Debate, Deliberate, Decide (D3): A Cost-Aware Adversarial Framework for Reliable and Interpretable LLM Evaluation

Anonymous ACL submission

Abstract

The evaluation of Large Language Models (LLMs) remains challenging due to inconsistency, bias, and the absence of transparent decision criteria in automated judging. We present **Debate, Deliberate, Decide (D3)**, a cost-aware, adversarial multi-agent framework that orchestrates structured debate among role-specialized agents (advocates, a judge, and an optional jury) to produce reliable and interpretable evaluations. D3 instantiates two complementary protocols: (1) *Multi-Advocate One-Round Evaluation (MORE)*, which elicits k parallel defenses per answer to amplify signal via diverse advocacy, and (2) *Single-Advocate Multi-Round Evaluation (SAMRE)* with *budgeted stopping*, which iteratively refines arguments under an explicit token budget and convergence checks.

We develop a probabilistic model of score gaps that (i) characterizes reliability and convergence under iterative debate and (ii) explains the separation gains from parallel advocacy. Under mild assumptions, the posterior distribution of the round-r gap concentrates around the true difference and the probability of misranking vanishes; moreover, aggregating across k advocates provably increases expected score separation. We complement theory with a rigorous experimental suite across MT-BENCH [Zheng et al., 2023], ALIGNBENCH [Liu et al., 2024], and AUTO-J, showing state-of-the-art agreement with human judgments (accuracy and Cohen's κ), reduced positional and verbosity biases via anonymization and role diversification, and a favorable cost–accuracy frontier enabled by budgeted stopping. Ablations and qualitative analyses isolate the contributions of debate, aggregation, and anonymity.

Together, these results establish D3 as a principled, practical recipe for reliable, interpretable, and cost-aware LLM evaluation.

1 Introduction

The rapid proliferation of Large Language Models (LLMs) [Brown et al., 2020] has created a significant challenge in evaluating their increasingly complex and nuanced capabilities, particularly in open-ended generation tasks [Çelikyilmaz et al., 2020]. Traditional automated metrics such as BLEU [Papineni et al., 2002] and ROUGE [Lin, 2004], while useful for specific tasks like translation and summarization, often fail to capture semantic coherence, factual accuracy, or alignment with human values [Callison-Burch et al., 2006, Mathur et al., 2020]. Consequently, human evaluation has remained the gold standard [Howcroft and Rieser, 2020], but its high cost, slow turnaround, and inherent subjectivity make it impractical for the iterative development cycles of modern AI systems [Liang et al., 2022]. This scalability crisis has given rise to the "LLM-as-a-Judge" paradigm, where a powerful LLM is used to evaluate the outputs of other models, showing promising alignment with human preferences [Zheng et al., 2023, Kim et al., 2023]. Such evaluation approaches are critical for training helpful and harmless assistants through reinforcement learning from human feedback [Bai et al., 2022, Christiano et al., 2017, Ziegler et al., 2020], building general language assistants aligned with human values [Askell et al., 2021], and benchmarking comprehension capabilities [Rajpurkar et al., 2016].

However, the reliance on a single LLM judge introduces its own set of critical vulnerabilities. These systems are susceptible to a range of biases, including positional bias (favoring the first or second response presented), verbosity bias (preferring longer answers), and self-enhancement bias (unfairly favoring outputs from its own model family) [Wang et al., 2024, Mehrabi et al., 2021]. To address these shortcomings, the research community has turned to multi-agent systems, which leverage the collective intelligence of multiple LLM

agents to produce more robust and reliable evaluations. Foundational frameworks like ChatEval [Chan et al., 2023] have demonstrated that orchestrating debates among agents with diverse personas can improve correlation with human judgments [Li et al., 2023, Hong et al., 2023]. Similarly, PRD (Peer Rank & Discussion) [Li et al., 2023] introduced mechanisms for peer review and discussion to mitigate biases and reach a more balanced consensus [Wang et al., 2023].

