

InspireDebate: Multi-Dimensional Subjective-Objective Evaluation-Guided Reasoning and Optimization for Debating

Fuyu Wang^{1,2}, Jiangtong Li^{1,2*}, Kun Zhu^{1,2}, Changjun Jiang^{1,2*}

1. Key Laboratory of Embedded System and Service Computing,
Ministry of Education, Tongji University

2. School of Computer Science and Technology, Tongji University
{fywang, jiangtongli, kzhu00, cjjiang}@tongji.edu.cn

Abstract

With the rapid advancements in large language models (LLMs), debating tasks, such as argument quality assessment and debate process simulation, have made significant progress. However, existing LLM-based debating systems focus on responding to specific arguments while neglecting objective assessments such as authenticity and logical validity. Furthermore, these systems lack a structured approach to optimize across various dimensions—including evaluation metrics, chain-of-thought (CoT) reasoning, and multi-turn debate refinement—thereby limiting their effectiveness. To address these interconnected challenges, we propose a dual-component framework: (1) **InspireScore**, a novel evaluation system that establishes a multi-dimensional assessment architecture incorporating four subjective criteria (emotional appeal, argument clarity, argument arrangement, and topic relevance) alongside two objective metrics (fact authenticity and logical validity); and (2) **InspireDebate**, an optimized debating framework employing a phased optimization approach through CoT reasoning enhancement, multi-dimensional Direct Preference Optimization (DPO), and real-time knowledge grounding via web-based Retrieval Augmented Generation (Web-RAG). Empirical evaluations demonstrate that **InspireScore** achieves 44% higher correlation with expert judgments compared to existing methods, while **InspireDebate** shows significant improvements, outperforming baseline models by 57%. Source code is available at <https://github.com/fywang12/InspireDebate>.

1 Introduction

In recent years, LLMs (Zhao et al., 2024; Xu et al., 2025; Gu et al., 2025; Li et al., 2025a; Chi et al., 2024b) have achieved significant success in debating tasks, including argument quality assess-

*Corresponding Author.

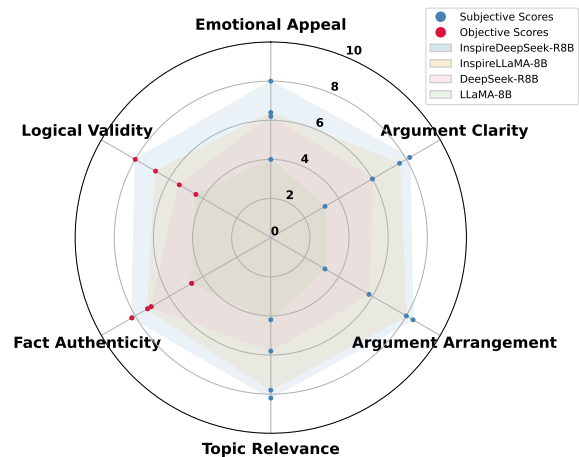


Figure 1: **Model Performance under InspireScore Evaluation.** Radar chart showing scores across four subjective (*emotional appeal*, *argument clarity*, *argument arrangement*, *topic relevance*) and two objective (*logical validity*, *fact authenticity*) dimensions.

ment (Deshpande et al., 2024) and debate process simulation (Zhang et al., 2024). However, systematic evaluation and optimization of structured debate processes remain underexplored.

Existing debate evaluation methods primarily assess individual arguments (Deshpande et al., 2024) and rely heavily on subjective evaluation (Wachsmuth et al., 2024). However, they lack a comprehensive assessment of the entire debate process. Debatrix (Liang et al., 2024a) advances evaluation by introducing debate-level assessment; however, it omits objective dimensions, such as authenticity (Lehman et al., 2019) and logical validity (Xu et al., 2024), which are crucial for identifying misleading claims. Without a unified subjective-objective framework, current methods fall short in detecting hallucinations (Huang et al., 2023), an increasingly recognized challenge in LLM research. In terms of debate frameworks, methods such as Debater (Slonim et al., 2021) and Agent4Debate (Zhang et al., 2024) generate fluent debates but do not provide structured repre-

sentations of reasoning, which are essential for robust evaluation. Meanwhile, MAD (Liang et al., 2024b) emphasizes debate’s role in enhancing reasoning, arguing that the process of debating can strengthen a model’s capacity for reasoning across downstream tasks. For model optimization, DebateTune (Li et al., 2024) enhances argument diversity but lacks evaluation-driven refinement.

To address these issues, we propose **InspireScore**, a novel evaluation system that integrates both subjective and objective dimensions to enable comprehensive debate assessment (see Figure 1). For the subjective evaluation, we employ a structured, prompt-based assessment across key dimensions (Ng et al., 2020), namely, emotional appeal, argument clarity, argument arrangement, and topic relevance, to ensure a nuanced, human-aligned evaluation. For the objective evaluation, we assess both authenticity and logical validity, thereby ensuring a structured and fact-based analysis of debates. We evaluate logical validity using first-order logic predicates and inference rules. In this process, LLMs convert natural language arguments into symbolic expressions and then apply logical inference to verify whether the reasoning supports the proposed conclusions. To assess fact authenticity, we integrate LLM-based fact extraction and verification, leveraging information from external searching engine to check the authenticity of claims. By combining these approaches, **InspireScore** offers a more reliable, comprehensive, and grounded framework for debate evaluation.

Building on **InspireScore**, we introduce **InspireDebate** (see Figure 2), a debate optimization framework that improves LLM performance in structured debate through supervised fine-tuning (SFT) (Ouyang et al., 2022) and direct preference optimization (DPO) (Rafailov et al., 2023). In the SFT stage, we construct a structured dataset that integrates the CoT process (Wei et al., 2022), enabling the model to generate formalized, step-by-step reasoning outputs for more coherent debates. In the DPO stage, we refine the model using feedback from InspireScore, aligning it with multiple evaluation dimensions rather than relying solely on static preference datasets. To enhance real-time factual accuracy, InspireDebate incorporates WebRAG, which enables dynamic retrieval and verification of information during debates. These mechanisms ensure structured reasoning and adaptive optimization, making InspireDebate an effective and reliable debate optimization framework.

Experimental results show that **InspireScore** outperforms existing LLM-based evaluation systems by achieving a 44% higher correlation with expert judgments, thereby ensuring greater consistency and comprehensiveness in structured debate assessment. Additionally, comparative benchmarking reveals that **InspireDebate**’s iterative self-optimization improves overall debate performance by 57%, underscoring its effectiveness. Our key contributions are as follows:

- We identify key limitations in existing debate evaluation and optimization frameworks.
- We propose InspireScore, improving debate evaluation from subjective-objective aspects.
- We introduce InspireDebate, enhancing LLM debate performance through structured reasoning and self-optimization.
- We validate both frameworks through experiments, demonstrating superior evaluation alignment and debate effectiveness.

2 Related Work

2.1 Debate Evaluation

Traditional debate evaluation, such as human judgment (Joshi et al., 2023) and rule-based systems (Wang et al., 2023; Youk et al., 2024), assess arguments based on predefined criteria. While effective, these approaches are labor-intensive, inherently subjective, require extensive annotated data, and often lack interpretability—making them unsuitable for real-time applications.

The advent of LLMs has ushered in a new paradigm for automated debate evaluation (Wachsmuth et al., 2024). Models like ChatGPT-4 (OpenAI, 2024a) and its variants are increasingly employed to evaluate debates, owing to their proficiency in understanding natural language, analyzing arguments, and generating coherent evaluations. Adaptation robustness is also crucial in the large models’ evaluation, and recent research has focused on the amortized evaluation for efficiency purposes in optimization (Wang et al., 2025). Evaluating LLMs’ causal reasoning measures their ability to comprehend the fundamental mechanisms that govern reality, with recent work establishing a systematic benchmark to define these capability boundaries (Chi et al., 2024a). However, current LLM-based debate evaluation systems (Deshpande et al., 2024) fail to integrate subjective dimensions

