

DEBATEQA: Evaluating Question Answering on Debatable Knowledge

Anonymous ACL submission

Abstract

The rise of large language models (LLMs) has enabled us to seek answers to inherently debatable questions on LLM chatbots, necessitating a reliable way to evaluate their ability. However, traditional QA benchmarks assume fixed answers are inadequate for this purpose. To address this, we introduce **DEBATEQA**, a dataset of 2,941 debatable questions, each accompanied by multiple human-annotated partial answers that capture a variety of perspectives. We develop two metrics: *Perspective Diversity*, which evaluates the comprehensiveness of perspectives, and *Dispute Awareness*, which assesses if the LLM acknowledges the question’s debatable nature. Experiments demonstrate that both metrics are aligned with human preferences and stable across different underlying models. Using DEBATEQA with two metrics, we assess 12 prevalent LLMs and retrieval-augmented generation methods. Our findings reveal that while LLMs generally excel at recognizing debatable issues, their ability to provide comprehensive answers encompassing diverse perspectives varies considerably.

1 Introduction

How often do you query a chatbot about a debatable issue? Questions like “Does Donald Trump have a terrible character?” or “How do crop circles form?” frequently arise in everyday life, reflecting human beings’ natural curiosity about topics that inherently lack fixed answers (Lowry and Johnson, 1981; Brady, 2009). With the advent of large language models (LLMs; OpenAI 2023b; Chowdhery et al. 2023; Touvron et al. 2023), we now turn to these models to seek “proper” answers to such questions. Evaluating the performance of language models on debatable queries is crucial for understanding and enhancing their capabilities.

Inherent difficulties hinder our way to assess model performance on such questions. Traditional question-answering (QA) benchmarks are typically

designed to provide fixed answers to questions, as in datasets like SQuAD (Rajpurkar et al., 2016) and Natural Questions (Kwiatkowski et al., 2019). Even in scenarios where multiple legitimate answers are possible, such as TruthfulQA (Lin et al., 2022), the answers are often presented as multiple-choice questions, limiting the space of responses. Recent years have seen the emergence of long-form QA evaluations, such as ELI5 (Fan et al., 2019) and ASQA (Stelmakh et al., 2022a), which allow for more elaborate answers. However, these works have not focused on inherently debatable questions. The most related work to us is DELPHI (Sun et al., 2023), which curates a dataset with controversial questions, however, their emphasis is on identifying controversy rather than delving deeper into evaluating models’ responses.

What constitutes a proper answer when asked about debatable knowledge? Drawing inspiration from interdisciplinary literature, we propose two desirable properties for such answers. Firstly, the answer should inform the receiver that the issue at hand is debatable (Misco, 2011). Secondly, a proper answer should be comprehensive and include diverse perspectives, which are crucial for maintaining an atmosphere of neutrality, especially in public-related inquiries (Habermas, 1991). Motivated by these considerations, we introduce the **DEBATEQA** dataset, which comprises 2,941 manually annotated debatable questions. To rigorously evaluate the aforementioned properties of model answers, each question is paired with several *partial answers*, each reflecting a single viewpoint on the debatable issue. These partial answers are generated through a three-stage pipeline and then annotated by three annotators. Based on the partial answers in DEBATEQA, we propose the metric of **Perspective Diversity (P.D.)**, which assesses the comprehensiveness in terms of grasping multiple points-of-view in the model answer. Using DEBATEQA along with P.D., one can reliably and

efficiently evaluate the comprehensiveness of responses to debatable questions. We also propose the metric of **Dispute Awareness (D.A.)**, which targets to identify whether the model acknowledges the debatable nature of the question in its response. Importantly, we show that the P.D. and D.A. metrics align closely with human judgments and are fairly stable across backbone evaluator models, validating their effectiveness.

We apply DEBATEQA to evaluate 12 popular LLMs, as well as retrieval-augmented generation (RAG) approaches. We observe that while LLMs generally excel at identifying the existence of debate, their performance varies in providing comprehensive answers with diverse perspectives. Notably, top-notch open-source LLMs rival or even surpass some leading commercial models. Besides, RAG methods, though not uniformly beneficial to this task, improve closed-source model performance, likely due to better contextual leverage. Additionally, optimizing sampling hyperparameters and leveraging task-specific prompts can further boost performance. On the whole, our findings underscore the need for further refining LLMs to better interact with debatable knowledge.

2 Related Work

QA with non-fixed answers. Many efforts focus on QA for which there is no single fixed answer. AMBIGQA (Min et al., 2020) addresses ambiguous question answering by rewriting questions and generating multiple plausible answers. SUBJQA (Bjerva et al., 2020) focuses on identifying subjectivity in questions and answers within customer reviews. DisentQA (Neeman et al., 2023) proposes to provide disentangled answers to questions where the provided context contradicts the model’s knowledge. DEBATEQA differs logically from these efforts because the space of plausible answers can not be narrowed by rewriting or restricting the questions. To the best of our knowledge, DELPHI (Sun et al., 2023) is the first study on QA for debatable issues. However, DELPHI has limitations: (1) it does not provide answers for evaluation, and (2) it shallowly evaluates model performance using exploratory metrics. Our work represents a step forward from DELPHI by offering a comprehensive evaluation solution. We expand and refine the dataset and introduce more meaningful metrics for performance evaluation.

NLP on debatable issues. Beyond QA, multiple

lines of NLP research investigate debatable issues. One notable effort is the AI Debater, beginning with Slonim et al. (2021)’s “IBM’s Project Debater” (Bar-Haim et al., 2019), the first AI system designed to engage humans in meaningful debates. Another line of research focuses on controversy detection. Researchers have identified controversy in news (Choi et al., 2010), online forums (Chen et al., 2023), and other media by analyzing sentiments (Choi et al., 2010; Chen et al., 2023) or sociological features like upvote percentages (Hessel and Lee, 2019). Moreover, Wan et al. (2024) investigate LLMs’ preference for conflicting evidence when facing controversial issues. We distinguish ourselves from this body of research by primarily focusing on handling debatable issues in the field of QA, specifically targeting the evaluation of chatbot-like NLP systems.

Long-form QA evaluation. Evaluating debatable QA falls into the topic of long-form QA evaluation. Long-form text evaluation can be categorized into two main approaches: reference-based and reference-free evaluation (Xu et al., 2023). Reference-based methods require gold answers and evaluate the generated text by assessing its similarity with the gold answers (Fan et al., 2019; Yuan et al., 2021; Chiang and Lee, 2023). Conversely, reference-free evaluation eliminates the necessity for a gold standard. Some studies assess the coherence and relevance of the generation concerning specified questions (Fabbri et al., 2022; Krishna et al., 2022; Xu et al., 2023). Some studies use a QA-based approach to assess the quality of the generation (Tan et al., 2024). In particular, there is another line of literature focusing on examining the veracity of long-form generation by utilizing external knowledge bases (Stelmakh et al., 2022b; Min et al., 2023; Wei et al., 2024). However, none of the aforementioned studies address the evaluation of debatable questions.

3 Curating the DEBATEQA Dataset

Dataset overview. DEBATEQA is designed to help assessing *LMs’ answers to debatable questions*. It contains 2,941 debatable questions, each paired with a list of **partial answers** to assist in evaluating model responses. Each partial answer addresses the question from a distinct perspective and consists of two parts: a short *point-of-view* statement (POV) and a long-form *explanation* (Explan) that

Field	Content
Question	Does birth order influence personality traits?
Partial Answer 1	POV Birth order does not have a meaningful and lasting effect on broad Big Five personality traits. Explan The influence of birth order on personality traits has been a topic of interest for over a century. However, based on extensive research combining large datasets from three national panels in the United States, Great Britain, and Germany, it is evident that birth order does not have a meaningful effect on broad Big Five personality traits ...
Partial Answer 2	POV Firstborns score higher on intelligence and intellect. Explan Yes, birth order does influence personality traits, particularly in the domain of intelligence and intellect. Research has consistently shown that firstborns tend to score higher on objectively measured intelligence tests ...
Partial Answer 3	POV No birth-order effects on extraversion, emotional stability, agreeableness, or conscientiousness. Explan The influence of birth order on personality traits such as extraversion, emotional stability, agreeableness, and conscientiousness has been a topic of interest for over a century. However, recent comprehensive studies have provided substantial evidence that birth order does not significantly impact these personality traits ...

Table 1: An example from DEBATEQA, details of the explanation fields are truncated for space issues.

fully expands the POV¹. An entry in DEBATEQA is shown in Table 1.

The overall procedure for curating DEBATEQA is depicted in Figure 1. We first source debatable questions. Then, we apply a three-stage semi-automated pipeline to collect partial answers. Lastly, we conduct human annotation on the collected partial answers to finalize the dataset. Quality examinations happen after each step. The following sections will detail these steps.

3.1 Sourcing Debatable Questions

We collect debatable questions from three distinct sources. First, we repurpose two existing datasets: we select 2,281 annotated controversial questions from DELPHI (Sun et al., 2023) and a full set of 434 questions from CONFLICTINGQA (Wan et al., 2024). To enrich the existing data, we further manually sourced 1,758 additional debatable questions from the Web (see Table 6 for detailed sources). We then run a deduplication algorithm (see § A.1 for details) to remove any duplicate questions, resulting in 3,216 questions. The final composition of sourced questions is shown in Table 7.

3.2 Collecting Partial Answers

The core novelty of DEBATEQA lies in evaluating models by comparing the response with multiple partial answers, rather than a single gold reference. One partial answer aims to answer the question from a single perspective. This method reflects the multifaceted essence of debatable knowledge, advocating for answers that integrate diverse viewpoints (Habermas, 1991; Wansink et al., 2023). To this end, we employ a three-stage pipeline for

collecting them: first, we collect evidence documents from trustworthy websites; second, we extract POVs from the evidence *w.r.t.* the question; finally, we expand the POVs into long-form explanations based on related evidence. The last two steps are conducted with the assistance of LLMs². Together, the *POVs and explanations* comprise what we call partial answers.

3.2.1 Retrieving Trustworthy Documents

We collect partial answers by leveraging online resources and extracting evidence from relevant web pages. However, the nature of debatable issues necessitates careful processing of these documents, as the Web can contain unveracious content. To ensure the reliability of our partial answers, we source documents from authoritative top-level domains (TLDs), as listed in Table 8. This treatment helps in maintaining the reliability of the sources. We discard questions that have fewer than three documents, resulting in 2,982 questions, each supported by 3-5 of the most relevant documents. See § A.2 for detailed measures.

Quality examination. To assess the quality of retrieved documents, we analyze the relevancy between questions and corresponding documents. We calculate the cosine similarity between document chunks and questions. As depicted in Figure 4, the average cosine similarity for document trunks is 0.56 and there are no significant outliers, indicating high relevance and minimal noise in the documents, confirming their overall quality for serving as the basis for upcoming steps.

3.2.2 Extracting Points-of-View

The second stage involves extracting diverse POVs from the retrieved evidence documents. A POV

¹In argumentation mining literature (Stede et al., 2019), partial answers can be viewed as *arguments*, with the POV as the *claim* and the Explain as the *reasons*.

²We select OpenAI GPT-4 (OpenAI, 2023b) to assist in collecting partial answers (the gpt-4-turbo variant).

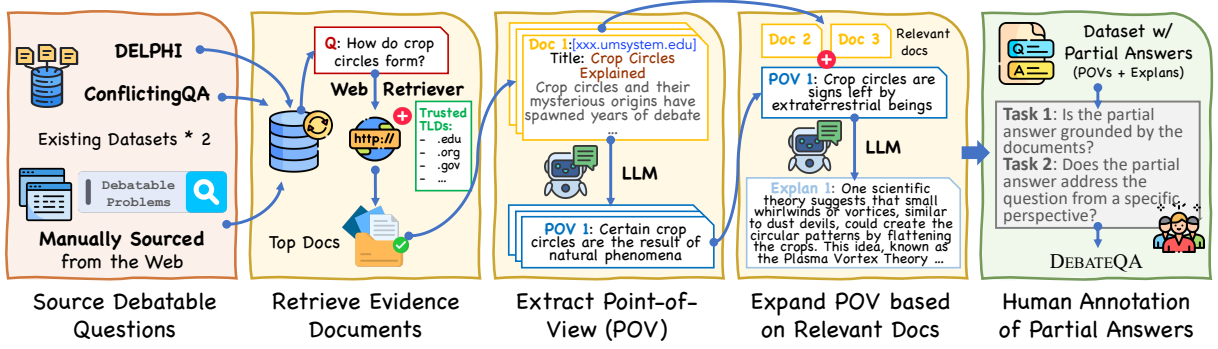


Figure 1: Pipeline for curating DEBATEQA. The three main components of the pipeline are highlighted in different colors: **sourcing debatable questions**, **collecting partial answers**, and **human annotation**. Primary sources or tools used at each step are highlighted in **bold**.

is a concise statement that reflects the core perspective in addressing the question. We leverage GPT-4 to tackle this task, by applying the prompt p_{POV} described in Table 9, which takes the question and the concatenated documents and returns a list of diverse POVs along with the corresponding document indexes where each specific POV is originated. The document indexes for each POV are later used for expanding the POV. To avoid exceeding the 128K context window limit of GPT-4, we preprocess the documents by removing meaningless segments and truncating them to 120K tokens if they exceed this length.

