse effects on developing brains, which are important considerations in public health discussions.', 'It overlooks the potential for increased access to cannabis leading to more individuals using it, which could result in public health risks, particularly among vulnerable populations who may not have the resources to make informed choices.']

**Sociologist:**

[Coarse-grained Predicted Score]: 8

[Fine-grained Predicted Score]: 0.7817

[Strengths]: ['The argument emphasizes personal autonomy and individual rights, which align with contemporary societal values that prioritize personal freedom and self-determination.', 'The legalization of cannabis could generate economic benefits for communities, as it has the potential to create jobs and boost local economies through tax revenues and business opportunities.', 'It highlights the importance of personal responsibility, suggesting that individuals should be trusted to make informed choices about their bodies, which can contribute to a more empowered and educated society.']

[Weaknesses]: ['The argument does not address potential societal implications, such as the impact on public health, safety, and the potential for increased substance abuse among vulnerable populations.', 'It overlooks cultural differences and the varying attitudes towards cannabis across different communities, which can lead to conflicts in societal norms and values.', 'The argument lacks consideration of the legal and regulatory frameworks necessary to ensure responsible use and prevent negative community effects, such as increased crime rates or increased traffic accidents.']

**Economist:**

[Coarse-grained Predicted Score]: 7

[Fine-grained Predicted Score]: 0.7119

[Strengths]: ['Legalizing cannabis can lead to significant economic benefits, including increased tax revenue from sales and reduced enforcement costs, which can be reinvested in public services.', 'The legalization of cannabis may encourage a regulated market, which can enhance consumer safety and quality control, leading to a more stable market for cannabis products.', 'Allowing personal choice in consumption can promote individual freedom and autonomy, aligning with economic principles of consumer sovereignty and market efficiency.']

[Weaknesses]: ['The argument does not address potential negative externalities, such as increased substance abuse or health issues that could arise from widespread legalization, which may affect overall economic well-being.', 'It overlooks the complexities of taxation and regulation that could arise from legalization.', 'The claim does not consider the potential market dynamics that could arise from legalization, such as monopolies or oligopolies forming within the cannabis industry, which could limit competition and lead to higher prices for consumers.']

**[Aggregated Overall Predicted Score]: 0.7627**

---

Table 20: Case study.

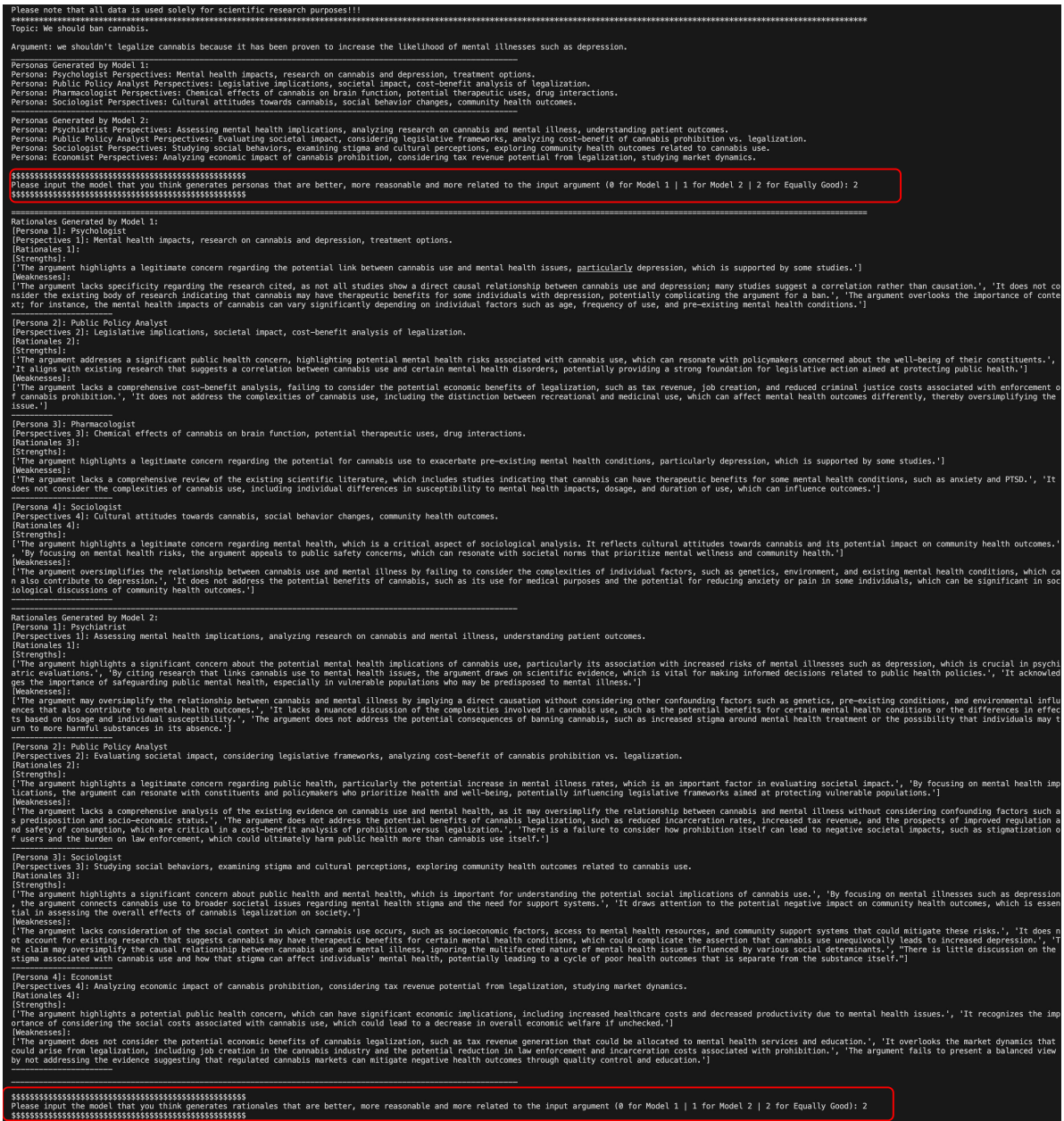


Figure 9: The interface of human evaluations.