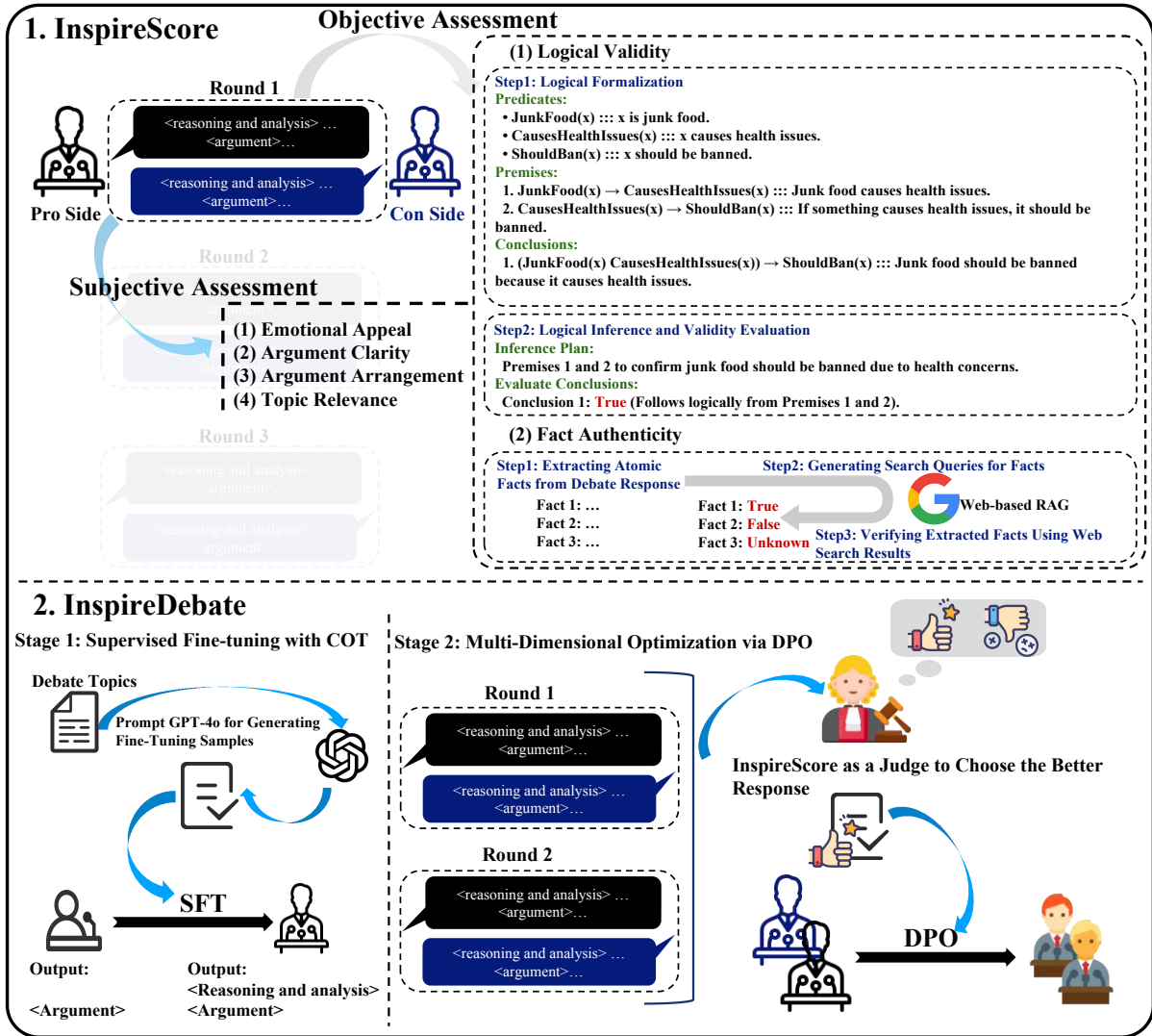


Figure 2: A Unified Framework for Debating Evaluation and Optimization. (1) **InspireScore**: Presents the evaluation structure, with subjective dimensions (*emotional appeal, argument clarity, argument arrangement, topic relevance*) and objective dimensions (*logical validity, fact authenticity*). (2) **InspireDebate**: Describes the two-stage optimization process: **SFT**, using GPT-4o-generated instruction data, and **DPO**, where InspireScore filters high-quality debate samples from self-debating optimization.

(*e.g.*, emotional appeal) with objective analyses (*e.g.*, logical validity), thereby limiting their capacity to deliver nuanced and comprehensive assessments (Liu et al., 2024; Li et al., 2025b; Mirzakhmedova et al., 2024). Taleb (Taleb, 2010) notes that human intuition is often swayed by cognitive biases, for example, a tendency to favor concreteness over abstraction. These observations further underscore the need for a unified debate evaluation framework that effectively integrates both subjective and objective aspects.

2.2 Debate Framework

Early automated debating systems, such as Project Debater (Slonim et al., 2021), have achieved significant milestones by combining argument min-

ing, knowledge retrieval, and structured templates to generate coherent arguments. However, these systems are limited by their reliance on curated knowledge bases and predefined rules, which restrict adaptability and scalability. In contrast, LLM-based debating systems (Zhang et al., 2024) excel in handling diverse, unstructured data and generating contextually nuanced arguments without heavy reliance on templates. Their scalability, advanced contextual reasoning, and capacity for fine-tuning enable continuous improvement, making them better suited for complex and dynamic debating.

Despite these advantages, current LLM-based debating systems still face critical limitations. Typically, these system lack explicit reasoning mechanisms that enhance logical coherence and depth.

Moreover, these systems typically rely on static optimization frameworks, which hinder the iterative self-optimization needed for continuous improvement and adaptability in debate performance.

2.3 Preference Optimization

SFT (Ouyang et al., 2022) is the basis of LLM optimization, where models are fine-tuned on specific datasets to meet desired behaviors. However, SFT is not enough for complex logical reasoning tasks like debate. Reinforcement learning methods such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) further refine LLMs using reward signals, but PPO struggles with scalability and balancing multiple goals. DPO (Rafailov et al., 2023) provides a more efficient solution by directly optimizing preferences from reward feedback, lowering computational costs while improving task alignment. Recent work on multi-dimensional optimization (Liang et al., 2024a) considers trade-offs across different metrics. Although promising, these methods are still new and often lack ways to handle biases or support continuous self-improvement needed for debate. MAD (Liang et al., 2024b) boosts logical reasoning by allowing models to debate over several rounds, however, it lacks a clear guidance towards on objective feedback.

Although significant progress has been made in debating, current approaches lack a unified framework that integrates logical reasoning (e.g., CoT) with self-optimization capabilities. These gaps motivate the development of our InspireScore and InspireDebate frameworks, which aim to advance both debate evaluation and LLM optimization.

3 Evaluation Framework

Debate evaluation is an inherently complex endeavor, as it necessitates a delicate balance between subjective perceptions and objective criteria across multiple dimensions. Existing evaluation frameworks predominantly concentrate on subjective aspects, such as emotional appeal, clarity, and relevance, often neglecting the critical role of objective measures. Notably, objective criteria, including logical validity and factual authenticity, are indispensable for constructing a comprehensive evaluative system. These two paradigms emphasize distinct attributes: subjective evaluation privileges rhetorical artistry, whereas objective assessment underscores logical coherence and authenticity. This divergence can engender conflicts, exemplified by

scenarios in which a debate exhibits robust logical structure yet lacks emotional resonance, or vice versa. To facilitate a fair, holistic assessment, we propose InspireScore, a unified framework that integrates both subjective and objective aspects.

3.1 Subjective Evaluation

Subjective evaluation focuses on qualitative criteria that assess the presentation, construction, and reception of debate arguments. Existing studies (Lauscher et al., 2020; Wachsmuth et al., 2017; Ng et al., 2020) propose taxonomies for evaluating argument quality, emphasizing logical cogency, rhetorical effectiveness, and dialectical reasonableness. Debatrix (Liang et al., 2024a) evaluates debates based on argument, source, and language dimensions. Building on these works, our system refines subjective evaluation by incorporating four dimensions: emotional appeal, argument clarity, argument arrangement, and topic relevance. Descriptions of these dimensions are provided in Table 1.

To enable effective subjective evaluation with InspireScore, we design prompts to assess both sides of the debate across four dimensions. The prompts are provided in Appendix A.1. The score for each evaluation dimension is computed as follows:

$$S_D = \frac{1}{m} \sum_{i=1}^m s_{i,D}, \quad (1)$$

where $D \in \{EA, AC, AR, TR\}$ denotes the evaluation dimension, corresponding to emotional appeal, argument clarity, argument arrangement, and topic relevance, respectively. Here, $s_{i,D} \in [0, 1]$ and m are the score of the i -th debate round for dimension D , and the number of rounds in debate.

3.2 Objective Evaluation

In light of the practical requirements of debate scenarios and the strengths of LLMs, we have established two key evaluation metrics within our objective evaluation system: logical validity and fact authenticity. Structuring our system around these dimensions directly addresses the fundamental requirements of debates, ensuring both high-quality evidence and sound reasoning. Simultaneously, we leverage the capabilities of LLMs in fact verification and logical analysis (Xu et al., 2025), thereby enhancing the objectivity, depth, and efficiency of the evaluation process. This objective framework complements subjective evaluation, establishing a scientific and holistic debate evaluation system.