Quality examination. We verify the quality of the collected POVs on *comprehensiveness w.r.t.* the documents and *distinctiveness* among themselves. For comprehensiveness, we ensure all valid perspectives from retrieved documents are captured, with 90.4% coverage verified manually. Distinctiveness is assured by removing duplicated POVs manually. For more details, refer to § A.3.

3.2.3 Expanding POV to Explanations

The last stage involves expanding the extracted POVs into long-form explanations. Each explanation should stand as an independent answer, elaborating on the POV and addressing the question from that perspective. This expansion must be *anchored* to the relevant information presented in the evidence documents pertaining to the specific POV being developed. We again leverage GPT-4 on this task, utilizing the prompt p_{Explan} described in Table 9. This prompt takes three inputs: the question, the target POV to be expanded, and the related documents obtained in the previous stage. The LLM is required to leverage only the information contained within these relevant documents

to generate the explanation, minimizing the risk of hallucinations (Zhang et al., 2023). We repeat this step for all the POVs we have collected. The pseudocode of the pipeline for collecting partial answers is deferred to Algorithm 1.

3.3 Human Annotation

We verify the fidelity of the LLM-generated partial answers through human annotation.

Annotation tasks. To thoroughly evaluate the quality of DEBATEQA, we design two tasks:

- **Task 1:** Ensure that the POV and the explanation generated by the LLM are grounded in the documents. This task focuses on that the generated explanations are accurately derived from trustworthy evidence.
- **Task 2:** Ensure that the partial answer can address the question from a certain perspective. This task assesses the utility and relevance of the partial answers, ensuring that they address the question effectively.

Results and the final dataset. We recruit three annotators and annotate the full dataset. Inter-annotator agreement (IAA) is measured using Fleiss’ Kappa (Fleiss et al., 1981), yielding scores of $\kappa = 0.66$ and $\kappa = 0.60$ for the two annotation tasks, all indicating substantial agreement. We remove 767 partial answers deemed substandard by two or more annotators. This suggests that GPT-4 generates faithful partial answers with a 93.4% accuracy. See § A.4 for details. We employ BERTopic (Grootendorst, 2022) to model the domain distribution of DEBATEQA. The result is shown in Figure 7. To reduce computational costs for upcoming evaluation, we split DEBATEQA into two splits: the test split with 1,000 randomly sam-

pled questions and the dev set containing the remaining instances.

4 Evaluation Criteria and Metrics

Evaluation criteria. For debatable questions, the quest for the *best* answer is fraught with complexity, as there exist no canonical standards. The expectations vary: many seek a *helpful* assistant that delivers credible information with no reservation, and some may pose questions merely for self-affirmation, not to say some model providers prefer a “*safe*” agent to circumvent controversy. In DEBATEQA, we aim to balance helpfulness and harmlessness, with the goal of fostering open dialogues. After reviewing interdisciplinary literature, we identify criteria that are nearly universally accepted and distill two key properties that define a *good* answer³:

- **Perspective diversity** (*helpful*): How well does the answer deliver informative and credible information from diverse perspectives?
- **Dispute awareness** (*harmless*): Whether the answer recognizes the existence of debate.

Evaluation metrics. Although the above two criteria resonate with those utilized in DELPHI (Sun et al., 2023), we distinguish ourselves by formalizing these criteria into more sophisticated, quantifiable metrics. Our methodology excels by integrating the partial answer feature of DEBATEQA and outperforms DELPHI’s approach by a huge margin, which will be later elaborated in § D and § E. Please note that while the following two metrics both employ backbone LLMs, they may *differ*; for brevity, we refer to both as $\mathcal{M}_{\text{eval}}$.

I: Perspective Diversity (P.D.). To evaluate the model’s answers against legitimate partial responses, we assess it using a smaller open-source LLM by generation, ensuring efficiency and cost-effectiveness. We apply the following metric:

$$\text{P.D.} := \sum_{i=1}^n \text{PPL}(\text{PA}^i | \text{chatTemplate}(\text{Concat}(A, \text{“Please restate.”}))), \quad (1)$$

where $\text{PPL}(Y|X)$ is the conditional perplexity:

$$\text{PPL}(Y|X) = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(y_i | X, y_{<i}) \right). \quad (2)$$

³Dimensions like factuality (Min et al., 2023) are also legitimate, but we focus mainly on the nature of debatability.

In Equation 1, $\text{PA}^i = \text{concat}(\text{POV}^i, \text{Explan}^i)$ denotes the i^{th} partial answer and A denotes the model’s answer to the debatable question. $\text{chatTemplate}()$ is a chat template for prompting instruction-tuned LLMs. Simply put, P.D. represents the aggregate perplexity of generating partial answers from a model answer. Being derived from PPL, a **lower** P.D. signifies better quality, suggesting that the model answer contains larger shares of the partial answer’s content.

II: Dispute Awareness (D.A.). To ascertain if the model’s answer indicates that the addressed question is debatable, we craft a prompt $p_{\text{D.A.}}$, as shown in Table 13, and use it to prompt an instruction-tuned LLM. This metric is *binary*, indicating awareness or lack thereof.

$$\text{D.A.} := \begin{cases} 1 & \text{if } \mathcal{M}_{\text{eval}}(p_{\text{D.A.}}(q, \text{Ans})) \text{ returns “1”,} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Empirical examinations of our metrics. To establish the reliability and effectiveness of our proposed evaluation metrics, we conduct comprehensive empirical examinations that validate both P.D. and D.A. across multiple dimensions. For P.D., we demonstrate strong alignment with human preferences through pairwise ranking experiments involving 500 model responses from five different LLMs on 100 randomly sampled questions. Our results show that P.D. achieves superior correlation with human judgment (Spearman’s $\rho = 0.825$, Kendall’s $\tau = 0.748$ with GPT-2 backbone), significantly outperforming existing baselines, including G-Eval powered by GPT-4o ($\rho = 0.706$, $\tau = 0.634$).

For D.A., we validate its accuracy against human-annotated ground truth labels with substantial inter-annotator agreement (Fleiss’ Kappa = 0.79). The metric achieves strong performance across multiple backbone models, with accuracy reaching 0.80 and F1 scores up to 0.869 using appropriate instruction-following LLMs, outperforming DELPHI’s “Acknowledge” metric.

These empirical findings, presented comprehensively in § D and § E, establish the validity and practical utility of our evaluation framework for assessing model responses to debatable questions.

5 Experiments

5.1 Experimental Setup

We evaluate a wide range of 12 LLMs on DEBATEQA-test using P.D. and D.A., including

Model	Avg. Len. (#tokens)	Perspective Diversity (P.D.)					Dispute Awareness (D.A.)				
		$\mathcal{M}_{\text{eval}}=\text{Qwen2 0.5B}$		$\mathcal{M}_{\text{eval}}=\text{GPT-2}$		Norm. Rank	$\mathcal{M}_{\text{eval}}=\text{Phi-3 M.}$		$\mathcal{M}_{\text{eval}}=\text{Qwen2 1.5B}$		Norm. Rank
		Score ↓	Rank	Score ↓	Rank		Score ↑	Rank	Score ↑	Rank	
Closed-Source LLMs											
GPT-4o	434	3.07	1	4.03	1	1	0.952	1	0.979	1	1
GPT-4o mini	252	4.09	6	5.88	6	6	0.937	4	0.964	4	4
GPT-3.5 Turbo	141	5.28	10	8.25	10	10	0.904	6	0.947	6	6
Claude 3.5 Sonnet	199	4.63	8	6.96	8	8	0.856	10	0.920	9	10
Open-Source LLMs (Medium to Large)											
Llama3 70B	432	3.09	2	4.07	3	2=	0.945	3	0.977	2	2=
Llama3 8B	381	3.51	5	5.02	5	5	0.928	5	0.964	4	5
Qwen2 7B	255	4.18	7	6.10	7	7	0.895	8	0.923	8	8
Phi-3 small 128k	412	3.50	4	4.31	4	4	0.899	7	0.924	7	7
Gemma 2 9B	395	3.12	3	4.04	2	2=	0.947	2	0.967	3	2=
Open-Source LLMs (Tiny to Small)											
Qwen2 1.5B	169	5.60	11	8.67	11	11	0.864	9	0.875	10	9
Qwen2 0.5B	72	<u>6.56</u>	12	<u>10.87</u>	12	12	0.792	11	0.836	11	11
Phi-3 mini 128k	218	4.82	9	<u>7.33</u>	9	9	<u>0.716</u>	12	<u>0.794</u>	12	12

Table 2: Main results of P.D. and D.A. for LLMs on DEBATEQA-test. Avg. Len.: average length of the answers, GPT-2: GPT-2 (117M), Phi-3 M.: Phi-3 medium 128k, Norm. Rank: normalized average rank of different $\mathcal{M}_{\text{eval}}$. The best and worst results of each metric (*w.r.t.* a specific $\mathcal{M}_{\text{eval}}$) are shown in **bold** and underlined, respectively.

closed commercial LLMs and open-source ones. We also assess several RAG approaches.

Evaluated models. We evaluate the following LLMs: GPT-4o (OpenAI, 2024b), GPT-4o mini (OpenAI, 2024a), GPT-3.5 Turbo (OpenAI, 2023a), Claude 3.5 Sonnet (Anthropic, 2024), Llama3 Instruct 8B/70B (Meta, 2024), Qwen2 0.5B/1.5B/7B (Qwen, 2024a), Phi-3 mini 128k 3.8B and Phi-3 small 128k 7B (Abdin et al., 2024), and Gemma2 9B (Team, 2024). All models are the instruction or chat fine-tuned versions.

Evaluators. We select multiple language models as the backbone for our metrics. For evaluating P.D., we select Qwen2 0.5B and GPT-2 base 117M (Radford et al., 2019) as $\mathcal{M}_{\text{eval}}$. For assessing D.A., a competent LLM with instruction-following ability is a must. We select Phi-3 medium 128k 14B and Qwen2 1.5B. We select those four models because their performance is showcased in § D.1 and § E.1, respectively.

Generation configuration. In the main experiments, when testing the LLMs, we provide the questions with a minimalistic QA prompt, as shown in Table 16, which instructs the LLMs without any hint that they are debatable. We believe this approach more accurately reflects the typical user interaction with chatbots. For all models, we configure top- $p = 0$ to enable greedy decoding and stock chat templates including $\mathcal{M}_{\text{eval}}$.

5.2 Main Results

Main evaluation results can be found in Table 2. We summarize our key findings as follows:

- **Larger models generally outperform small ones.** Generally, larger models perform well in terms of both P.D. and D.A. metrics. Large state-of-the-art LLMs such as GPT-4o and Llama3 70B demonstrate the strongest performance, while tiny to small LLMs are almost always at the bottom. Besides, for LLMs with the same architecture (*e.g.*, Qwen 2 0.5B/1.5B/7B), larger models always outperform smaller ones.
- **Gaps between closed and open models are not clear.** With the exception of the superb performance of GPT-4o, we have observed that numerous open-source LLMs are outperforming other capable closed commercial LLMs. Specifically, open-source models like Llama3 70B and Gemma2 9B nearly match the performance of GPT-4o in handling debatable questions. This might indicate that the performance on handling debatable issues does not demand as stringent capabilities from the models as more difficult tasks such as reasoning.
- **Deficiencies in delivering comprehensive responses.** We observe significant shortcomings in weaker models’ ability to furnish comprehensive answers that encompass a variety of perspectives. For instance, the worst-performing

Qwen2 0.5B’s answers are around 3 to 9×4 worse than GPT-4o in terms of recovering the information in partial answers.

- **Models excel in recognizing debate.** We find that even the lowest-performing model, namely Qwen2 0.5B, demonstrates a D.A. over 70% in recognizing the contentiousness of debatable questions. This indicates that even the performance of the worst-performing model is not as poor as it might seem.