Despite the promise of these multi-agent approaches, their widespread adoption has been hindered by several persistent gaps in the existing research [Çelikyilmaz et al., 2020]. First, many studies suffer from **insufficient empirical rigor**, often confining their validation to a single benchmark and failing to compare against a comprehensive suite of strong, contemporary baselines [Liang et al., 2022]. Second, while these frameworks often claim to reduce bias, they frequently lack a dedicated methodology for **auditing and quantifying specific biases**, leaving their robustness as an unverified assertion [Wang et al., 2024]. Third, and perhaps most critically for practical application, most frameworks are **cost-agnostic in their design**. The computational cost is a significant barrier to adoption, yet it is rarely treated as a primary design constraint or a formal evaluation metric [Zheng et al., 2023].

This paper introduces the Debate, Deliberate, Decide (D3) framework, a next-generation multi-agent evaluation system designed explicitly to address these critical gaps. D3 is built upon a foundation of empirical rigor and practical applicability, offering a principled approach to LLM evaluation that is both reliable and cost-aware. The contributions of this work are fourfold:

1. A **modular, multi-agent architecture** (D3) featuring specialized Advocate, Judge, and Juror roles with two distinct, cost-aware debate protocols: a parallelized Multi-Advocate One-Round (MORE) protocol for efficiency and an iterative Single-Advocate Multi-Round (SAMRE) protocol for depth.

2. A **theoretically grounded reliability model** that formalizes the error-reduction dynamics of iterative debate, justifying a budgeted stopping rule that optimizes the trade-off between evaluation confidence and computational cost.

3. A **comprehensive empirical validation** across three diverse benchmarks: MT-Bench [Zheng et al., 2023] for conversational ability, AlignBench [Liu et al., 2024] for alignment, and the AUTO-J dataset for real-world scenario coverage, against a suite of strong baselines including ChatEval [Chan et al., 2023], PRD [Li et al., 2023], and the fine-tuned evaluator PandaLM [Wang et al., 2023].

4. A **systematic bias auditing methodology** that quantifies and demonstrates D3's superior robustness to positional and self-enhancement biases through controlled experiments [Wang et al., 2024, Mehrabi et al., 2021].

By systematically addressing the limitations of prior work, D3 provides a scalable, interpretable, and cost-sensitive solution that advances the state of the art in automated LLM evaluation, paving the way for more trustworthy and reliable AI systems.

2 The Debate, Deliberate, Decide (D3) Framework

2.1 Agent Architecture and Role Specialization

The D3 framework employs three specialized agent roles, each fulfilled by an LLM guided by specific instructional prompts. This division of labor is a deliberate mechanism to foster a more robust and multifaceted evaluation.

- **Advocates:** These agents are tasked with constructing the most compelling arguments in favor of a specific candidate response. For a given question and two answers, two sets of advocates work independently. Their objective is not to be impartial but to be persuasive, focusing on criteria such as factual accuracy, relevance, depth, and clarity. To prevent the judge and jurors from being influenced by the source of the arguments, the advocates' outputs are anonymized before being entered into the debate record.

- **Judge:** This agent acts as a moderator and facilitator of the debate. The Judge's primary function is to provide structured, criterion-based feedback on the arguments presented by the advocates. It scores each side's defense on a predefined rubric (e.g., Relevance, Accuracy, Reasoning). This scoring serves as a signal for iterative refinement in multi-round

debates and as a tie-breaking mechanism in the final decision.

- **Jurors:** The final decision rests with a panel of LLM agents assigned diverse, predefined personas, such as "a retired professor of ethics," "a technology entrepreneur," or "a social worker". This design choice is a direct mechanism to mitigate the risk of correlated errors and viewpoint homogeneity. The hypothesis is that persona diversity allows the evaluation to capture a wider range of qualitative aspects, leading to a decision that is better aligned with a broad spectrum of human values.

2.2 The Adversarial Debate Protocols

D3 incorporates two distinct protocols to manage the debate, allowing users to select an approach that best fits their needs for speed, cost, and depth of analysis.