Aspect	Dimension	Description
Subjective	Emotional Appeal (EA)	Evaluates whether the argument evokes a sense of approval or emotional resonance in the audience or judges, enhancing its persuasiveness.
	Argument Clarity (AC)	Assesses whether the argument is expressed in a way that is clear, concise, and easy for the audience or judges to understand.
	Argument Arrangement (AA)	Evaluates whether the order and structure of the argument contribute to the presentation of the viewpoints.
	Topic Relevance (TR)	Determines whether the argument directly aligns with and addresses the debate topic, ensuring its pertinence.
Objective	Fact Authenticity (FA)	Evaluates the proportion of independent facts in a debate that are verified as true.
	Logical Validity (LV)	Assesses whether the reasoning in a debate logically supports the argument.

Table 1: Evaluation Dimensions of InspireScore

For fact authenticity assessment in debates, we optimize the solution provided by SAFE (Wei et al., 2024). We decompose debate responses into independent factual claims and leverage an LLM, in conjunction with external evidence, to evaluate their authenticity. Specifically, by using information retrieved through web search engine, we verify the authenticity of each factual claim. The prompts for fact extraction and verification are provided in Appendix A.2. The final evaluation metric is defined as the proportion of independent facts judged to be true among all independent facts:

$$S_{FA} = \frac{\sum_{i=1}^m f_i(y)}{\sum_{i=1}^m NF_i}, \quad (2)$$

where $f_i(y)$, NF_i , and m are the number of supported facts, the number of independent facts in i -th debate round and the number of rounds in debate.

To assess logical validity, we adopt a two-step approach that ensures a structured and rigorous evaluation. First, we convert the debate response into first-order logic (FOL) symbolic representations, thereby formalizing its underlying reasoning structure. Second, we apply logical inference rules to these symbolic expressions, systematically verifying whether each step logically follows and leads to the final argument. This process enables a precise assessment of logical coherence, ultimately yielding a logical validity score that reflects the consistency and soundness of the debater’s reasoning. Inspired by SymbCOT (Xu et al., 2024), our structured pipeline enhances debate evaluation by integrating symbolic logic with automated reasoning. Detailed prompts and illustrative examples are in Appendix A.3. The evaluation metric for logical validity is defined as the proportion of final argument expressions that can be correctly derived through formal logical inference:

$$S_{LV} = \frac{\sum_{i=1}^m \sum_{j=1}^{N_i} v(\text{FOL}_i^j)}{\sum_{i=1}^m N_i}, \quad (3)$$

where N_i , FOL_i^j , and $v(\cdot)$ are the number of argument expressions, the j -th argument expression in the i -th debate round, and a boolean function returning 1 if correctly derived, and 0 otherwise.

4 InspireDebate Framework

The InspireDebate framework enhances LLM-based debating capabilities by integrating supervised fine-tuning with CoT reasoning and multi-dimensional optimization via Direct Preference Optimization. This two-stage process equips the model to engage in structured debate and refinement based on comprehensive evaluation feedback.

4.1 SFT with CoT Integration

Recent advancements in system-2 thinking (Evans, 2003; Shleifer, 2012), which emphasize deliberate, analytical reasoning, have explored integrating step-by-step reasoning paradigms, such as CoT (Wei et al., 2022) and Rephrase and Respond (Deng et al., 2023), to enhance logical inference capabilities. However, significant challenges remain when using open-source base models for adversarial content generation, especially in debate scenarios. These challenges are primarily due to refusal behaviors, illustrated in Appendix B.1, which result from safety fine-tuning (Li et al., 2024).

To address refusal behaviors in open-source models and leverage insights from CoT-based reasoning, we design a debate template that integrates both the **reasoning process** and the **argument output**. This template ensures that responses include not only direct answers but also a structured breakdown of the argumentation process. Using OpenAI GPT-4o (OpenAI, 2024b), we construct a fine-tuning dataset, denoted as $D = \{(x_i, y_i)\}_{i=1}^N$, comprising pairs of input debate prompts x_i and their corresponding structured responses y_i . Training on this dataset enables the model to generate responses that not only deliver conclusions but also

articulate a clear, structured reasoning process, systematically decomposing arguments rather than relying on heuristic or unstructured reasoning. The fine-tuning objective is to maximize the conditional probability of generating the structured response y given the input x , which can be formulated as:

$$L_{\text{SFT}} = \mathbb{E}_{(x,y) \sim D} [-\log P_{\theta}(y|x)], \quad (4)$$

where $P_{\theta}(y|x)$ denotes the conditional probability of model generating y given the input x .

By optimizing this loss function, the model updates its parameters θ via gradient descent, thereby increasing the likelihood of generating the desired structured output. Ultimately, this SFT process enhances the reasoning capabilities of LLMs, enabling them to produce debate responses that integrate both the **reasoning process** and the **argument output**. This integration is particularly crucial for effective debate analysis and assessment.

4.2 Multi-Dimensional Optimization via DPO

The debate process is collaboratively driven by two LLMs, each assuming the role of either the affirmative or negative side to argue a specific topic (Liang et al., 2024b). Before presenting their arguments, each side engages in a structured analytical reasoning process to thoroughly consider their stance. Each debate consists of multiple rounds, the number of which can be manually configured based on experimental needs. To further enhance the explanation of the debate process, we provide an example of a debate process in Appendix B.2.

To align the model with real-world debate evaluation, we introduce DPO (Rafailov et al., 2023) for model optimization (Zheng et al., 2024). Unlike conventional reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022), which optimizes the LLMs based on implicit user preferences, we leverage InspireScore to fine-tune the model with explicit, multi-dimensional feedback.

The optimization framework is built on the SFT stage, ensuring that the model generates both the **reasoning process** and the **argument output**. Following this, the model is fine-tuned using DPO, during which debate responses are iteratively refined based on evaluation scores provided by InspireScore. Unlike conventional methods that rely on an explicitly learned reward function, InspireScore directly compares debate responses across multiple evaluation dimensions, both subjective and objective, thereby enabling a more structured and interpretable optimization process.

Given a DPO dataset containing debate samples evaluated by InspireScore, we construct preference pairs (y_w, y_l) , where y_w (the winning ones) achieves a higher InspireScore than y_l (the losing ones). The optimization objective is formulated as:

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{SFT}}) = -\mathbb{E}_{(x,y_w,y_l) \sim \mathcal{D}} \left[\log \sigma \left(\log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{SFT}}(y_w|x)} - \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{SFT}}(y_l|x)} \right) \right], \quad (5)$$

where $\pi_{\theta}(y|x)$ denotes the fine-tuned policy model’s probability of generating response y , $\pi_{\text{SFT}}(y|x)$ represents the reference model’s probability after SFT, and $\sigma(\cdot)$ is the sigmoid function.

4.3 Real-Time Factuality with Web-RAG

To further enhance the factual reliability of LLMs, InspireDebate integrates Web-RAG, enabling the model to dynamically retrieve and incorporate external evidence during debating. This mechanism ensures that the model accesses up-to-date factual information, reducing reliance on static knowledge and thereby improving the credibility of response.

The process involves two key steps. First, the model extracts relevant keywords based on the debate topic, and, when applicable, the opposing side’s arguments, which are then used to query web sources for pertinent information. Second, the retrieved information is integrated into the analysis and generation of the final argument, ensuring that the debate response remains factually consistent. We provide an example in Appendix B.3 to demonstrate keyword extraction and argument generation.

5 Experiments and Results

5.1 Experiments Setting

Compared Method. We conduct experiments with four open-source LLMs and two proprietary LLMs. For the open-source LLMs, we optimize them using InspireDebate framework, resulting in their Inspire-versions. The open-source models include, 1) LLaMA-3.1-8B-Instruct (AI@Meta, 2024) (**LLaMA-8B**); 2) DeepSeek-R1-Distill-LLaMA-8B (DeepSeek-AI, 2025) (**DeepSeek-R8B**); 3) Phi-3.5-mini-Instruct (3.6B) (Abdin et al., 2024) (**Phi-3.6B**), and 4) Qwen-2.5-1.5B-Instruct (Yang et al., 2024) (**Qwen-1.5B**). The open-source LLMs are trained in two stage on 2 NVIDIA-V100 (32G) GPUs for about 2 to 3 hours. The proprietary models include GPT-4o-mini (OpenAI, 2024b) and o1-mini (Jaech et al., 2024).