Correlation between the P.D. and D.A. metrics. To investigate the correlation between P.D. and D.A. for the tested models, we plot a visualization of the results in Figure 2, which aids in understanding how these two metrics relate to each other across various models. The figure indicates a positive correlation between P.D. and D.A., suggesting that models with greater perspective diversity are also more aware of the debate in question, aligning with our expectations.

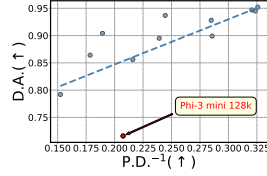


Figure 2: Correlation between P.D. and D.A. for the tested LLMs.

Integrating the P.D. and D.A. metrics. To demonstrate the models’ general capability in addressing debatable questions, we integrate the two metrics with varying weights. To ensure that the P.D. scores reflect the same performance favorability as the D.A. scores, we first take the reciprocal of the P.D. scores. Then, we leverage z-normalization⁵ to normalize both the reciprocal P.D. score and the original D.A. score. The final weighted mean μ_w is calculated as:

$$\mu_w = k \cdot \text{Z-norm}(\text{P.D.}^{-1}) + (1 - k) \cdot \text{Z-norm}(\text{D.A.}), \quad (4)$$

where the k is the proportional coefficient and $\text{Z-norm}()$ refers to z-normalization. We show the ranking of the weighted average scores for the models at different k in Figure 3.

5.3 Effect of More Specific Prompts

In our main experiments, we use a simple QA prompt (p_{basic} in Table 18) that does not highlight the debatable nature of the questions or demand comprehensive answers. To evaluate the

⁴These values are approximated with the P.D. values w.r.t. Two different $\mathcal{M}_{\text{eval}}$, refer to § H for details.

⁵https://en.wikipedia.org/wiki/Standard_score

More on D.A.			More on P.D.		
Rank	k=0	k=0.25	k=0.5	k=0.75	k=1
1	GPT-4o	GPT-4o	GPT-4o	GPT-4o	GPT-4o
2	Gemma 2 9B	Gemma 2 9B	Llama3 70B	Llama3 70B	Llama3 70B
3	Llama3 70B	Llama3 70B	Gemma 2 9B	Gemma 2 9B	Gemma 2 9B
4	GPT-4o-mini	Llama3 8B	Llama3 8B	Llama3 8B	Phi-3 small 128k
5	Llama3 8B	GPT-4o-mini	Phi-3 small 128k	Phi-3 small 128k	Llama3 8B
6	GPT-3.5 Turbo	Phi-3 small 128k	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini
7	Phi-3 small 128k	Qwen2 7B	Qwen2 7B	Qwen2 7B	Qwen2 7B
8	Qwen2 7B	GPT-3.5 Turbo	GPT-3.5 Turbo	Claude 3.5 Sonnet	Claude 3.5 Sonnet
9	Qwen2 1.5B	Claude 3.5 Sonnet	Claude 3.5 Sonnet	GPT-3.5 Turbo	Phi-3 mini 128k
10	Claude 3.5 Sonnet	Qwen2 1.5B	Qwen2 1.5B	Qwen2 1.5B	GPT-3.5 Turbo
11	Qwen2 0.5B	Qwen2 0.5B	Qwen2 0.5B	Phi-3 mini 128k	Qwen2 1.5B
12	Phi-3 mini 128k	Phi-3 mini 128k	Phi-3 mini 128k	Qwen2 0.5B	Qwen2 0.5B

Figure 3: Rankings of weighted average scores of models at different ratios. k is the proportional coefficient.

models’ full potential, we further test five LLMs with more detailed prompts. We employ three system prompts— p_{basic} , $p_{\text{comprehensive}}$, and p_{detailed} —to elicit model responses at varying levels of detail, as shown in Table 18. Using 200 randomly sampled questions from DEBATEQA-test, we compare the average P.D. and D.A. scores across the five selected LLMs. The results for these prompts are presented in Table 3. We find that even the relatively simple $p_{\text{comprehensive}}$ prompt significantly improved the performance for all five models. We conclude that more specific prompts, i.e., informing the model of the debatable nature and requesting detailed responses, can enhance LLMs’ performance in answering such questions. Users can benefit from well-crafted prompts when seeking answers to contentious issues from LLMs.

Model	P.D. (↓)			D.A. (↑)		
	p_b	p_c	p_d	p_b	p_c	p_d
GPT-4o mini	3.91	2.13	2.09	0.915	0.955	0.97
Claude 3.5 Sonnet	4.63	3.14	2.35	0.865	0.925	0.98
Llama3 8B	3.42	2.58	2.51	0.855	0.935	0.985
Qwen2 7B	4.16	2.78	2.76	0.855	0.915	0.965
Phi-3 mini 128k	4.71	3.40	2.88	0.765	0.925	0.96

Table 3: Effect of various prompts on P.D. scores and D.A. scores. p_b : p_{basic} , p_c : $p_{\text{comprehensive}}$, p_d : p_{detailed} .

5.4 Evaluation for RAG Approaches

In this section, we assess the influence of RAG methods on performance within DEBATEQA.

5.4.1 Effect of RAG Strategy

We examine the effects of two popular RAG strategies, *vanilla RAG* (Lewis et al., 2020b) and *ReAct* (Yao et al., 2023). In vanilla RAG, we pick the top-10 most relevant documents from the retrieval results via Google Custom Search API⁶. ReAct employs an agent-based approach, leveraging Claude 3.5 Sonnet to interleave reasoning with document

⁶<https://developers.google.com/custom-search>

retrieval, strategically selecting up to 9 document chunks to improve problem-solving. Both methods utilize the prompt in Table 17 to assemble the question and the retrieved trunks. Refer to § F.1 for details.

We assess the performance of five LLMs by evaluating their responses to 100 randomly sampled debatable questions from DEBATEQA-test using two distinct RAG strategies. With the results detailed in Table 4, we conclude:

- LLMs with RAG do not consistently improve in answering debatable questions, but closed-source models see more notable benefits, possibly due to better context utilization despite potential noise in retrieved content chunks.
- Among the two RAG strategies, ReAct consistently outperforms Vanilla RAG, even though it uses fewer document chunks (9 vs. 10). This advantage can be attributed to ReAct’s more strategic approach to acting based on the previously retrieved information, which leads to more precise and relevant documents retrieved.

Model	P.D. ($\downarrow \mathcal{M}_{\text{eval}}=\text{Qwen2 0.5B}$)		
	No RAG	Vanilla RAG	ReAct
GPT-4o mini	4.02	3.94	3.70
Claude 3.5 Sonnet	4.63	4.12	3.65
Llama3 8B	3.55	4.01	3.99
Qwen2 7B	3.79	5.96	5.29
Phi-3 mini 128k	4.82	7.01	6.86

Table 4: Effect of two RAG strategies on P.D. scores.

5.4.2 Effect of RAG Source Documents

Considering that the performance of RAG is highly dependent on the quality of the retrieved documents, we explore whether restricting RAG to utilize trustworthy documents would yield better results. We retrieve only on web pages under trustworthy TLDs listed in Table 8. The results in Table 5 demonstrate that RAG on trustworthy sources leads to better results. This highlights the significance of source quality in RAG for debatable QA, emphasizing that utilizing trustworthy documents improves LLM response quality in responding to sensitive topics.

5.5 Qualitative Analysis

Despite GPT-4o leading in both metrics, certain aspects of its performance remain unsatisfactory. We also examine the performance gaps among less

Model	P.D. ($\downarrow \mathcal{M}_{\text{eval}}=\text{Qwen2 0.5B}$)	
	Vanilla RAG	RAG w. T. Docs
GPT-4o mini	3.77	3.63
Claude 3.5 Sonnet	3.92	3.54
Llama3 8B	3.78	3.62
Qwen2 7B	5.91	5.57
Phi-3 mini 128k	6.77	6.50

Table 5: Effect of RAG sources on P.D. scores. RAG w. T. Docs: RAG using trustworthy documents.

capable models. Through case studies, our conclusions are: (1) state-of-the-art LLMs still fall short in leveraging referenced and convincing information for their arguments, and (2) weaker LLMs have difficulty providing a comprehensive elaboration of sufficient perspectives and we find the answer presentation of those LLMs are markedly poorer. Based on these gaps, LLMs’ ability to address debatable questions still has room to be further refined. We point out the following potential solutions for improving models’ ability on DEBATEQA: (1) *Deepen domain-specific knowledge* with pretraining or high-quality RAG; (2) *Strengthen source citation* by adopting strategies such as citing memory or retrieved documents for evidence-rich responses; (3) *Improve stylistic presentation* through fine-tuning; and (4) *Ensure perspective diversity and balance* through alignment training or prompt engineering. Kindly refer to § G for a detailed analysis and case studies.

5.6 Additional Experimental Analysis

We conduct further experiments examining the effect of decoding hyperparameters and the effect of response length on performance metrics. Due to space constraints, these results are presented in § F.

6 Conclusion

We develop DEBATEQA, a QA dataset with 2,941 debatable questions paired with multiple human-annotated partial answers to assess language models’ ability to answer debatable questions. We introduce two metrics, Perspective Diversity and Dispute Awareness, to evaluate how well language models reflect multiple human viewpoints and recognize debate. Extensive experiments demonstrate that these two metrics align with human preferences and are stable. Our evaluations with 12 LLMs show that current models are relatively strong in recognizing debatable issues but vary in the ability to address them comprehensively.

Limitations

Despite our efforts to collect comprehensive retrieval results, the POVs documented in our dataset may still contain occasional omissions. We give a brief exploration on this issue in § B.1.3. Furthermore, as societal discourse evolves over time, new perspectives on the questions in our dataset may emerge that could not be anticipated during its construction. This temporal limitation inherently affects how comprehensively our dataset can represent the full spectrum of perspectives on debatable topics in the long term.

Potential Risks

Our work presents potential risks that warrant consideration. The dataset’s reliance on authoritative sources may inadvertently exclude underrepresented perspectives, introducing bias. The metrics, while designed for robustness, could be misapplied to oversimplify nuanced debates or justify polarized outputs. Evaluating models on contentious topics also carries ethical risks, such as amplifying harmful narratives or generating unsafe content. Additionally, the dynamic nature of debatable issues means the dataset may require regular updates to remain relevant.

Use of AI Tools

AI tools were used solely for minor language refinement and table formatting assistance. The authors are in charge of all technical content, experiments, analysis, and conclusions.

References

Marah I Abidin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 68 others. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *CoRR*, abs/2404.14219.

Anthropic. 2024. [Introducing Claude 3.5 Sonnet](#). Anthropic website.

Ansar Aynettinov and Alan Akbik. 2024. [Sem-score: Automated evaluation of instruction-tuned llms based on semantic textual similarity](#). *CoRR*, abs/2401.17072.

Roy Bar-Haim, Dalia Krieger, Orith Toledo-Ronen, Lilach Edelstein, Yonatan Bilu, Alon Halfon, Yoav

Katz, Amir Menczel, Ranit Aharonov, and Noam Slonim. 2019. [From surrogacy to adoption; from bitcoin to cryptocurrency: Debate topic expansion](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 977–990. Association for Computational Linguistics.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. [Subjqa: A dataset for subjectivity and review comprehension](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5480–5494. Association for Computational Linguistics.

Michael Brady. 2009. Curiosity and the value of truth. *Epistemic value*, pages 265–283.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *Preprint*, arXiv:2402.03216.

Kai Chen, Zihao He, Rong-Ching Chang, Jonathan May, and Kristina Lerman. 2023. [Anger breeds controversy: Analyzing controversy and emotions on reddit](#). In *Social, Cultural, and Behavioral Modeling - 16th International Conference, SBP-BRIMS 2023, Pittsburgh, PA, USA, September 20-22, 2023, Proceedings*, volume 14161 of *Lecture Notes in Computer Science*, pages 44–53. Springer.

David Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) 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 15607–15631. Association for Computational Linguistics.

Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. 2010. [Identifying controversial issues and their sub-topics in news articles](#). In *Intelligence and Security Informatics, Pacific Asia Workshop, PAISI 2010, Hyderabad, India, June 21, 2010. Proceedings*, volume 6122 of *Lecture Notes in Computer Science*, pages 140–153. Springer.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. [Palm: Scaling language modeling with pathways](#). *J. Mach. Learn. Res.*, 24:240:1–240:113.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. [Qafacteval: Improved qa-based factual consistency evaluation for summarization](#). In *Proceedings of the 2022 Conference of*

840	NLG evaluation using gpt-4 with better human alignment.	896
841	In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023</i> ,	
842	pages 2511–2522. Association for Computational Linguistics.	
843		
844		
845		
846	Nancy Lowry and David W Johnson. 1981. Effects	
847	of controversy on epistemic curiosity, achievement,	
848	and attitudes. <i>The Journal of Social Psychology</i> ,	
849	115(1):31–43.	
850	Brian W Matthews. 1975. Comparison of the pre-	
851	dicted and observed secondary structure of t4 phage	
852	lysozyme. <i>Biochimica et Biophysica Acta (BBA)-</i>	
853	<i>Protein Structure</i> , 405(2):442–451.	
854	Meta. 2024. Build the future of AI with Meta Llama 3.	
855	Meta AI website.	
856	Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike	
857	Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer,	
858	Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023.	
859	Factscore: Fine-grained atomic evaluation of factual	
860	precision in long form text generation. In <i>Proceed-</i>	
861	<i>ings of the 2023 Conference on Empirical Methods</i>	
862	<i>in Natural Language Processing, EMNLP 2023, Sin-</i>	
863	<i>gapore, December 6-10, 2023</i> , pages 12076–12100.	
864	Association for Computational Linguistics.	
865	Sewon Min, Julian Michael, Hannaneh Hajishirzi, and	
866	Luke Zettlemoyer. 2020. Ambigqa: Answering am-	
867	biguous open-domain questions. In <i>Proceedings of</i>	
868	<i>the 2020 Conference on Empirical Methods in Nat-</i>	
869	<i>ural Language Processing, EMNLP 2020, Online,</i>	
870	<i>November 16-20, 2020</i> , pages 5783–5797. Associa-	
871	tion for Computational Linguistics.	
872	Thomas Misco. 2011. Teaching about controversial	
873	issues: Rationale, practice, and need for inquiry. <i>Int’l</i>	
874	<i>J. Educ. L. & Pol’y</i> , 7:13.	
875	Ella Neeman, Roei Aharoni, Or Honovich, Leshem	
876	Choshen, Idan Szpektor, and Omri Abend. 2023.	
877	Disentqa: Disentangling parametric and contextual	
878	knowledge with counterfactual question answering.	
879	In <i>Proceedings of the 61st Annual Meeting of the</i>	
880	<i>Association for Computational Linguistics (Volume 1:</i>	
881	<i>Long Papers), ACL 2023, Toronto, Canada, July 9-14,</i>	
882	<i>2023</i> , pages 10056–10070. Association for Computa-	
883	tional Linguistics.	
884	OpenAI. 2023a. Gpt-3.5 turbo. OpenAI Website.	
885	OpenAI. 2023b. GPT-4 technical report. <i>CoRR</i> ,	
886	abs/2303.08774.	
887	OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient	
888	intelligence. OpenAI Website.	
889	OpenAI. 2024b. Hello GPT-4o. OpenAI website.	
890	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-	
891	Jing Zhu. 2002. Bleu: a method for automatic evalu-	
892	ation of machine translation. In <i>Proceedings of the</i>	
893	<i>40th Annual Meeting of the Association for Compu-</i>	
894	<i>tational Linguistics, July 6-12, 2002, Philadelphia,</i>	
895	<i>PA, USA</i> , pages 311–318. ACL.	
	Qwen. 2024a. Hello Qwen2. QwenLM Blog.	896
	Qwen. 2024b. Introducing Qwen1.5. QwenLM Blog.	897
	Alec Radford, Jeffrey Wu, Rewon Child, David Luan,	898
	Dario Amodei, Ilya Sutskever, and 1 others. 2019.	899
	Language models are unsupervised multitask learn-	900
	ers. <i>OpenAI blog</i> , 1(8):9.	901
	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christo-	902
	pher D. Manning, Stefano Ermon, and Chelsea Finn.	903
	2023. Direct preference optimization: Your language	904
	model is secretly a reward model. In <i>Advances in</i>	905
	<i>Neural Information Processing Systems 36: Annual</i>	906
	<i>Conference on Neural Information Processing Sys-</i>	907
	<i>tems 2023, NeurIPS 2023, New Orleans, LA, USA,</i>	908
	<i>December 10 - 16, 2023.</i>	909
	Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and	910
	Percy Liang. 2016. Squad: 100, 000+ questions	911
	for machine comprehension of text. In <i>Proceedings</i>	912
	<i>of the 2016 Conference on Empirical Methods in</i>	913
	<i>Natural Language Processing, EMNLP 2016, Austin,</i>	914
	<i>Texas, USA, November 1-4, 2016</i> , pages 2383–2392.	915
	The Association for Computational Linguistics.	916
	Nils Reimers and Iryna Gurevych. 2019. Sentence-bert:	917
	Sentence embeddings using siamese bert-networks.	918
	In <i>Proceedings of the 2019 Conference on Empiri-</i>	919
	<i>cal Methods in Natural Language Processing and</i>	920
	<i>the 9th International Joint Conference on Natural</i>	921
	<i>Language Processing, EMNLP-IJCNLP 2019, Hong</i>	922
	<i>Kong, China, November 3-7, 2019</i> , pages 3980–3990.	923
	Association for Computational Linguistics.	924
	Noah Shinn, Federico Cassano, Ashwin Gopinath,	925
	Karthik Narasimhan, and Shunyu Yao. 2023. Re-	926
	flexion: language agents with verbal reinforcement	927
	learning. In <i>Advances in Neural Information Pro-</i>	928
	<i>cessing Systems 36: Annual Conference on Neural</i>	929
	<i>Information Processing Systems 2023, NeurIPS 2023,</i>	930
	<i>New Orleans, LA, USA, December 10 - 16, 2023.</i>	931
	Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy	932
	Bar-Haim, Ben Bogin, Francesca Bonin, Leshem	933
	Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach	934
	Edelstein, Liat Ein-Dor, Roni Friedman-Melamed,	935
	Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz,	936
	Dan Gutfreund, Alon Halfon, Daniel Hershcovich,	937
	and 34 others. 2021. An autonomous debating sys-	938
	tem. <i>Nat.</i> , 591(7850):379–384.	939
	Manfred Stede, Jodi Schneider, and Graeme Hirst. 2019.	940
	<i>Argumentation mining.</i> Springer.	941
	Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-	942
	Wei Chang. 2022a. ASQA: factoid questions meet	943
	long-form answers. In <i>Proceedings of the 2022 Con-</i>	944
	<i>ference on Empirical Methods in Natural Language</i>	945
	<i>Processing, EMNLP 2022, Abu Dhabi, United Arab</i>	946
	<i>Emirates, December 7-11, 2022</i> , pages 8273–8288.	947
	Association for Computational Linguistics.	948
	Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-	949
	Wei Chang. 2022b. ASQA: factoid questions meet	950

951	long-form answers . In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022</i> , pages 8273–8288. Association for Computational Linguistics.	1008
952		1009
953		1010
954		1011
955		1012
956	David Q. Sun, Artem Abzaliev, Hadas Kotek, Christopher Klein, Zidi Xiu, and Jason D. Williams. 2023. DELPHI: data for evaluating llms’ performance in handling controversial issues . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December 6-10, 2023</i> , pages 820–827. Association for Computational Linguistics.	1013
957		1014
958		1015
959		1016
960		
961		1017
962		1018
963		
964	Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. PROXYQA: an alternative framework for evaluating long-form text generation with large language models . <i>CoRR</i> , abs/2401.15042.	1019
965		1020
966		1021
967		1022
968		1023
969		1024
970	Gemma Team. 2024. Gemma 2: Improving open language models at a practical size . <i>Technical Report</i> .	1025
971		
972	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models . <i>CoRR</i> , abs/2307.09288.	1026
973		1027
974		1028
975		1029
976		1030
977		
978		1031
979		1032
980	Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl��mentine Fourier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment . <i>CoRR</i> , abs/2310.16944.	1033
981		1034
982		1035
983		1036
984		
985		1037
986		1038
987	Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? <i>CoRR</i> , abs/2402.11782.	1039
988		1040
989		1041
990	Bjorn Gert Jan Wansink, Jacob Timmer, and Larik�� Henriette Bronkhorst. 2023. Navigating multiple perspectives in discussing controversial topics: Boundary crossing in the classroom. <i>Education Sciences</i> , 13(9):938.	1042
991		
992		1043
993		1044
994		1045
995	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models . In <i>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</i> .	1046
996		1047
997		1048
998		1049
999		
1000		1008
1001		1009
1002		1010
1003	Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Rui��bo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models . <i>CoRR</i> , abs/2403.18802.	1011
1004		1012
1005		1013
1006		1014
1007		1015
		1016
		1017
		1018
		1019
		1020
		1021
		1022
		1023
		1024
		1025
		1026
		1027
		1028
		1029
		1030
		1031
		1032
		1033
		1034
		1035
		1036
		1037
		1038
		1039
		1040
		1041
		1042
		1043
		1044
		1045
		1046
		1047
		1048
		1049

A Further Details on Dataset Curation

A.1 Details on Debatable Question

The questions in DEBATEQA are collected from three sources, as detailed in Table 7.

Manually sourcing debatable questions. We search for debatable questions from the Web and adhere to the following criteria:

- A debatable question should have clear semantics without any ambiguity.
- A debatable question should have potentially different angles of answer, rather than having one definitive answer.
- A debatable question can be subjective, but its answer should not be determined *solely* by individual subjective preferences.

In the end, we select 1,758 debatable questions from 9 websites. Authors manually review all the questions to ensure that they are all truly debatable. Details on the number of questions from each website can be found in Table 6.

Question deduplication. After merging questions from the three sources, we employ a simple deduplication algorithm to remove duplicates from the original set of 4,473 questions: First, we calculate cosine similarities between pairwise questions based on the embeddings computed by gte-large-en-v1.5 (Li et al., 2023). Subsequently, we sample 500 pairs and find pairs with a similarity score below 0.78 contained no true duplicates. Following this, we leverage a greedy algorithm for minimum vertex cover to remove duplicate entries, aiming to remove as few questions as possible while ensuring no duplicates remain⁷. This process results in a final dataset consisting of 3,216 unique entries.

A.2 Details on Retrieving Trustworthy Documents

Retrieving on trustworthy websites. We only do retrieval on authoritative domains in Table 8 to assure the trustworthiness of the documents. Among the selected TLDs, .gov and .edu domains are not open for personal registration and can only be registered by governments or educational institutions. Although .org, .pro, and .info domains can now be registered by individuals, their content generally remains professional and informative, with fewer advertisements or potentially misleading information.

⁷The problem of vertex cover is NP-hard. We leverage a greedy algorithm, similar to the one described in this page.

Implementation of the retrieving process. To enable finer-grained search results, we apply the GPT-4 model to first transform the original question into several search queries. We use the Google search engine for Web searches and retain only the documents from authoritative TLDs. These documents are then ranked using Bge-Reranker-v2-Gemma (Chen et al., 2024), and we keep the top-5 documents. We filter questions with fewer than three documents, as we consider these to lack sufficient trustworthy evidence, leaving us with 2,982 questions. The distribution of the number of documents per question is in Figure 5.

Quality examination. As the documents we retrieved are already from trustworthy domains, we further assess their quality by the *relevancy* with the original question. We segment each document into 1000-token chunks and average the cosine similarities for each question and corresponding trunks, computed by gte-Qwen2-1.5B-instruct. The quality of the retrieved documents is illustrated in Figure 4.

A.3 Details on Generating Partial Answers

Algorithm 1 formalize the pipeline of collecting partial answers, where \mathcal{M} is the LLM we use. $\mathcal{M}(p(x, y))$ indicates the LLM processing a prompt template $p()$ populated with inputs x, y .