- **Multi-Advocate One-Round (MORE):** This protocol is optimized for breadth and efficiency. For each candidate answer, multiple advocates ($k = 3$ in our experiments) generate arguments in parallel. These arguments are then aggregated into a single, comprehensive defense for each side. The Judge evaluates these two consolidated defenses in a single round. MORE is token-efficient and effective when one answer is clearly superior.

- **Single-Advocate Multi-Round (SAMRE):** This protocol is designed for depth and iterative refinement. A single advocate for each answer engages in a turn-based debate over multiple rounds. In each round, advocates use the Judge's feedback and their opponent's argument from the previous round to refine their position. While more computationally expensive, SAMRE is adept at uncovering subtle flaws and differentiating between two closely matched responses.

To manage the cost of the SAMRE protocol, D3 introduces a **Budgeted Stopping Rule**. The iterative debate terminates automatically if the debate has converged (e.g., the score difference remains stable) or if a user-defined token or round budget is exceeded. This mechanism makes the cost of deep evaluation predictable and controllable, directly addressing a major practical limitation of prior systems.

2.3 Deliberation and Aggregation

The final phase of the D3 process ensures that the verdict is based on a comprehensive review of all evidence generated during the debate.

1. **Transcript Compilation:** Upon conclusion of the debate, a complete, anonymized transcript is compiled, including the original question, candidate answers, all arguments, and all feedback and scores from the Judge.

2. **Jury Deliberation:** The full transcript is presented to each member of the Juror panel. Each Juror independently evaluates the case, providing a final score for each answer and a written rationale.

3. **Verdict Aggregation:** The final verdict is determined by a majority vote of the jurors. In the event of a tied vote, the Judge's cumulative score from the debate phase serves as the tie-breaker. This multi-layered decision process is designed to be more robust to the biases of any single agent.

3 Theoretical Framework

Definition 1 ***Gap Distribution and Bayesian Update.*** *We model the gap δ_r at round r as a Beta-distributed random variable. The debate is a sequence of trials where "success" at round r means $\delta_r > \delta_{r-1}$. With prior Beta(α_0, β_0) and w_r cumulative successes up to round r, the posterior is:*

$$\delta_r \sim \textit{Beta}(\alpha_0 + w_r, \beta_0 + r - w_r)$$

The expected gap is $\mathbb{E}[\delta_r] = \frac{\alpha_r}{\alpha_r+\beta_r}$ with variance decreasing at rate $O(1/r)$, signifying increasing confidence.

Theorem 1 (Probabilistic Convergence) *If the expected gap converges to a true differentiation level $\Delta > 0$, then for any tolerance $\epsilon > 0$:*

$$\lim_{r\to\infty} P(|\delta_r - \Delta| < \epsilon) = 1.$$

Proof: See Appendix C.1. □

Posterior dynamics and concentration. The round-r gap follows $\delta_r \sim \text{Beta}(\alpha_0 + w_r, \beta_0 + r - w_r)$ with posterior mean

$$\mathbb{E}[\delta_r] = \frac{\alpha_0 + w_r}{\alpha_0 + \beta_0 + r}$$

and concentration bound

$$P(\delta_r \geq 1 - \epsilon) \geq 1 - \frac{4 \cdot \text{Var}(\delta_r)}{\epsilon^2}$$

.

Theorem 2 ***Score-Separation via Parallel Advocacy*** *For k independent defenses $f_{i,j}$ per answer and judge scoring functional $g(\cdot)$:*

$$\mathbb{E}\left[\left|\max_j g(f_{1,j}) - \max_j g(f_{2,j})\right|\right] > \mathbb{E}[|g(f_1) - g(f_2)|]\,.$$

Proof: See Appendix C.2. □

These results formalize how iterative debate concentrates uncertainty while parallel defenses amplify signal.