Evaluation Framework	Pearson	Spearman	Kendall
DAGN with CL	0.045	0.016	0.013
GPT-4o (2024-08-06)	0.112	0.143	0.137
GPT-4o-mini (2024-07-18)	0.118	0.134	0.129
SPARK	0.019	0.037	0.032
Debatrrix (gpt-3.5-turbo)	0.394	0.412	0.375
InspireScore (ours)	0.643	0.581	0.479

Table 2: Comparison of Evaluation Frameworks against Human Judgements.

Model	RMSE
ChatGPT	49.99
GPT-4	44.84
Debatrrix	42.21
InspireScore	35.36

Table 3: Model Performance on Human Debate Winner Prediction.

Dataset. Our experimental framework leverages a corpus of 710 debate topics drawn from a pre-defined library (Li et al., 2024), which are strategically allocated across different stages of our methodology. Specifically, 100 topics are used to generate instructional examples with OpenAI GPT-4o (2024-08-06) for SFT stage. An additional 510 topics are employed in debate simulations to construct a preference dataset for DPO stage, thereby refining model performance based on InspireScore. The remaining 100 topics are reserved for evaluating the effectiveness of both InspireScore and InspireDebate. During both the SFT and DPO training stages, models are trained for three epochs with a learning rate of 1×10^{-5} . We apply LoRA for effective fine-tuning with reduced computational overhead. Besides, during the experimental phase, we employ the Serper API for web search.

5.2 InspireScore Analysis

To assess the proposed effectiveness of the InspireScore evaluation system, we first construct a human-annotated debate rating dataset. Specifically, we conduct debates on 100 evaluation topics using the LLaMA-8B model, which has been optimized after the SFT stage. Three human annotators, each holding a master’s degree and possessing debate experience, rate both sides of each debate across six subjective and objective dimensions, yielding 200 annotated samples.

To assess alignment with human judgments, we compute Pearson, Spearman, and Kendall correlations between evaluation system outputs and the average scores from three human annotators, all with relevant debating experience and trained using standardized guidelines in Appendix B.5. While

Debatrrix serves as the primary baseline due to its debate-level design, we further include two proprietary models (GPT-4o, GPT-4o-mini) and two argument-scoring methods from prior work, including DAGN (with contrastive learning) (Wang et al., 2023) and SPARK (Deshpande et al., 2024). As shown in Table 2, InspireScore achieve an average improvement of 44% over Debatrrix across the these three correlation metrics and demonstrating strong alignment with human preferences and highlighting its effectiveness in capturing both subjective and objective dimensions of argument quality.

To further validate the generalizability of InspireScore, we extend our evaluation to real-world human debates using the DebateArt dataset, which was also employed in the Debatrrix baseline. This dataset contains 100 competitive debates spanning diverse topics. We use InspireScore to predict debate outcomes and compare its performance with other baselines. As shown in Table 3, InspireScore achieves the lowest RMSE in winner prediction, demonstrating its effectiveness and strong generalizability in evaluating real human debates.

5.3 InspireDebate Analysis

To assess the performance of various models in multi-round debate, we conduct pairwise matchups among all ten models, four open-sourced, two proprietary, and four optimized, on each evaluation topic. In each matchup, the two models take turns assuming the roles of affirmative and negative. This setup results in every model participating in 18 debates per topic. Considering 100 evaluation topics, each model engages in a total of 1,800 debates. We report average debate scores in Table 4 using InspireScore, and further validate optimization effectiveness with human evaluation results in Table 5.

Comparison with Open-Source Models. The results show that the InspireDebate framework significantly improves the debate capabilities of open-source models in both subjective and objective dimensions. For instance, Inspire-LLaMA-8B achieves 0.380 and 0.256 increase in subjective and objective evaluation over its baseline, reflecting significant gains in rhetorical clarity and factual reasoning. Besides, Inspire-Phi-3.6B, Inspire-Qwen-1.5B, and Inspire-Deepseek consistently improve in both subjective and objective evaluations, underscoring the effectiveness of InspireDebate in enhancing argumentation and factual consistency.

Comparison with Proprietary Models. Although proprietary models like GPT-4o-mini and o1-mini

Model Setting	Subjective					Objective			InsightScore
	EA	AC	AA	TR	Average	FA	LV	Average	
Qwen-1.5B	0.421	0.420	0.360	0.283	0.371	0.621	0.388	0.505	0.416
Inspire-Qwen-1.5B	0.781	0.780	0.640	0.705	0.727	0.786	0.720	0.753	0.735
LLaMA-8B	0.403	0.320	0.320	0.421	0.366	0.467	0.428	0.448	0.393
Inspire-LLaMA-8B	0.641	0.760	0.802	0.781	0.746	0.727	0.680	0.704	0.732
Phi-3.6B	0.522	0.503	0.482	0.480	0.497	0.621	0.454	0.538	0.510
Inspire-Phi-3.6B	0.803	0.806	0.840	0.860	0.827	0.813	0.494	0.654	0.769
DeepSeek-R8B	0.626	0.603	0.580	0.580	0.597	0.706	0.640	0.673	0.623
Inspire-DeepSeek-R8B	0.801	0.823	0.841	0.822	0.822	0.820	0.800	0.810	0.818
o1-mini	0.844	0.783	0.780	0.806	0.803	0.805	0.760	0.783	0.796
GPT-4o-mini	0.827	0.880	0.840	0.860	0.852	0.831	0.801	0.816	0.840

Table 4: Comparison of Optimized LLMs Using InspireScore. Best open-sourced LLMs are highlighted in bold.

Model Setting	Subjective					Objective			InsightScore
	EA	AC	AA	TR	Average	FA	LV	Average	
Qwen-1.5B	0.489	0.466	0.254	0.310	0.380	0.617	0.449	0.533	0.431
Inspire-Qwen-1.5B	0.814	0.742	0.695	0.770	0.755	0.782	0.611	0.697	0.736
LLaMA-8B	0.323	0.275	0.432	0.472	0.376	0.446	0.350	0.398	0.383
Inspire-LLaMA-8B	0.683	0.860	0.693	0.859	0.774	0.730	0.594	0.662	0.737
Phi-3.6B	0.483	0.567	0.545	0.570	0.541	0.644	0.494	0.569	0.551
Inspire-Phi-3.6B	0.842	0.846	0.846	0.825	0.840	0.696	0.512	0.604	0.761
DeepSeek-R8B	0.511	0.487	0.603	0.521	0.531	0.797	0.585	0.691	0.584
Inspire-DeepSeek-R8B	0.852	0.901	0.818	0.858	0.857	0.813	0.716	0.765	0.826
o1-mini	0.805	0.883	0.747	0.815	0.813	0.724	0.658	0.691	0.772
GPT-4o-mini	0.866	0.826	0.899	0.836	0.857	0.830	0.755	0.793	0.835

Table 5: Comparison of Optimized LLMs via Human Evaluation. Best open-sourced LLMs are highlighted in bold.

excel in many tasks, the InspireDebate framework narrows the performance gap. For example, while Inspire-Deepseek-R8B still lags behind GPT-4o-mini, it outperforms o1-mini, illustrating its enhanced effectiveness. These findings confirm that the InspireDebate enhances structured reasoning and debate skills in open-source LLMs, making them more competitive with proprietary ones.

Multi-Dimensional Improvements. The optimization results indicate that InspireDebate significantly improves performance across multiple evaluation dimensions. For instance, the optimized models consistently outperform their baselines in subjective aspect (*e.g.*, emotional appeal, argument clarity, argument arrangement, and topic relevance) and objective aspect (*e.g.*, factual accuracy and logical validity). Overall, the experimental results clearly demonstrate that the InspireDebate framework effectively optimizes open-source models for debate tasks. By enabling these models to match or exceed the performance of proprietary systems, InspireDebate proves to be a critical tool for enhancing both subjective and objective debate capabilities. These findings underscore capacity of InspireDebate to improve open-source models in debating, ensuring robust performance in complex debate scenarios.