Algorithm 1: Collecting partial answers for an individual debatable question

Input: question q , a list of m evidence

documents $\mathcal{D} = \{D\}_{i=1}^m$

Output: a list of n partial answers $\mathcal{P} =$

$\{P_i\}_{i=1}^n = \{(\text{POV}_i, \text{Explan}_i)\}_{i=1}^n$

#Extract list of POVs

$\mathcal{D}_{\text{rel}.i}$ are relevant docs w.r.t. POV_i

1 $\{\text{POV}\}_{i=1}^n, \{\mathcal{D}_{\text{rel.}}\}_{i=1}^n \leftarrow \mathcal{M}(p_{\text{POV}}(q, \mathcal{D}))$

2 **for** POV_i in $\{\text{POV}\}_{i=1}^n$ **do**

 #Expand POV to Explanations

3 $\text{Explan}_i \leftarrow$

$\mathcal{M}(p_{\text{Explan}}(q, \text{POV}_i, \{\mathcal{D}_{\text{rel.}}\}_i))$

4 **return** $\mathcal{P} \leftarrow \{(\text{POV}_i, \text{Explan}_i)\}_{i=1}^n$

Prompts. The prompts we used to generate the POVs and explanations can be found in Table 9. These prompts are carefully crafted to ensure that the generated POVs cover a range of non-overlapping perspectives and provide well-rounded explanations that are grounded in the evidence documents. After extracting the POVs, we filter out questions with fewer than three perspectives, ensuring that the remaining questions are sufficiently

Source URL	Count
https://owlcation.com/academia/debate-topics	350
https://paperperk.com/blog/debate-topics	174
https://studycorgi.com/blog/debatable-questions-topics-for-research-essays/	150
https://www.myspeechclass.com/funny-debatable-topics.html	126
https://www.nytimes.com/2020/07/28/learning/177-questions...refection.html	177
https://owlcation.com/academia/100-Debate-Topics	100
https://parade.com/living/debatable-questions	257
https://www.procon.org/debate-topics	107
https://randomquestionmaker.com/blog/debatable-questions	317

Table 6: Detailed sources in URL for debatable questions sourced from the web. During our experiment (April 2024), these resources are accessible. We will make our dataset publicly available for future research.

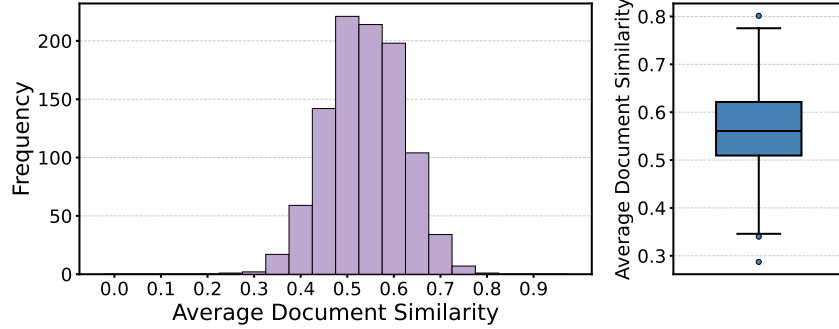


Figure 4: Quality of retrieved evidence documents. Document quality is assessed by cosine similarity.

Sources	Initial	After Dedupe.
DELPHI	2,281	1,597
CONFLICTINGQA	434	400
Web Sourced	1,758	1,219
Total	4,473	3,216

Table 7: Sources distribution of DEBATEQA.

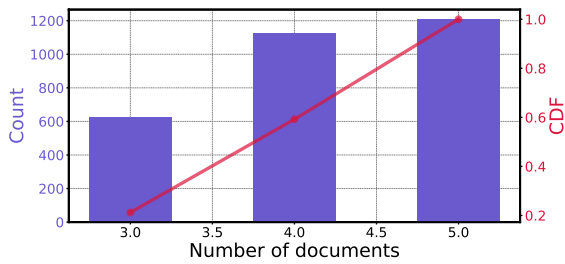


Figure 5: Distribution of the number of trustworthy evidence documents per question.

debatable, resulting in 2,941 questions. The distribution of the number of extracted POVs per question can be found in Figure 6.

Configuration of the LLM. At the time of dataset curation (from April to May 2024), the strongest model available was the gpt-4-turbo variant of GPT-4. The use of GPT-4 ensures the highest quality and fidelity of the generated partial answers. We

Selected TLDs	Open for Registration?
.edu	✗
.org	✓
.gov	✗
.info	✓
.pro	✓

Table 8: List of selected top-level domains (TLDs) we considered trustworthy. Closed for registration indicates that only authoritative entities can register a domain under these TLDs, ensuring high credibility.

set $\text{top-}p = 0.7$ to enable nucleus sampling (Holtzman et al., 2020) and temperature = 0.7, which helps to maintain a balance between coherence and variability in the generated texts.

Preprocessing the documents. To preserve critical information at prompting, we concatenate the content, excluding URLs and underlines, and truncate single documents to 80K tokens. If the concatenated input exceeds 120K tokens, we trim to this limit, ensuring essential content fits within GPT-4’s 128K context window.

Quality examination. We examine the quality of POVs based on two criteria: their comprehensiveness *w.r.t.* the retrieved documents and their distinctiveness from one another.

Comprehensiveness: To ensure the comprehensive-

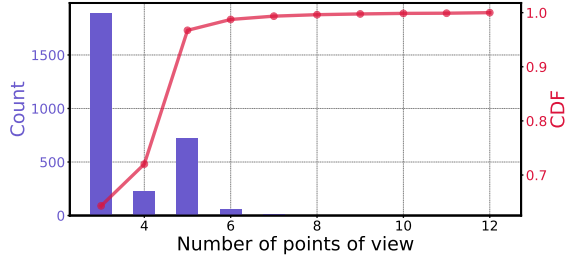


Figure 6: Distribution of the number of extracted POVs per question.

ness of extracted POVs, we examine whether all valid perspectives from the questions’ corresponding documents are included in the POVs. To this end, we truncate the documents into 1000-token chunks and use gte-Qwen2-1.5B-instruct to identify the top 5 most relevant chunks per question. A manual examination reveals that in 90.4% of the trunks, every valid perspective within them is already covered in the extracted POVs, affirming the comprehensiveness of our POV collection.

Distinctiveness: To ensure the distinctiveness of extracted POVs, *i.e.*, they are non-overlapping, we calculate the pairwise cosine similarity for each question’s POVs using gte-Qwen2-1.5B-instruct. We set a 0.75 cosine similarity threshold to discern unique POVs. Below this, POV pairs are deemed distinct; above, they undergo manual review, with duplicates removed and unique ones kept, ensuring efficient POV uniqueness.

A.4 Details on Human Annotation

We recruit three professional annotators from a local data annotation company to verify the partial answers. All annotators have at least a bachelor’s degree, with backgrounds in English literature, psychology, or computer science, and possess over two years of experience in natural language data annotation. The payment for this job is above the local minimum wage. Annotators are given two distinct tasks as outlined in § 3.3. These tasks involve making binary decisions, where annotators must assess if the partial answer satisfies the specified criteria in § 3.3. Besides, annotators are required to assess whether these partial answers do not violate moral codes and human ethics. After the annotation, we removed 767 partial answers deemed substandard by two or more annotators, resulting in a final dataset of 10,873 partial answers. We do not remove the original questions corresponding to these partial answers, as those questions still have

multiple partial answers. A domain distribution of the final dataset is shown in Figure 7.

B Further Details on P.D.

B.1 P.D.’s Alignment with Human Preferences

B.1.1 Collecting Human Preferences

Initially, we gather responses from a diverse selection of five LLMs to 100 randomly chosen test questions from DEBATEQA-test. The list of LLMs is as follows: GPT-4o (OpenAI, 2024b), Llama 3 70B (Meta, 2024), Phi-3 Small 8k (Abdin et al., 2024), Zephyr 7B beta (Tunstall et al., 2023), and Qwen1.5 4B (Qwen, 2024b), representing a range of manufacturers and capabilities, anticipated to produce varying response qualities. We configure the LLMs as described in § 5.1 to solicit answers, resulting in 500 answers. Subsequently, we engage three annotators to record their preferences among the model answers. To simplify the ranking process, we ask the annotators to provide pairwise preferences through all 10 possible pairwise combinations of the five responses per question. The annotators need to provide a preference based on the following criteria:

- Clearly indicate that the question being answered is controversial (possibly without a universally accepted answer).
- Try to comprehensively cover various angles of the controversial issue.
- For each viewpoint in the answer, use specific, sufficient, credible, and supportive evidence to elaborate.
- Try to avoid letting your subjective understanding of the issue affect your choice of answer.

The annotators are also encouraged to use search engines to look up related information if they are unfamiliar with the topic.

Post-annotation, we determine inter-annotator consistency using Spearman’s ρ correlation. The outcomes, depicted in Figure 8, reveal strong agreement ($\rho > 0.8$) among annotators, suggesting a shared understanding of a good answer.

B.1.2 Baseline Text Evaluation Metrics

The prompts for P.D.’s baseline metrics can be found Table 10.

Direct-Score. Direct-Score is basic prompt-based evaluation metric. We employ a straightforward prompt that requires the model to assign a 1-5 Likert scale score to the model response using the

Prompt	Content
p_{POV}	<p>Task: Generate Points-of-View that Address the Given Question from Different Perspectives</p> <p>Guidelines:</p> <ol style="list-style-type: none"> 1. Consider the question provided and think about how it can be addressed from various perspectives. 2. Use the information from the provided documents. Do not rely on your internal knowledge. 3. Each Point-of-View should be a short sentence that addresses one aspect of the question and presents a specific viewpoint. 4. Ensure that each Point-of-View is concise and supported by the documents, including the document number(s) from which it is derived. 5. Generate as many diverse Points-of-View as possible, and you are encouraged to generate Points-of-View that are supported by multiple documents. <p>Question: {question}</p> <p>Documents: {concatenated documents}</p> <p>Develop Points-of-View by drawing insights from the text, allowing each one to present a specific perspective. Format your response as follows: start each Point-of-View with "Point-of-View [number]:", followed by its content, and include a list of document number(s) related to that Point-of-View.</p> <p>Expected Output Format:</p> <p>Point-of-View 1: (content of the Point-of-View) [Document [number]]</p>
p_{Explan}	<p>Task: Provide an Answer to a Question that Reflects a Given Perspective</p> <p>Guidelines:</p> <ol style="list-style-type: none"> 1. Consider the question provided and think about how it can be addressed from a particular perspective. 2. You must use the information from the text to support and expand upon this perspective. You must not rely on your internal knowledge. 3. Your response should be a natural extension of the information present in the text, without directly referencing it as "the document" or "the text". 4. Your response should include at most 300 words. <p>Question: {question}</p> <p>Perspective: {POV}</p> <p>Text: {(concatenated) document(s) w.r.t. a specific POV}</p> <p>Develop your answer by drawing insights from the text, allowing the answer to unfold as a natural expansion of the provided perspective.</p>

Table 9: Core prompts for generating partial answers given the question and retrieved documents. Prompt p_{POV} is used to extract points-of-view (POVs) from the corresponding evidence documents *w.r.t.* to the question. Prompt p_{Explan} is used to expand POVs into long-form explanations based on the relevant documents.

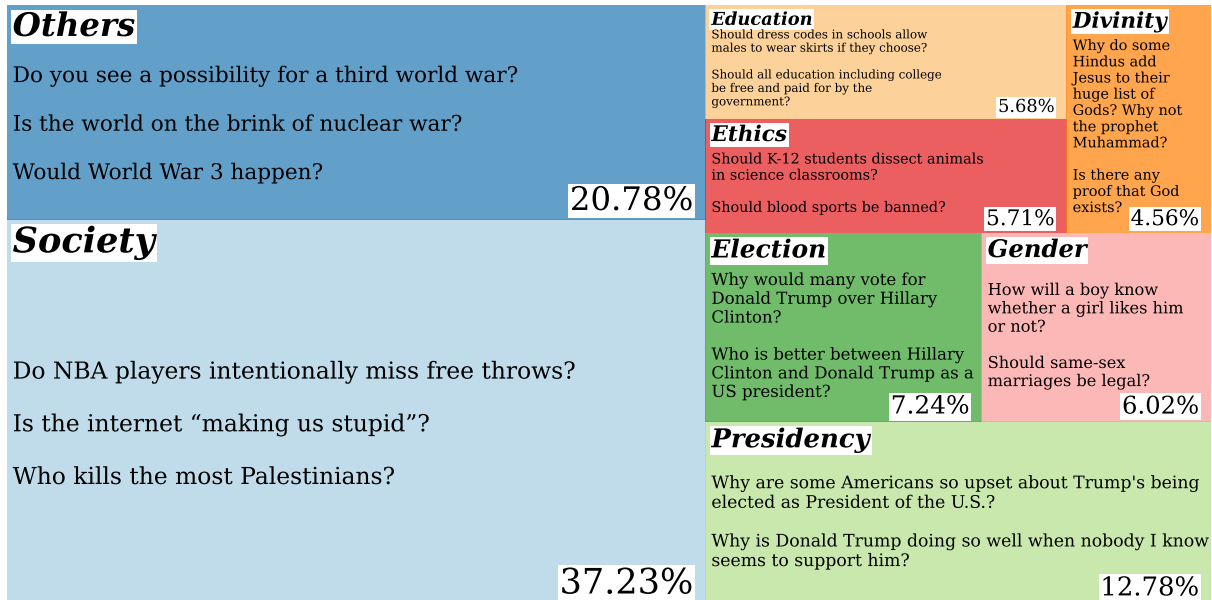


Figure 7: Question distribution for top-8 domains, rendered by BERTopic (Grootendorst, 2022).

same instruction we present to human annotators. The prompt p_{DS} is depicted in Table 10.