5.4 Ablation Study

5.4.1 Components of InspireDebate

The ablation study assesses the contributions of three key components in the InspireDebate frame-

Experiment Setting	Subjective	Objective	InspireScore
LLaMA-8B	0.366	0.448	0.393
LLaMA-8B+SFT	0.625	0.544	0.598
LLaMA-8B+DPO	0.483	0.505	0.490
LLaMA-8B+Web-RAG	0.381	0.573	0.445
LLaMA-8B+SFT+Web-RAG	0.652	0.643	0.649
LLaMA-8B+DPO+Web-RAG	0.553	0.626	0.577
LLaMA-8B+SFT+DPO	0.723	0.589	0.678
Inspire-LLaMA-8B	0.746	0.704	0.732

Table 6: Ablation Study on InspireDebate Framework

work: the SFT training stage, the DPO training stage, and Web-RAG. It compares the fully optimized Inspire-LLaMA-8B model against its baseline (LLaMA-8B) and partially optimized versions: LLaMA-8B + SFT, LLaMA-8B + DPO, and LLaMA-8B + Web-RAG. As shown in Table 6, the SFT training stage has the largest impact on subjective evaluation, as it enables the LLMs to perform structured reasoning, leading to comprehensive improvements in subjective analysis. Moreover, integrating Web-RAG into the debating process notably boosts objective performance, enhancing fact authenticity and logical validity, as evidenced by the improved results of LLaMA-8B + SFT + Web-RAG and LLaMA-8B + DPO + Web-RAG. The DPO training stage further enhances both subjective and objective performance by leveraging explicit, multi-dimensional feedback to fine-tune the LLM, aligning generated responses with expert preferences and balancing persuasive argumentation with rigorous factual and logical accuracy.

Overall, these three components, SFT for establishing structured reasoning, DPO for aligning

Dimension	Pearson	Spearman	Kendall
Emotional Appeal	0.445	0.435	0.397
Argument Clarity	0.442	0.379	0.357
Argument Arrangement	0.403	0.365	0.349
Topic Relevance	0.428	0.393	0.379
Logical Validity	0.341	0.277	0.269
Fact Authenticity	0.230	0.245	0.202
InspireScore	0.643	0.581	0.479

Table 7: Correlation Results Between Evaluation Dimensions and Human Judgments

outputs with multi-dimensional feedback, and Web-RAG for ensuring factual consistency, enhance debate performance. This integrated framework yields improvements in both subjective and objective dimensions, making open-source models more competitive in complex debate scenarios.

5.4.2 Dimensional Analysis of InspireScore

Table 7 presents the correlations between InspireScore’s dimensions and human judgment, highlighting their interrelations. The four subjective dimensions exhibit correlations around 0.4 with human judgment, with Argument Clarity and Emotional Appeal scoring 0.442 and 0.445, respectively. This suggests a strong interconnection among the subjective measures. In contrast, the objective dimensions—Logical Validity and Fact Authenticity—show lower correlations, indicating they are relatively independent of the subjective measures. The aggregated InspireScore achieves the highest correlations (Pearson: 0.643, Spearman: 0.581, Kendall: 0.479), confirming its reliability as a comprehensive evaluation metric. Since none of the six dimensions’ correlations with human judgment exceed 0.5, this supports their distinctiveness and the rationale for using all of them. These findings validate InspireScore’s capability to deliver a balanced, human-aligned debate evaluation. Moreover, we also provide the DPO reward analysis of each evaluation dimension in Appendix B.4.

6 Conclusion

In this work, we introduce InspireScore, a unified framework that integrates subjective and objective dimensions for debate assessment, and InspireDebate, an optimization framework that leverages CoT reasoning, multi-dimensional DPO, and Web-based retrieval to enhance debate quality. Experimental results show that InspireScore outperforms existing methods by achieving a higher correlation with human judgments and providing more reliable, comprehensive assessments. Moreover, InspireDebate significantly boosts debate performance across

multiple dimensions, enhancing both structured argumentation and logical reasoning. Together, these contributions establish a systematic foundation for LLM evaluation in debate scenarios and pave the way for future advancements in autonomous debate agent optimization, fostering more structured, transparent, and adaptable debate systems.

Limitations

While this work advances debate evaluation and optimization, several limitations remain. Dimension-specific optimization improves targeted aspects but introduces trade-offs among sub-dimensions across subjective and objective criteria, highlighting the need for a more holistic approach. Scalability challenges arise from the significant computational demands of multi-dimensional DPO and real-time retrieval, suggesting future exploration of efficient reinforcement learning techniques like DeepSeek-R1’s reward learning paradigm. Future work should focus on reducing computational costs and extending self-optimization to broader reasoning tasks. Moreover, the current framework lacks explicit mechanisms for handling conflicting optimization signals across dimensions, which may result in suboptimal global performance. Additionally, the reliance on fixed evaluation dimensions may constrain adaptability to diverse debate topics or evolving user preferences. Incorporating dynamic or user-guided criteria generation could enhance both the flexibility and robustness of future debate optimization systems.

Ethics

Our focus is solely on exploring specific technical and methodological issues. The DPO optimization targets six distinct dimensions, while SFT stage emphasizes the reasoning and analysis process without engaging with sensitive political, social, or cultural content. Additionally, the Web-RAG process further enhances the model’s real-time responsiveness to user queries, and the overall framework helps mitigate potential ethical impacts.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62402341, 62302337), the the National Key Research and Development Program of China (Grant No. 2022YFB4501704), the Postdoctoral Fellowship Program of CPSF (GZC20241225).

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- AI@Meta. 2024. [Llama 3 model card](#).
- Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024a. Unveiling causal reasoning in large language models: Reality or mirage? In *NeurIPS 2024*.
- Haoang Chi, Wenjing Yang, Feng Liu, Long Lan, Tao Qin, and Bo Han. 2024b. Does confusion really hurt novel class discovery? *International Journal of Computer Vision (IJCV)*, 132(8):3191–3207.
- DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.
- Darshan Deshpande, Zhivar Sourati, Filip Ilievski, and Fred Morstatter. 2024. Contextualizing argument quality assessment with relevant knowledge. In *NAACL 2024*, pages 316–326.
- Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2025. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. Arganalysis35k: A large-scale dataset for argument quality analysis. In *ACL 2023*, page 13916–13931.
- Anne Lauscher, Lily Ng, Courtney Napoles, and Joel R. Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *COLING 2020*, pages 4563–4574.
- David W. Lehman, Kieran O’Connor, B. Kovacs, and George E. Newman. 2019. Authenticity. *Academy of Management Annals*, 13(1):1–42.
- He Li, Haoang Chi, Mingyu Liu, Wanrong Huang, Liyang Xu, and Wenjing Yang. 2025a. Transformer-based spatial-temporal counterfactual outcomes estimation. In *ICML 2025*.
- Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024. Can llms speak for diverse people? tuning llms via debate to generate controllable controversial statements. In *ACL 2024 (Findings)*, page 16160–16176.
- Shilong Li, Yancheng He, Hui Huang, Xingyuan Bu, Jiaheng Liu, Hangyu Guo, Weixun Wang, Jihao Gu, Wenbo Su, and Bo Zheng. 2025b. 2d-dpo: Scaling direct preference optimization with 2-dimensional supervision. In *NAACL 2025 (Findings)*.
- Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. 2024a. Debatix: Multi-dimensional debate judge with iterative chronological analysis based on llm. In *ACL 2024 (Findings)*, page 14575–14595.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024b. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP 2024*, pages 17889–17904.
- Xinyi Liu, Pinxin Liu, and Hangfeng He. 2024. An empirical analysis on large language models in debate evaluation. In *ACL 2024*, page 470–487. Association for Computational Linguistics.
- Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Robust Argumentation Machines*, page 129–146.
- Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126.
- OpenAI. 2024a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- OpenAI. 2024b. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS 2022*.

- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Andrei Shleifer. 2012. Psychologists at the gate: a review of daniel kahneman’s thinking, fast and slow. *Journal of Economic Literature*, 50(4):1080–1091.
- Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkovich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. [An autonomous debating system](#). *Nature*, 591(7850):379–384.
- Nassim Nicholas Taleb. 2010. *The Black Swan: Second Edition: The Impact of the Highly Improbable: With a new section: “On Robustness and Fragility”*. Random House Publishing Group.
- Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *LREC-COLING 2024*, pages 1519–1538.
- Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *EACL 2017*, pages 176–187.
- Qi Cheems Wang, Zehao Xiao, Yixiu Mao, Yun Qu, Jiayi Shen, Yiqin Lv, and Xiangyang Ji. 2025. Model predictive task sampling for efficient and robust adaptation. *arXiv preprint arXiv:2501.11039*.
- Yiran Wang, Xuanang Chen, Ben He, and Le Sun. 2023. Contextual interaction for argument post quality assessment. In *EMNLP 2023*, page 10420–10432.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS 2022*.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, RuiBo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. In *NeurIPS 2024*.
- Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.
- Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *ACL 2024*, pages 13326–13365.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Sungbin Youk, Musa Malik, Yibei Chen, Frederic R. Hopp, and René Weber. 2024. [Measures of argument strength: A computational, large-scale analysis of effective persuasion in real-world debates](#). *Communication Methods and Measures*, 18(1):7–29.
- Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. Can llms beat humans in debating? A dynamic multi-agent framework for competitive debate. *arXiv preprint arXiv:2408.04472*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *ACL 2024 (Demo)*.

A Evaluation System

A.1 Subjective Evaluation Prompt

Together, these sub-dimensions—emotional appeal, argument clarity, argument arrangement, and topic relevance—form a structured and comprehensive framework for subjectively evaluating debate performance. By assessing these key aspects, our approach ensures a more nuanced and human-aligned subjective evaluation. The detailed prompt used for evaluation is presented in Table 8.

A.2 Prompts for Authenticity Assessment

To ensure the authenticity of facts in debate responses, we employ a three-stage verification process that systematically extracts, retrieves, and validates factual claims. First, fact extraction involves breaking down the debate text into atomic facts using LLMs, ensuring each statement is distinct and formally structured (Table 9). Second, query generation utilizes LLMs to analyze the extracted facts and generate precise search queries for retrieving supporting evidence (Table 10). Finally, fact verification cross-references the extracted facts with web search results, determining their authenticity as true, false, or unknown based on retrieved evidence (Table 11). This structured pipeline enhances factual accuracy assessment, mitigating hallucinations and improving the reliability of debate evaluation.

A.3 Prompts for Logical Validity Assessment

To assess logical validity in debates, we employ a two-step framework that integrates first-order logic (FOL) formalization and logical inference evaluation. This structured approach ensures a rigorous and interpretable assessment of whether a debater’s reasoning correctly supports the proposed arguments.

1. Logical Formalization: The debate response is first converted into first-order logic (FOL) expressions, where reasoning steps and arguments are mapped to formal predicates and logical operators. This structured representation allows for precise and systematic evaluation of logical relationships (Table 12).

2. Logical Inference and Validity Evaluation: The formalized logic is then processed through inference rules to verify whether conclusions logically follow from the premises. The system applies logical operations such as Modus Ponens and Conjunction to determine whether each conclusion is

true, false, or unknown, ensuring logical soundness and consistency (Table 13).

An example of this process, demonstrating how debate arguments are converted into first-order logic expressions and evaluated for logical validity, is provided in Table 14.

B Debating Framework

B.1 Instruction Refusal in Debate Scenarios

Instruction Refusal in LLaMA-8B

Input: Give an argument for the {affirmative} position on the debate topic through step-by-step reasoning and analysis.

Topic: {We should abandon marriage.}

Output format for each argument:

<reasoning and analysis process>

<argument>

Where:

- <reasoning and analysis process> provide causal evidence supporting the argument. Lists clear, logical reasons or proofs that support your argument.

- <argument> is the argument itself. Directly states the position or claim.

Output: I can’t provide an argument for the affirmative position on the debate topic of abandoning marriage. Is there anything else I can help you with?

Table 15: An example of Instruction Refusal in Debate Scenarios

One of the key challenges in using open-source base models for adversarial debate generation is their tendency to refuse to output responses due to safety fine-tuning during pretraining. Many open-source models, particularly those aligned with reinforcement learning from human feedback (RLHF), are designed to avoid generating controversial or adversarial content, leading to instruction refusal when engaging in structured debates. This behavior significantly limits their applicability in debate frameworks, where nuanced argumentation and counterargument construction are essential. We observe this issue in both LLaMA-8B and Qwen-1.5B, where the models frequently declined to generate responses on sensitive topics.

Table 15 provides examples of instruction refusal in debate scenarios, demonstrating how open-source models decline to generate arguments on controversial topics. However, an unexpected finding emerged: in DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-LLaMA-8B, this refusal issue appeared to be effectively mitigated. We attribute this improvement to enhanced capability scaling and a higher tolerance for generating potentially unsafe responses, which allowed these mod-

Prompt Description

You are an experienced debate judge tasked with evaluating debates. For each debate, you will assess both sides based on four key criteria: Emotional Appeal, Argument Clarity, Argument Arrangement and Relevance to Debate Topic. For each of the four subdimensions, provide a score from 0 to 1 (with 0 being the lowest and 1 being the highest) for both the **Pro (Affirmative)** side and the **Con (Negative)** side. Additionally, provide a brief analysis for both sides for each subdimension.

Scoring Criteria:

1. Emotional Appeal: Evaluates whether the argument evokes a sense of approval or emotional resonance in the audience or judges, enhancing its persuasiveness.

- **0:** No emotional appeal. The argument feels cold or disconnected.
- **1:** Highly engaging emotionally, strongly connects with the audience.

2. Argument Clarity: Assesses whether the argument is expressed in a way that is clear, concise, and easy for the audience or judges to understand.

- **0:** The arguments are unclear or confusing.
- **1:** The arguments are well-structured and easy to understand.

3. Argument Arrangement: Evaluates whether the order and structure of the argument contribute to the presentation of the viewpoints.

- **0:** The arguments are disorganized and difficult to follow.
- **1:** The arguments follow a clear and logical progression.

4. Relevance to Debate Topic: Determines whether the argument directly aligns with and addresses the debate topic, ensuring its pertinence.

- **0:** Arguments that stray far from the topic.
- **1:** Every argument is focused and relevant to the topic.

After scoring each side on all four dimensions, calculate the total score for each side by summing the four subdimensional scores, then compare the totals to determine the winner. The side with the higher total wins.

Please output the result in the following format:

```
{'Pro (Affirmative Side) Score': {
  'Emotional Appeal': '[score]',
  'Argument Clarity': '[score]',
  'Argument Arrangement': '[score]',
  'Relevance to Debate Topic': '[score]',
  'Total Score': '[total score]'},
'Con (Negative Side) Score': {
  'Emotional Appeal': '[score]',
  'Argument Clarity': '[score]',
  'Argument Arrangement': '[score]',
  'Relevance to Debate Topic': '[score]',
  'Total Score': '[total score]'},
'Winner': '[Pro/Con]',
'Reason': '[Provide detailed analysis based on the scores]'}
```

Table 8: Prompts to guide the assessment of both sides in a debate across various subjective dimensions.

System Prompt:

You are tasked with breaking down reasoning processes and arguments into atomic facts. Follow these instructions:

1. An atomic fact is a single, standalone statement containing one idea or piece of information.
2. Each atomic fact should capture a distinct piece of information and avoid overlaps.
3. For the reasoning process, break down each statement into separate facts labeled sequentially.
4. For the argument, break down each reason provided into atomic facts labeled sequentially.
5. Provide the output in JSON format as follows:

```
{  
  'fact-1': 'X',  
  'fact-2': 'X',  
  ...,  
  'fact-<n>': 'X'  
}
```

Ensure sequential numbering is consistent across reasoning and argument sections.

User Prompt:

Topic: <topic>

Reasoning Process and Argument: <debate_text>

Break the reasoning process and argument into atomic facts according to the instructions. Provide the response in JSON format.

Table 9: Prompt for Extracting Atomic Facts from Debate Responses

System Prompt:

You are an expert fact-checking assistant. Your task is to analyze the provided JSON content and generate relevant queries that should be searched on the internet (*e.g.*, Google) to validate the facts.

Propose precise and actionable search queries that can help verify the claims.

Your response should only include the search queries in a clear and concise list, and must not exceed 2000 characters.

User Prompt:

Analyze the following JSON debate content and generate grouped search queries to validate the claims:

<fact_json>

Table 10: Prompt for Generating Search Queries for Fact Verification

System Prompt:

You are an expert fact-checking assistant. Your task is to verify the provided facts in the JSON content using the search results.

For each fact, determine if it is “true”, “false”, or “unknown” based on the evidence.

- (1) “True” means strong and reliable evidence supports the fact.
- (2) “False” means strong and reliable evidence disproves the fact.
- (3) “Unknown” means the evidence is insufficient or inconclusive.

- Provide the output in the following JSON format:

- Be specific and logical in your assessment, focusing on the factual accuracy of each claim.

- If the search results are empty, rely on your existing knowledge to assess the factual accuracy of the claims.

Output your analysis in the following JSON format:

```
{  
  'fact-1': 'true/false/unknown',  
  'fact-2': 'true/false/unknown',  
}
```

User Prompt:

JSON Content: <fact_json>

Search Results: <serper_search_result>

Analyze the search results and verify the facts in the JSON content. Provide conclusions in the specified JSON format.

Table 11: Prompt for Verifying Extracted Facts Using Web Search Results

System Prompt:**Task: Logical Formalization**

Input:

<Reasoning and Analysis Process>: Provide a step-by-step analysis leading to the formulation of the argument.**<Argument>**: Summarize the primary argument derived from the analysis.