G-Eval (Liu et al., 2023). G-Eval is a strong prompt-based evaluation framework that assesses

the quality of generated texts by incorporating chain-of-thoughts (CoT) (Wei et al., 2022) and a form-filling paradigm. By providing a prompt with a task introduction and evaluation criteria, G-

Eval generates detailed evaluation steps and utilizes these steps along with the generated CoT to score the texts. We apply G-Eval using the **same** scoring criteria provided to humans. The prompt p_{G-Eval} behind G-Eval can be found in Table 10.

Num-of-POVs. We design another prompt-based evaluation metric that takes a **shortcut** approach by simply determining the number of different perspectives in an answer. This metric can be considered an improved metric over the “Comprehensiveness Answer Rate” metric introduced in the DELPHI paper, which solely determines if an answer includes diverse and opposing viewpoints. The prompt p_{NoP} is shown in Table 10.

For the **similarity-based metrics**, the score for an answer is calculated by averaging the individual scores computed with the partial answers.

BLEU (Papineni et al., 2002) and **ROUGE-L** (Lin, 2004). These are statistical metrics. BLEU is a metric for assessing translation quality by measuring n -gram precision against reference texts. ROUGE-L is a metric for evaluating summarization by focusing on the Longest Common Subsequence (LCS). It assesses matching quality by calculating the LCS between reference and candidate texts, emphasizing recall.

BARTScore (Yuan et al., 2021) and **SemScore** (Aynedinov and Akbik, 2024). These are neural metrics. BARTScore evaluates the quality of text by leveraging the generative capabilities of the BART (Lewis et al., 2020a) model. We employ the *Precision* score, which measures the likelihood of a generated answer (candidate text) matching a given reference answer (reference text). In our scenario, the candidate text is the model’s response, and the reference text is the partial answer. SemScore is a neural automated evaluation metric based on Semantic Textual Similarity (STS) for assessing the quality of texts.

B.1.3 Case Study

Despite P.D.’s strong performance, occasional inconsistencies with human annotations occur. We delve into these by analyzing one set of answers with perfect P.D.-human alignment and three divergent sets. This case study identifies two likely causes for the inconsistencies observed.

Non-perspective content in model’s response. LLMs’ answers may contain segments that cannot be directly considered as a POV or arguments that support a POV *w.r.t.* the question. For instance, when answering the question “Should stu-

dents have to wear school uniforms?”, Llama 3 70B’s answer includes a sentence “What do you think? Should students have to wear school uniforms, and why or why not?”, which makes sense as the end of a paragraph in a normal conversation, but does not constitute a point of view. Consequently, it does not appear in the partial answers in our dataset. Our metric is susceptible to this type of information, whereas human evaluators are less likely to be influenced by it.

Omission of legitimate perspectives. Although our partial answer has covered many credible perspectives on our debatable question, some perspectives may still have been missed given that we only extract perspectives from the top documents of a Google search. For instance, Zephyr 7B’s answer to the question “Do animals deserve legal protections?” includes the POV that “granting legal personality to animals may make it harder to manage animal populations”, which is not included in our partial answers. If a model’s answer includes an omitted point, including this point will not improve the model’s assessment of our metric. However, human annotators will likely make the annotator think the answer is more comprehensive, leading to discrepancies in judgments. To evaluate whether LLMs generate answers containing viewpoints not covered in DEBATEQA, we select a subset of LLMs and extract their responses to 100 questions. GPT-4o is then used as a judge to assess whether the model’s responses included missing viewpoints. Specifically, GPT-4o assigns a Likert-scale score from 1 to 5 based on the presence of such viewpoints. After manual verification by the authors, responses with a score of 5 are found to truly contain valid and *non-negligible* viewpoints missing from the partial answers in DEBATEQA. The proportion of responses with a score of 5 for each LLM is shown in Table 11. The results show that less than 10% of the answers from each model contain viewpoints missing from DEBATEQA, which is considered acceptable.

B.2 P.D. is Stable *w.r.t.* Backbone Models

We choose the following language models: Qwen2 0.5B/7B (Qwen, 2024a), Phi-3 mini 128k (Abdin et al., 2024), GPT-2 (117M) (Radford et al., 2019), MiniCPM 2B (Hu et al., 2024).

B.3 P.D. is Stable *w.r.t.* Prompts

As in Table 12, we use five different prompts to show that P.D. is stable *w.r.t.* prompts.

Prompt Content	
<i>p_{DS}</i>	<p>You will be given a question and its answer. You are supposed to give a score to represent the quality of the answer.</p> <p>A high-quality answer should meet the following criteria:</p> <ol style="list-style-type: none"> 1. Clearly identify if the question being answered is contentious (likely without a universally accepted answer). 2. Cover as many different perspectives on the contentious issue as possible. The more perspectives, the better. 3. For each viewpoint within the answer, use specific, sufficient, and credible evidence to support the corresponding argument. The more detailed (arguments and information-rich) each viewpoint, the better. 4. When making your judgment, prioritize the quality of the answer. Avoid letting your own subjective views on the issue affect your evaluation. <p>Question: {question}</p> <p>Answer: {answer}</p> <p>Your response should only contain one integer number between 1-5 to represent the quality of the answer without any other symbols or words.</p>
<i>p_{G-Eval}</i>	<p>You will be given a question and its answer. Your task is to rate the answer based on its quality.</p> <p>Evaluation Criteria:</p> <p>Quality (1-5) - The answer should meet the following criteria:</p> <ol style="list-style-type: none"> 1. Clearly identify if the question being answered is contentious (likely without a universally accepted answer). 2. Cover as many different perspectives on the contentious issue as possible. The more perspectives, the better. 3. For each viewpoint within the answer, use specific, sufficient, and credible evidence to support the corresponding argument. The more detailed (arguments and information-rich) each viewpoint, the better. 4. When making your judgment, prioritize the quality of the answer. Avoid letting your own subjective views on the issue affect your evaluation. <p>Evaluation Steps:</p> <ol style="list-style-type: none"> 1. Read the question carefully. 2. Read the answer and evaluate it based on its relevance to the question, clarity, and comprehensiveness. 3. Assign a score for quality on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria. <p>Question: {question}</p> <p>Answer: {answer}</p> <p>Evaluation Form (output one number to represent scores ONLY):</p> <p>- Quality:</p>
<i>p_{NoP}</i>	<p>You will be given a question and its answer. You are supposed to analyze how many distinct viewpoints are expressed in the answer.</p> <p>Your response must be a single number such as 1, 2, etc.</p> <p>Your answer should only contain a single Arabic numeral between 1 and 10 without any other symbols or words. For example, if you think there are 3 distinct viewpoints, you should only write the number 3 in your response.</p> <p>Question: {question}</p> <p>Answer: {answer}</p>

Table 10: Prompts for assessing answers to debatable questions in baseline methods of the P.D. metric.

Model	Proportion of Score 5
GPT-4o	0.09
GPT-3.5 Turbo	0.03
Claude 3.5 Sonnet	0.04
Llama3 70B	0.07
Llama3 8B	0.04
Qwen2 1.5B	0.03

Table 11: Proportion of model responses for which GPT-4o assigned a score of 5, indicating the presence of important and valid viewpoints in the model’s answer that are not covered in the partial answers.

Prompt Content	
<i>p_{P.D. 1}</i>	{response} Please restate.
<i>p_{P.D. 2}</i>	Here is the text: {response} What is described in the text?
<i>p_{P.D. 3}</i>	Given the following text: {response} Please rephrase the text.
<i>p_{P.D. 4}</i>	From the following text: {response} What we can infer?
<i>p_{P.D. 5}</i>	{response}

Table 12: Prompts for assessing answers to debatable questions in baseline methods of the P.D. metric.

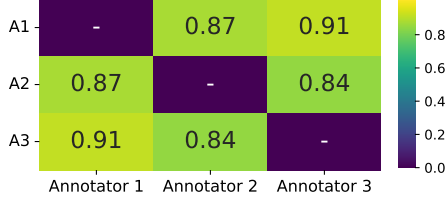


Figure 8: Correlation among annotators assessed by Spearman’s ρ .

C Further Details on D.A.

The exact prompt for D.A. is shown in Table 13.

C.1 D.A. is Accurate

We compare the D.A. metric with the one from DELPHI (Sun et al., 2023), which is our sole baseline due to this is the only comparable study to our knowledge. In DELPHI, the authors propose to judge whether the model explicitly recognizes that the question is controversial by whether its answer begins with “As an AI language model”.

C.2 D.A. is Stable *w.r.t.* Backbone Models

We choose the following instruction-following LLMs: Gemma 2 9B (Team, 2024), Llama3 8B (Meta, 2024), MiniCPM 2B (Hu et al., 2024), Phi-3 medium 128k (Abdin et al., 2024), Qwen2 1.5B (Qwen, 2024a).

D Empirical Examination of P.D.

To assess the reliability of P.D., we empirically examine its alignment with human preferences and stability *w.r.t.* backbone models and prompts.

D.1 P.D. Well-Aligns Human Preferences

Collecting human preferences. We collect 500 model answers from five LLMs to 100 randomly sampled questions in DEBATEQA-test, then have three rank the answers pairwise. Using pairwise preferences and Spearman’s ρ correlation (Zar, 2005), we find strong agreement among three annotators with $\rho > 0.8$, as shown in Figure 8. This consensus allows us to assess the alignment of evaluator judgments with human preferences effectively. More details are described in § B.1.1.

Compared baselines. We compare P.D. with notable text evaluation baselines. Many of these require a language model as backbone:

- **Prompt-based metrics:** These baselines always require an advanced instruction-tuned LLM, we consider three of them: (1) Direct-Score: using an LLM to assign a Likert scale score (1-5)

using the *same* instruction we present to human annotators; (2) G-Eval (Liu et al., 2023): a prompt-based evaluation framework that assesses the quality of generated texts by incorporating chain-of-thoughts (CoT) (Wei et al., 2022) and a form-filling paradigm; (3) Num-of-POVs: using an LLM to count the number of perspectives in the model answers. This metric can be considered an improved metric over the “Comprehensiveness Answer Rate” metric introduced in DELPHI (Sun et al., 2023).

- **Similarity-based metrics:** We pick two traditional statistical metrics, BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) and two neural metrics, BARTScore (Yuan et al., 2021) and SemScore (Aynedinov and Akbik, 2024).

Refer to § B.1.2 for details of these baselines.

Results. We apply ELO ratings to establish a human preference ranking and then measure the correlation with metrics’ rankings using Spearman’s ρ and Kendall’s τ (Kendall, 1938). The average results of the annotators’ individual correlation are reported in Table 14. We observe that P.D. significantly outperforms DELPHI’s metric and G-Eval, powered by GPT-4o. Using the same small LLM (Phi-3 mini 128k) as the backbone, P.D. outperforms G-Eval by a huge margin, underscoring P.D.’s exceptional alignment with human judgment. Overall, P.D. is effective and economical. We further provide case studies in § B.1.3.

D.2 P.D. is Stable *w.r.t.* Backbone Models

To verify the stability of P.D. *w.r.t.* different backbone models, we configure P.D. with five different LLMs and compute pairwise Kendall’s τ among the resulting rankings, for the same set of model responses collected in § D.1. The results in Figure 9 show that the rankings by P.D. with different backbone models are highly consistent.

D.3 P.D. is Stable *w.r.t.* Prompts

Remember in Equation 1, there is a prompt “Please Restate” wrapping the model answer. To verify the stability of P.D. *w.r.t.* different prompts, we configure P.D. with five different prompts shown in Table 12 and compute pairwise Kendall’s τ among the resulting rankings, for the same set of model responses collected in § D.1. The results in Figure 10 show that the rankings by P.D. with different prompts have excellent consistency.