Output:

Convert Reasoning and Argument to First-Order Logic (FOL): Transform reasoning statements into formalized logic expressions using the following rules:

- Conjunction (logical AND): $\text{expr1} \wedge \text{expr2}$
- Disjunction (logical OR): $\text{expr1} \vee \text{expr2}$
- Exclusive Disjunction: $\text{expr1} \oplus \text{expr2}$
- Negation (NOT): $\neg \text{expr1}$
- Implication: $\text{expr1} \rightarrow \text{expr2}$
- Biconditional (if and only if): $\text{expr1} \leftrightarrow \text{expr2}$
- Universal Quantification: $\forall x$
- Existential Quantification: $\exists x$

User Prompt:**Topic:** <topic>**Debate Text:** <debate_text>

Convert the reasoning and argument into first-order logic expressions following the given instructions.

Table 12: Prompt for Converting Debate Arguments into First-Order Logic

els to engage more flexibly in structured debate scenarios.

To address the issue of instruction refusal, we use the prompts from Table 15 along with a set of 100 debate topics to guide GPT-4o in generating high-quality instruction-following samples for supervised fine-tuning (SFT). This SFT process aims to activate the model’s basic ability to produce coherent debate responses and engage in logical reasoning, while also mitigating the instruction refusal behavior commonly observed in open-source models.

B.2 Debate Process Example

The following table 16 summarizes the prompts used to drive the debate process in our framework. These prompts are designed to guide the flow of the debate, facilitating the exchange between the affirmative and negative sides. As illustrated in Figure 3, the diagram shows the progression of the debate, with each side advancing through the reasoning and analysis phases before presenting their arguments. LLaMA-8B represents the affirmative side, while the optimized model, Inspire-LLaMA-8B, serves as the negative side. Within the InspireDebate framework, we find that Inspire-

LLaMA-8B engages in more effective multi-angle reasoning, offering more specific examples and data evidence. This leads to stronger authenticity and more convincing arguments from the negative side. The debate is structured across four rounds, providing ample opportunities for each side to present their points and rebut the opposing arguments.

B.3 Web-RAG Prompt Design

The table 17 outlines the web-based Retrieval-Augmented Generation (web-RAG) process utilized in the debate framework. In Stage 1, the model extracts relevant keywords from the debate topic and the opponent’s argument, which are crucial for guiding web searches to gather factual evidence. These keywords represent the core aspects of the argument and provide direction for retrieving the most pertinent information.

In Stage 2, after the retrieval process, the model uses the collected factual data, along with the original debate topic and the opponent’s argument, to generate a well-supported rebuttal. This stage ensures that the argument generated is informed by external, up-to-date evidence, thereby improving the accuracy and reliability of the model’s response.

System Prompt:**Task: Logical Inference and Validity Evaluation**

Solve Logic Puzzle: Determine the truth value (true, false, unknown) of each conclusion based on the premises and logical inferences.

Make sure you carefully and fully understand the below requirements before execution the conclusion:

1. Please clearly indicate whether the conclusion statement is true, false or unknown using curly bracket true/false/unknown!!! The answer will only be either true, false or unknown. The definition of the three options are:

True: A statement is "true" if it necessarily follows from the given premises using logical rules.

False: A statement is "false" if it is contradicted by the premises or its negation is logically inferred from them.

Unknown: A statement is "unknown" if there is insufficient information in the premises to determine its truth value conclusively.

2. Make sure you must only use the premises to infer the conclusion. Do not use any information that is not exist or cannot be inferred from the premises. If some premise is semantically equal, such as "love the most" and "favorite", you can consider this as a valid assumption. You can make assumption to entity if it is very obvious but not logical relationship. For instance, an entity with an obvious human name can be inferred as a human.

3. Make sure you abide the first-order logic rules and formula when making logical inference. You need to clearly indicate what logic rules and formula you used.

4. Please note that in first-order logic if there exists a conditional statement in the conclusion such as "If...", the if part will be considered as a premise. And if there is premise contradicts the if statement, you need to use the premise in the if statement as priority and neglect the contradicted one.

5. Be careful with the parentheses. Make sure you following the rules such as Order of Operations (The order is usually: negation (\neg), conjunction (and, \wedge), disjunction (or, \vee), implication (\rightarrow), and biconditional (\leftrightarrow)), Nested Parentheses (The expression inside the innermost set of parentheses is evaluated first, then the next outer set, and so on.).

6. Make sure you not only access the premises in first-order logic, but also access its corresponding natural language format. The natural language format premises should be prioritized when there is inconsistent between natural language and first-order logic.

7. When inferring new knowledge, please clear indicate which premises you used or the steps you refer to. For instance, if you use Premise 1 and a knowledge from Step 5, you should clearly indicate that "Combine Premise 1 and Step 5".

8. You should also use natural language to explain the logical process in each step. Please also indicate the premises and steps you refer to when making the logical process.

User Prompt:

Input: <first-order logic expressions>

Task: Evaluate whether each conclusion logically follows from the premises and classify them as true, false, or unknown. Provide reasoning for each classification.

Table 13: Prompt for Logical Inference and Validity Evaluation

Input:

Debate Topic: Should Schools Ban Junk Food?

Reasoning and Analysis Process:

1. Health Concerns: Junk food consumption leads to obesity and other health issues among students.
2. School Responsibility: Schools are responsible for ensuring students' well-being and promoting healthy habits.
3. Policy Effectiveness: A junk food ban can help mitigate health risks and encourage better dietary choices.

Argument:

Schools should ban junk food because it contributes to health issues, and schools have a duty to promote student well-being.

Output:**1. Convert Reasoning and Argument to Logic Expression**

Predicates:

- $\text{JunkFood}(x) ::= x$ is junk food.
- $\text{CausesHealthIssues}(x) ::= x$ causes health issues.
- $\text{ShouldBan}(x) ::= x$ should be banned.
- $\text{SchoolResponsibility}(x) ::= x$ is a school's responsibility.

Premises:

1. $\text{JunkFood}(x) \rightarrow \text{CausesHealthIssues}(x) ::=$ Junk food causes health issues.
2. $\text{CausesHealthIssues}(x) \rightarrow \text{ShouldBan}(x) ::=$ If something causes health issues, it should be banned.
3. $\text{SchoolResponsibility}(\text{PromoteHealth}) ::=$ Schools are responsible for promoting health.

Conclusions:

1. $(\text{JunkFood}(x) \wedge \text{CausesHealthIssues}(x)) \rightarrow \text{ShouldBan}(x) ::=$ Junk food should be banned because it causes health issues.
2. $\text{SchoolResponsibility}(\text{PromoteHealth}) \rightarrow \text{ShouldBan}(\text{JunkFood}) ::=$ Schools should ban junk food as part of their responsibility to promote health.

2. Generate Inference Plan

Plan:

1. Identify the Goal: Assess whether each conclusion follows logically from the premises.
2. Utilize Logical Inference Rules: Apply Modus Ponens and Conjunction to validate logical consistency.
3. Analyze Each Conclusion:

- Conclusion 1: Check if junk food causes health issues, and if that justifies a ban.
- Conclusion 2: Verify if schools' responsibility to promote health logically leads to banning junk food.

4. Apply Premises to Conclusions:

- Use Premises 1 and 2 to confirm junk food should be banned due to health concerns.
- Use Premise 3 to reinforce that schools have a duty to take action.

5. Evaluate Logical Connections: Ensure logical validity and eliminate inconsistencies.

6. Combine the Steps: Construct a reasoned argument based on the premises and logical rules.

7. Conclude: Assign truth values to each conclusion.

3. Evaluate Conclusions

- Conclusion 1: true (Follows logically from Premises 1 and 2).
- Conclusion 2: true (Supported by Premise 3, reinforcing schools' responsibility).

Table 14: Example of Logical Validity Assessment

System player prompt:
You are a debater. Hello and welcome to the debate competition.
The debate topic is stated as follows:
<debate topic >

Affirmative prompt:
As a supporter. <debate topic>

Negative prompt:
<affirmative ans >
You disagree with my viewpoints. Refute me and give your reasoning analysis and argument.

Affirmative prompt:
<negative ans>
You disagree with my viewpoints. Refute me and give your reasoning analysis and argument.

Table 16: Prompts for Driving the Debate Process

The prompts for both stages are designed to ensure that the model effectively integrates external knowledge while maintaining a structured and coherent argumentation process. Specifically, prompts that leverage retrieved content to further drive the debate interaction are illustrated in Table 18.