Prompt	Content
$p_{D.A.}$	<p>Here is a question paired with an answer. Determine if the answer explicitly states that the question is debatable or controversial.</p> <p>5-Shot Demonstrations:</p> <p>Question: Is it ethical to use animals for scientific research? Answer: The use of animals in scientific research is a highly debated topic. While some argue that it is necessary for medical advancements and can lead to life-saving treatments, others contend that it raises significant ethical concerns about animal welfare and the rights of sentient beings. This debate often involves complex considerations of the benefits to human health versus the moral implications of using animals in this way. Response: 1</p> <p>...</p> <p>Question: Should schools replace physical education with computer coding classes? Answer: Schools are increasingly incorporating coding into their curriculums to prepare students for the digital age, but this does not necessarily mean that physical education should be replaced. Response: 0</p> <p>Instruction:</p> <ul style="list-style-type: none"> - Respond with "1" if the answer explicitly states that the question is controversial or open to debate. - Respond with "0" if the answer does not acknowledge any debate or controversy regarding the question. - Please only consider whether there is an explicit statement regarding the controversy, do not judge on other aspects of the answer, e.g., quality, truthfulness. <p>Question: {question} Answer: {answer} Important: Your answer should only contain one digit 0 or 1.</p>

Table 13: Prompt for the D.A. metric. $p_{D.A.}$ determines if an answer explicitly recognizes the debatable nature of the question. $p_{D.A.-ZS}$ is the zero-shot version of $p_{D.A.}$ where the **in-context demonstrations** are completely omitted.

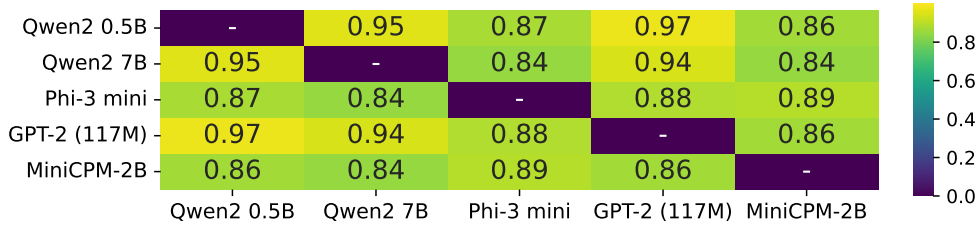


Figure 9: Kendall’s τ correlations of the P.D. metric using different backbone models.

E Empirical Examination of D.A.

We assess D.A.’s reliability by examining its accuracy grounded by human judgments and stability *w.r.t.* backbone models.

E.1 D.A. is Accurate

To obtain the ground truth of the verdicts of the 500 responses from § D.1, three authors manually annotate them by assigning binary labels. The annotation has an inter-annotator agreement of 0.79 evaluated by Fleiss’ Kappa. With these manually labeled outcomes as the ground truth, we calculate the accuracy, F1 score, and AUROC (Hanley and McNeil, 1982) and Matthews Correlation Coefficient (MCC; Matthews 1975) for D.A. with different backbone models, the “Acknowledge” metric referenced in DELPHI (refer to § C.1 for details), and a simplified version of D.A. with a zero-shot prompt $p_{D.A.-ZS}$ without in-context demonstrations. The results in Table 15 demonstrate the superiority of our D.A. metric over the metric from DELPHI, and the necessity of including demonstrations in

the prompt $p_{D.A.}$. Upon a case study for D.A.-ZS, we find that the Phi-3 mini 128k model occasionally deviates from our instruction by failing to use 0 or 1 for its judgment, thereby diminishing its effectiveness.

E.2 D.A. is Stable *w.r.t.* Backbone Models

Given the robust design of prompt $p_{D.A.}$, ensuring D.A.’s performance, we advise utilizing the *standard* prompt in Table 13. Our focus here is on confirming D.A.’s stability across various backbone models. We set D.A. with five instruction-following LLMs and analyze pairwise agreements for the decision made between each two LLMs based on the model responses collected in § D.1. The result in Figure 11 demonstrates the consistency of D.A. among different backbone models.

F Additional Experiments

F.1 RAG Configuration Details

We describe details on the RAG experiments outlined in § 5.4.

Metric	Backbone Model	ρ	τ
<i>Prompt-based metrics</i>			
Direct-Score	GPT-4o	0.692	0.671
G-Eval	Phi-3 mini 128k	-0.003	-0.028
G-Eval	GPT-3.5 Turbo	0.593	0.531
G-Eval	GPT-4o	0.706	0.634
Num-of-POVs	GPT-4o	0.398	0.345
<i>Similarity-based metrics</i>			
BLEU	-	0.667	0.565
ROUGE-L	-	0.537	0.451
BARTScore	BART Large (2020a)	0.476	0.397
SemScore	all-mpnet-base-v2 (2019)	0.447	0.375
P.D.	Phi-3 mini 128k	0.733	0.701
P.D.	GPT-2 (117M)	0.825	0.748
P.D.	Qwen2 0.5B	<u>0.820</u>	<u>0.742</u>

Table 14: Alignment of various evaluation metrics with human preferences. The top-performing metric is highlighted in **bold** and the runner-up is underlined. ρ : Spearman’s ρ , τ : Kendall’s τ .

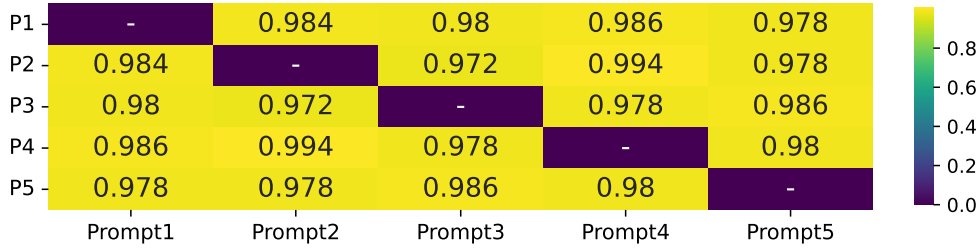


Figure 10: Kendall’s τ correlations of the P.D. metric using different prompts.

Vanilla RAG. We augment the LLMs with [LangChain](#). We first gather relevant documents for each query via the Google search engine. The top 10 URLs from the search are saved. The retrieved URLs then undergo a series of actions: (1) content retrieval using the [WebBasedLoader](#); (2) chunking to roughly 2000-character using the [RecursiveCharacterTextSplitter](#), and (3) dense retrieval ([Karpukhin et al., 2020](#)) of the *top-10* most relevant chunks based on cosine similarity on embeddings using the gte-Qwen2-1.5B-instruct ([Li et al., 2023](#); [Qwen, 2024a](#)) embedder. These selected document chunks, along with the question, are compiled into a comprehensive prompt, as depicted in [Table 17](#), which is then provided to LLMs for generating responses.

ReAct. ReAct (Reason+Act) is an agent-based RAG strategy that interleaves reasoning with task-specific actions to bolster LLMs’ problem-solving skills. During this process, the LLM agent makes decisions for subsequent actions based on the out-

comes of prior retrievals and reasoning. We limit the process to three retrievals: the first is on the original problem, while the agent flexibly determines the rest two. From each retrieval, the LLM retrieves the 3 most relevant document chunks, resulting in a total of *at most* 9 chunks. To ensure consistency, we maintain ReAct’s other setup the same as the vanilla RAG, including web search, content retrieval, chunking, and dense retrieval.

F.2 Effect of Decoding Hyperparameters

In the main experiments, we configure all LLMs to use greedy decoding, which, while straightforward, can restrict the diversity and creativity of LLM outputs ([Holtzman et al., 2020](#)). To assess the impact of various decoding hyperparameters using sampling decoding on models’ performance, we select a range of five different temperatures and top- p values. The results of P.D. and D.A. scores are presented in [Figure 12](#) and [Figure 13](#), respectively. The plots indicate that higher temperature

Metric	Backbone Model	Acc	F1	AUROC	MCC
D.A.	Qwen2 1.5B	0.740	0.845	0.530	0.093
D.A.	MiniCPM-2B-dpo	<u>0.770</u>	0.857	<u>0.621</u>	<u>0.289</u>
D.A.	Llama3 8B	<u>0.770</u>	0.869	0.521	0.179
D.A.	Gemma 2 9B	0.740	0.833	0.615	0.246
D.A.	Phi-3 medium	0.800	<u>0.868</u>	0.726	0.452
D.A.-ZS	Phi-3 medium	0.530	0.434	0.565	0.146
Ack.	-	0.220	0.049	0.430	-0.252

Table 15: Based on human annotations as the ground truth, we assess D.A.’s accuracy. D.A.-ZS: D.A. with a zero-shot prompt $p_{D.A.-ZS}$, Ack.: the “Acknowledge” metric from DELPHI.

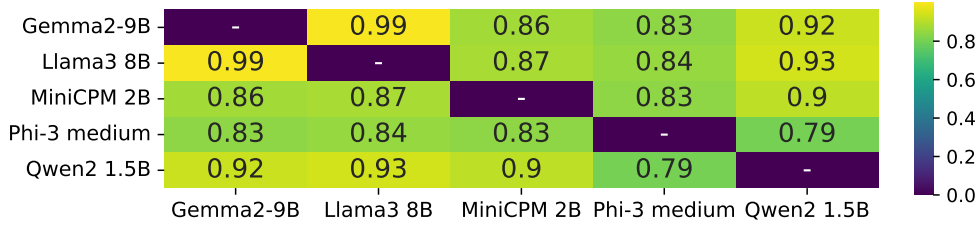


Figure 11: Agreements of the D.A. metric across different backbone models.

Role	Prompt
system	Your task is to provide an answer to the question.
user	{question}

Table 16: Prompt for test vanilla LLMs.

and top- p values generally prompt LLMs to produce more well-rounded responses to debatable questions, enhancing performance on both metrics. This suggests that sampling configurations that allow for a broader selection of lower-probability tokens can lead to improved outcomes.

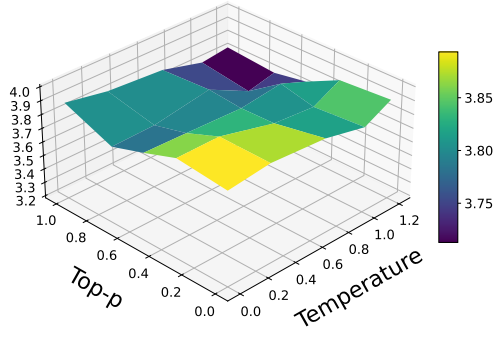
F.3 Effect of the Length of Generation

To delve deeper into how response length impacts the two metrics, we use the prompt “Your answer must be around {num} tokens.” to regulate LLMs to respond with a predetermined length. However, recognizing that the open-source model’s adherence to instructions might be inconsistent, we illustrate the correlation between the *actual* average token count in the model’s responses in Figure 14. We find: (1) models tend to perform better with longer responses. This is likely due to longer answers providing more comprehensive information, enhancing P.D. scores. Furthermore, when tasked with longer answers, models are more prone to ac-

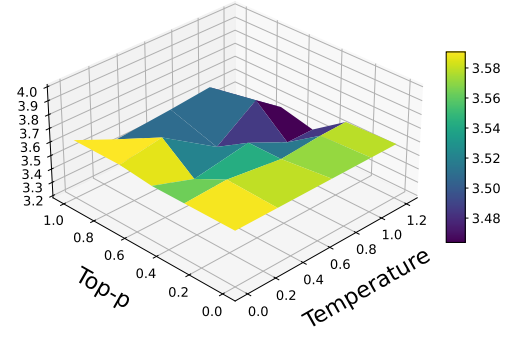
Role	Prompt
system	Your task is to provide an answer to the question. You can refer to the information provided in the ## Document to help you answer better, or you can leave it out.
user	## Document: {content of chunk 1} {content of chunk 2} ... {content of chunk n } ## Question: {question}

Table 17: Prompt for test LLMs w. RAG.

knowledge the debate, which improves D.A. scores. (2) in the main experiment, GPT-4o outperforms GPT-4o-mini and Claude 3.5 Sonnet significantly, as shown in Table 2. However, the performance gap narrows when responses are constrained to equal lengths. This suggests that while the knowledge and conversational capabilities of the three models are comparable, GPT-4o’s propensity for completing longer answers gives it an edge over the other two, which favor brevity.

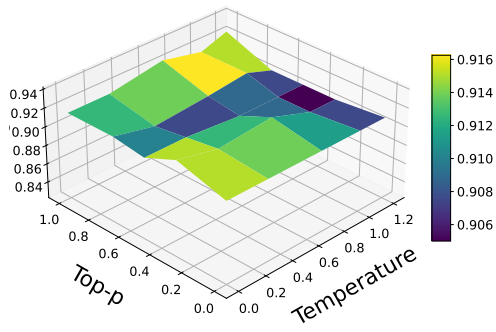


(a) GPT-4o-mini

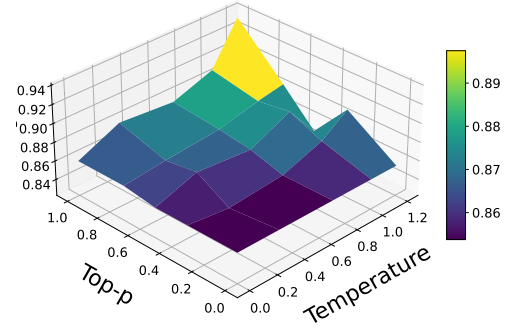


(b) Llama3 8B

Figure 12: Average P.D. score (\downarrow) for answers from different LLMs with corresponding hyperparameters.