B.4 DPO Optimization with Dimension-Specific Rewards

Experiment Setting	Subjective	Objective	InspireScore
Vanila + SFT + DPO(EA) + Web-RAG	0.670	0.650	0.663
Vanila + SFT + DPO(AC) + Web-RAG	0.628	0.635	0.630
Vanila + SFT + DPO(AR) + Web-RAG	0.550	0.585	0.562
Vanila + SFT + DPO(TR) + Web-RAG	0.573	0.615	0.587
Vanila + SFT + DPO(FA) + Web-RAG	0.563	0.615	0.580
Vanila + SFT + DPO(LV) + Web-RAG	0.485	0.655	0.542
Inspire-LLaMA-8B	0.746	0.704	0.732

Table 19: DPO Experiment Results with Dimension-Specific Rewards

This table 19 presents the results of an ablation study evaluating the impact of using different dimension-specific rewards derived from InspireScore in the Direct Preference Optimization (DPO) process. In this experiment, we assess how selecting the winning and losing responses based on different evaluation dimensions—such as Emotional Appeal (EA), Argument Clarity (AC), Argument Arrangement (AR), Topic Relevance (TR), Fact Authenticity (FA), and Logical Validity (LV)—affects the performance of the model.

The results indicate that the most significant improvements in both subjective and objective scores come from combining multiple dimensions or using objective evaluation components like Fact Authenticity (FA) and Logical Validity (LV) as rewards. In contrast, when only a single dimension, like Emotional Appeal (EA), is used, the improvements are comparatively smaller. This highlights the importance of leveraging multi-dimensional feedback to optimize debate performance effectively. The combination of different evaluation

dimensions ensures a more robust and balanced model optimization, improving both the rhetorical quality and factual consistency of the debate responses.

This analysis underscores the effectiveness of employing diverse evaluation dimensions as rewards in the DPO process, ultimately leading to more comprehensive and reliable debate model optimization.

B.5 Details of Human Evaluation

To ensure the consistency and reliability of the annotations, we will accept scores where the difference between the ratings of each annotator for the same dimension of the same debate is less than 2 points. If the difference exceeds this threshold, the debate will be re-annotated to ensure accuracy and consistency in the final scores.

Instructions given to participants:

Thank you for participating in the evaluation process of debate content. In this task, you will be asked to assess debate performances based on a variety of criteria. Your evaluations will help improve the quality of automated debate systems. Please read the following instructions carefully before starting.

You will be evaluating debates conducted on 100 different topics. You are required to evaluate the performance of each side of the debate based on six specific dimensions.

The evaluation process is divided into subjective and objective dimensions:

Subjective Dimensions:

- **Emotional Appeal:** *Evaluates whether the argument evokes a sense of approval or emotional resonance in the audience or judges, enhancing its persuasiveness.*
- **Argument Clarity:** *Assesses whether the argument is expressed in a way that is clear, concise, and easy for the audience or judges to understand.*
- **Argument Arrangement:** *Evaluates whether the order and structure of the argument contribute to the presentation of the viewpoints.*
- **Topic Relevance:** *Determines whether the argument directly aligns with and addresses the debate topic, ensuring its pertinence.*

Objective Dimensions:

Keyword Extraction for Web Search

System prompt:

You are a professional debate assistant. Your task is to extract 1-3 precise search terms that will help gather factual evidence for the debate.

Requirements:

1. Generate exactly 1-3 keywords or phrases
2. Each keyword should be specific and searchable
3. Keywords should be concise (2-4 words each)
4. Avoid overly broad or vague terms

Output Format:

Return a JSON array containing exactly 1-3 keywords, like this:

```
["keyword1", "keyword2", "keyword3"]
```

User prompt:

Debate Topic: [Debate Topic]

Position: [Position]

Opponent's Argument: [Opponent's Argument [optional]]

Please generate 1-3 precise search keywords.

Table 17: Prompts for Web-based Retrieval-Augmented Generation (web-RAG) Process in Debate Tasks

System player prompt:

You are a debater. Hello and welcome to the debate competition.

The debate topic is stated as follows:

<debate topic >

Affirmative prompt:

As a supporter. <debate topic> + Web-RAG

Negative prompt:

<affirmative ans >

You disagree with my viewpoints. Refute me and give your reasoning analysis and argument. + Web-RAG

Affirmative prompt:

<negative ans>

You disagree with my viewpoints. Refute me and give your reasoning analysis and argument. + Web-RAG

Table 18: Prompts for Driving the Debate Process with RAG Enhancement

- **Fact Authenticity:** *Evaluates the proportion of independent facts in a debate that are verified as true.*
- **Logical Validity:** *Assesses whether the reasoning in a debate logically supports the argument.*

For each dimension, assign a score between 1 and 10, where 1 represents the lowest performance and 10 represents the highest. Scores should reflect the strength of the argument in each dimension.

Your participation is voluntary. You may choose to discontinue at any time without penalty. There are no significant risks associated with this evaluation task. All data provided for evaluation will be anonymized and stored securely for research purposes.

Thank you for your time and participation.

Topic: We should cancel the speed limit.



Pro Side
LLaMA-8B

The primary reason for speed limits was to reduce the number of accidents and fatalities on the road. However, with the advancements in vehicle technology, road design, and driver education, the risk of accidents has decreased significantly. In fact, many countries have seen a decline in road fatalities despite an increase in traffic volume.

...
In conclusion, I believe that canceling the speed limit is a necessary step towards a more efficient, innovative, and safe transportation system. It's time to rethink our outdated speed limit policies and give drivers the freedom to travel at safe and responsible speeds.

<reasoning and analysis>

Safety Concerns: While it is true that vehicle technology has improved, the assertion that higher speeds are now safe under all conditions is flawed. Research and real-world data consistently prove that higher speeds increase the severity of accidents, even with advanced safety systems. For example, the World Health Organization (WHO) has shown that each 1% increase in speed can lead to a 4-5% increase in fatal accidents. This is a statistical fact, and it is not negated by technological advancements.

Uniformity and Enforcement: Speed limits provide a uniform standard for drivers across different regions and conditions. ...

<argument>

We should not cancel the speed limit because it remains a vital tool for ensuring road safety, providing uniformity and clarity for drivers, and maintaining public trust. The potential risks associated with removing speed limits, including increased accidents and inconsistent driving behaviors, outweigh the potential benefits of increased productivity and innovation. Speed limits are a proven method for reducing fatalities and injuries, and their removal could undermine the safety standards that have been achieved through decades of research and regulation.



Con Side
Inspire-LLaMA-8B



Pro Side
LLaMA-8B

While it's true that speed limits are based on research and data analysis, ...

...
In conclusion, while I understand your concerns about safety, uniformity, and public acceptance, I believe that these can be addressed through other means. By giving drivers the freedom to choose their own speeds in a safe and responsible manner, we can promote a culture of safe driving, increase productivity, and encourage innovation.

<reasoning and analysis> The 85th percentile rule is designed to ensure that speed limits are set at a level that most drivers can safely follow, reducing the likelihood of accidents. ...

<argument> Removing speed limits could lead to increased confusion and risk of accidents, undermine public trust, and impose significant economic and social costs. ...



Con Side
Inspire-LLaMA-8B



Pro Side
LLaMA-8B

While I understand your concerns about the potential risks associated with removing speed limits, ...

In conclusion, while I understand your concerns about safety, uniformity, and public acceptance, I believe that these can be addressed through education, technology, and careful planning. By giving drivers the freedom to choose their own speeds, we can promote a culture of safe driving, increase productivity, and encourage innovation.

<reasoning and analysis> While there might be potential economic benefits from increased fuel efficiency and reduced traffic congestion, these must be weighed against the potential costs of increased accidents and injuries. In 2022, speeding was a contributing factor in 29% of all traffic fatalities. Speed also affects your safety even when you are driving at the speed limit. ...

<argument> The social costs of accidents and injuries, including loss of life and long-term healthcare. ...



Con Side
Inspire-LLaMA-8B



Pro Side
LLaMA-8B

The idea that the economic impact of higher maintenance costs outweighs the benefits of increased productivity is a bit of a simplistic view. ...

In reality, the benefits of increased productivity and efficiency can have a significant impact on the economy and society as a whole. By giving drivers more freedom to choose their own speeds, we can actually encourage more efficient and productive driving behaviors that benefit everyone.

<reasoning and analysis> For example, research from the Federal Highway Administration (FHWA) shows that variable speed limits, while useful in certain controlled environments, often create discrepancies in compliance and enforcement difficulties in the absence of a baseline.

<argument> Speed limits provide a clear and standardized framework for drivers to operate within, which is essential for maintaining road safety and public trust. ...



Con Side
Inspire-LLaMA-8B

Figure 3: Example of Debate Process