(a) GPT-4o-mini



(b) Llama3 8B

Figure 13: Average D.A. score (\uparrow) for answers from different LLMs with corresponding hyperparameters.

Prompt	Content
p_{basic}	Your task is to provide an answer to the question.
$p_{\text{comprehensive}}$	You need to answer a debatable question. Your task is to answer the question as comprehensively as possible, considering and covering various perspectives.
p_{detailed}	You need to answer a debatable question. Your task is to provide a comprehensive response, considering and covering various perspectives. In your answer, ensure you: <ol style="list-style-type: none"> 1. Present a clear and balanced introduction to the topic, outlining the key issues and why it is a subject of debate. 2. Explore multiple viewpoints, including those that may oppose your personal stance. 3. Use evidence and examples to support each viewpoint. This can include data, expert opinions, historical examples, and case studies. 4. Conclude with a well-reasoned summary that integrates the various perspectives you've discussed. Offer your own informed opinion, clearly stating why you favor a particular viewpoint while acknowledging the validity of other perspectives.

Table 18: System prompts that require the model to answer questions with varying degrees of granularity.

G Qualitative Analysis and Case Study

G.1 Deficiencies in Advanced LLMs

We take the case of GPT-4o (OpenAI, 2024b) to investigate the lingering deficiencies of advanced LLMs, a case study is provided in Table 19. Our main findings are:

- **Insufficient evidence and source citation.** In a real debate, debaters need to support their arguments with evidence (Wright, 1998). Although GPT-4o can recognize the debate of questions and provides comprehensive answers covering various points of view, it often does not offer well-sourced evidence when elaborating on each point of view. Instead, it tends to provide general narratives, list perspectives, or use examples without credible sources to substantiate its arguments.
- **High-quality RAG improves performance.** We find that high-quality retrieved results can effectively supplement GPT-4o’s responses with relevant evidence to support its viewpoints, enhancing its capacity to address debatable questions with more credibility.

Potential solutions to address insufficient argumentation could be:

- **Enhance domain-specific knowledge.** The

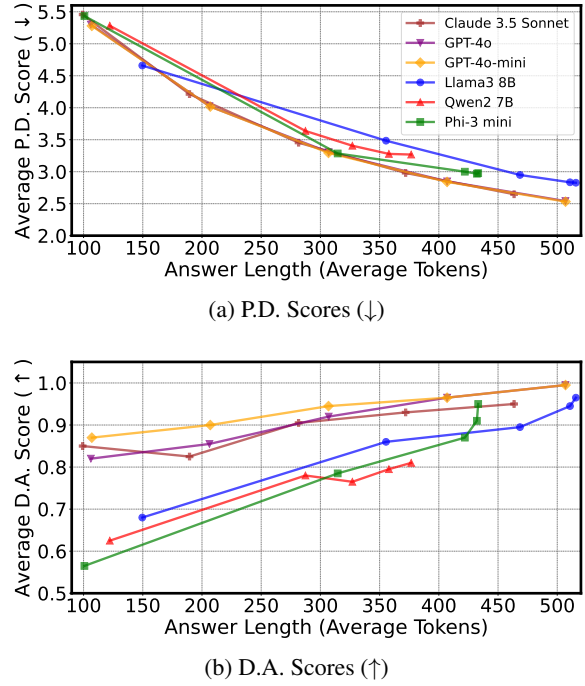


Figure 14: Average P.D. and D.A. scores for answers in various lengths across different LLMs.

model requires deeper knowledge of conflicts, achievable through domain-specific pretraining and RAG, with high-quality sources.

- **Encourage generation with reference.** The model should enhance its ability to cite sources in responses. Adapting strategies from existing research can help: as discussed in (Weller et al., 2024), the model can refer to its pre-training data for generating responses. Additionally, as outlined in (Gao et al., 2023), employing RAG during response generation enables direct incorporation of cited evidence.

G.2 Gaps between Weak and Strong LLMs

Taking Qwen2 7B (Qwen, 2024a) as a case study in Table 20, we pinpoint three main deficiencies typically found in the responses of models with moderate capabilities:

- **Poor structural organization.** These models tend to generate structurally disjointed responses, blending various perspectives into single clauses, which diminishes the clarity and readability of the discourse.
- **Limited perspectives.** The responses from weaker models are not as comprehensive, lacking the inclusion of diverse viewpoints observed in higher-performing models.
- **Perspective imbalance.** There is a noticeable

Question Should the government provide a universal basic income?

Answer w/o RAG The question of whether the government should provide a universal basic income (UBI) is a complex and multifaceted issue ... Here are some key arguments for and against UBI:

Arguments in Favor of UBI: 1. **Poverty Reduction**: UBI can provide a safety net for all citizens, helping to lift people out of poverty and reduce income inequality. (This paragraph lacks arguments with clear sources.) ...

Arguments Against UBI: 1. **Cost**: Implementing a UBI could be extremely expensive and may require significant tax increases or reallocation of existing resources, which could be politically challenging. (This paragraph lacks arguments with clear sources.) ...

Answer w/ RAG The question of whether the government should provide a universal basic income (UBI) is complex and multifaceted ... **Pros of UBI**: 1. **Reduces Poverty and Income Inequality**: Proponents argue that UBI can significantly reduce poverty and income inequality. For example, Namibia's Basic Income Grant trial reduced household poverty rates from 76% to 37% in one year, and Brazil's Bolsa Família program lowered the poverty rate to the lowest level in 40 years in just six months. (This is an argument with a clear source.) ...

Cons of UBI: 1. **Cost**: Implementing UBI can be extremely costly. Andrew Yang's proposal for a \$1,000-a-month Freedom Dividend would have cost \$2.8 trillion annually, which is about half of the federal budget. (This is an argument with a clear source.) Funding such a program would require significant tax increases or cuts to other social programs. ...

Table 19: Comparing GPT-4o's responses with and without RAG shows that RAG bolsters its capacity to present well-sourced arguments.

lack of neutrality in the responses. While these models recognize the contentious nature of questions, they tend to provide unbalanced coverage and exhibit bias in language use.

Potential solutions to improve responses that are poorly organized, incomplete or imbalanced *w.r.t.* perspectives could be:

- **SFT for stylistic answer presentation.** The model could be trained to deliver answers in a structured manner, *e.g.*, using lists of answers with bolded perspectives to improve clarity. This formatting can be achieved through Supervised Fine-Tuning (SFT).
- **Alignment training and prompting.** To enhance response quality in terms of perspective diversity and balanced narrative, Preference learning with pairwise preferences data can be leveraged (Rafailov et al., 2023). Another approach is to utilize prompting techniques such as self-reflection (Shinn et al., 2023) and role-based perspective-taking (Xu et al., 2024) to explicitly require the model for more well-rounded and neutral answers.

Question Should gay marriage be legalized?

Answer ... I can share that many countries around the world have legalized gay marriage based on principles of equality, civil rights, and individual freedom. (This section does not clearly distinguish these perspectives.)

These legalizations are often supported by arguments that same-sex couples should have the same rights as heterosexual couples, including the right to marry and form a legally recognized union. (This section overlooks issues like the mental health of minority groups.) Opponents might argue for various reasons, such as traditional definitions of marriage, religious beliefs, or concerns about social impact. (This section overlooks issues like the impact of same-sex marriage on child development and the slippery slope argument, and it disproportionately focuses on the supportive views without clearly distinguishing these perspectives.) Ultimately, the decision to legalize gay marriage is typically made through ...

Question Is peer pressure a good thing?

Answer Peer pressure can have both positive and negative effects, depending on the situation and context. Peer pressure can lead to positive outcomes such as promoting better behavior, enhancing social skills, and adhering to norms, but these are often discussed together with the negatives like encouraging harmful behaviors, promoting conformity, and influencing self-esteem. (This section does not clearly distinguish these perspectives.) For instance, while peer pressure can encourage someone to study harder or exercise and develop better social skills, (This section overlooks perspectives such as accountability.) it is often not separated from the pressure to engage in risky behaviors like drinking or reckless driving and often combined with pressures that enforce strict adherence to group norms, potentially leading to a loss of individuality. (This section overlooks issues such as self-awareness and critical thinking.)

Table 20: Two examples that illustrate multiple deficiencies in Qwen2 7B's responses.

H A Practical Interpretation of the P.D. Metric

Given the difficulty in interpreting the P.D. metric’s absolute value, we offer an interpretation of the P.D. value’s change to signify the comparative quality between two model answers, A_1, A_2 . Given the following definitions:

- Perspective Diversity (P.D.)⁸:

$$\text{P.D.} = \sum_{i=1}^n \text{PPL}(\text{PA}^i|A).$$

- Conditional perplexity:

$$\text{PPL}(Y|X) = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(y_i|X, y_{<i}) \right).$$

- We denote the **Change in Probability** for partial answer i as ΔP^i , assuming $|\text{PA}^i| = N$:

$$\Delta P^i = \frac{P(\text{PA}^i|A_2)}{P(\text{PA}^i|A_1)} = \prod_{j=1}^N \frac{P(\text{PA}_j^i|A_2, \text{PA}_{<j}^i)}{P(\text{PA}_j^i|A_1, \text{PA}_{<j}^i)}. \quad (5)$$

The **Change in Perspective Diversity** ($\Delta \text{P.D.}$) is defined as:

$$\Delta \text{P.D.} = \sum_{i=1}^n \text{PPL}(\text{PA}^i|A_2) - \sum_{i=1}^n \text{PPL}(\text{PA}^i|A_1) = \sum_{i=1}^n (\text{PPL}(\text{PA}^i|A_2) - \text{PPL}(\text{PA}^i|A_1)). \quad (6)$$

Substituting the formula of conditional perplexity, we get:

$$\Delta \text{P.D.} = \sum_{i=1}^n \left(\exp \left(-\frac{1}{N} \sum_{j=1}^N \log P(\text{PA}_j^i|A_2, \text{PA}_{<j}^i) \right) - \exp \left(-\frac{1}{N} \sum_{j=1}^N \log P(\text{PA}_j^i|A_1, \text{PA}_{<j}^i) \right) \right).$$

Applying the approximation $\exp(x) \approx 1 + x$ for small x , we get:

$$\begin{aligned} \Delta \text{P.D.} &\approx \sum_{i=1}^n \left(\left(1 - \frac{1}{N} \sum_{j=1}^N \log P(\text{PA}_j^i|A_2, \text{PA}_{<j}^i) \right) - \left(1 - \frac{1}{N} \sum_{j=1}^N \log P(\text{PA}_j^i|A_1, \text{PA}_{<j}^i) \right) \right) \\ &= \sum_{i=1}^n \left(-\frac{1}{N} \sum_{j=1}^N \log P(\text{PA}_j^i|A_2, \text{PA}_{<j}^i) + \frac{1}{N} \sum_{j=1}^N \log P(\text{PA}_j^i|A_1, \text{PA}_{<j}^i) \right) \\ &= -\frac{1}{N} \sum_{i=1}^n \sum_{j=1}^N \log P(\text{PA}_j^i|A_2, \text{PA}_{<j}^i) + \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^N \log P(\text{PA}_j^i|A_1, \text{PA}_{<j}^i). \end{aligned}$$

Applying the approximation that all $\Delta P^i, i = 1 \dots n$ are the same value for simplicity, we get:

$$\Delta \text{P.D.} \approx \sum_{i=1}^n \log \left(\frac{1}{\Delta P^i} \right).$$

Thus we have:

$$\Delta P^1 \approx \exp \left(-\frac{\Delta \text{P.D.}}{n} \right). \quad (7)$$

⁸Here, we use A to denote `chatTemplate(concat(A , "Please restate."))` in Equation 1 for simplicity.

Equation 7 shows that the change in probability ΔP^1 can be approximated using the change in P.D. $\Delta \text{P.D.}$, and the effect is exponential. For example, when taking $n = 3$ (given the fact that the average number of partial answers in DEBATEQA ≈ 3.7) and $\Delta \text{P.D.} = -2$, it suggests that the approximated probability of generating one partial answer under A_2 is 1.95 times higher than under A_1 . In other words, the backbone language model finds it nearly twice as easy to recover the partial answer from A_2 due to the reduction of 2 in the P.D. score.