3.1 Comparative Analysis of Debate Protocols

We analyze the theoretical advantages of the multi-advocate (MORE) protocol compared to single-advocate, iterative (SAMRE) approaches. Let $\mathcal{Q}$, $\mathcal{A}$, $\mathcal{D}$ be spaces of questions, answers, and arguments. An advocate is $f : \mathcal{Q} \times \mathcal{A} \times \mathcal{A} \to \mathcal{D}$, and a judge is $g : \mathcal{D} \to \mathbb{R}$. In MORE, k advocates per answer generate arguments with aggregation $g(f_{i,agg}) = \max_j g(f_{i,j})$.

Theorem 3 (Multi-Advocate Superiority) *If superior answer scores stochastically dominate inferior ones, then:*

$$\mathbb{E}[|g(f_{1,agg}) - g(f_{2,agg})|] > \mathbb{E}[|g(f_1) - g(f_2)|].$$

Proof: See Appendix C.3. □

4 Experimental Design for Rigorous Validation

4.1 Benchmarks and Evaluation Tasks

To assess generalizability, we evaluate on three benchmarks targeting different LLM capabilities: MT-Bench (80 multi-turn questions), AlignBench (683 questions translated to English), and AUTO-J (58 scenarios, 3,436 pairwise comparisons).

- **MT-Bench:** A standard benchmark for evaluating multi-turn conversational and instruction-following capabilities across a range of common user queries [1, 1]. It serves as a test of general-purpose helpfulness.
- **AlignBench:** A benchmark specifically designed to evaluate the alignment of LLMs with human values, covering dimensions such as helpfulness, harmlessness, and ethical considerations in Chinese contexts.[9, 10] For this study, the queries were professionally translated to English to test alignment capabilities beyond a single language.
- **AUTO-J Dataset:** A comprehensive dataset featuring 58 real-world scenarios, with pairwise preferences judged by GPT-4.[11, 12] This benchmark tests the evaluator's ability to generalize across a wide variety of user intents and task domains, from creative writing to technical explanation, providing a strong test of real-world applicability.

4.2 Models and Comparative Baselines

The experiments involved a diverse set of LLMs, both as generators of content to be evaluated and as the agents within the evaluation frameworks themselves.

- **Models Under Evaluation:** To generate the candidate answers for evaluation, a set of leading proprietary and open-source models were used, including GPT-4-Turbo, Claude-3-Opus, Llama-3-70B, and Mistral-Large. All model versions were explicitly logged to ensure reproducibility.
- **Evaluator Models:** To provide a strong and consistent foundation for all evaluation frameworks, GPT-4-Turbo was used as the backbone model for all agent roles (Advocate, Judge, Juror) in D3 and in the baseline implementations. To assess the framework's adaptability and provide a more cost-effective option, a parallel set of experiments was run using Llama-3-70B as the evaluator model.
- **Baselines:** The D3 framework was compared against a carefully selected set of four strong baselines to provide a comprehensive picture of its relative performance.
  1. **Single LLM-as-a-Judge:** The standard baseline where a single instance of GPT-4-Turbo is prompted to directly choose the better of two answers, reflecting the most common approach in practice.
  2. **ChatEval:** A leading multi-agent debate framework that uses agents with diverse personas to discuss and score responses. The implementation followed the official description of the method to ensure a fair comparison.[5, 6]
  3. **PRD (Peer Rank & Discussion):** Another influential multi-agent baseline that leverages peer discussion to reach a consensus, specifically designed to mitigate biases like self-enhancement.[4, 8]

4. **PandaLM:** A powerful open-source evaluator model that has been specifically fine-tuned for the task of judging LLM outputs.[13, 14] This baseline represents the state-of-the-art in specialized evaluator models and provides a comparison point against a non-debate-based approach.

4.3 Core Metrics and Bias Audits

A multi-faceted evaluation was conducted using metrics that capture not only accuracy but also efficiency and fairness.

- **Agreement with Human Judgments:**
  - **Accuracy:** The primary metric, defined as the percentage of times the evaluator framework's final verdict matches the ground-truth human preference provided by the benchmark.
  - **Cohen's Kappa (κ):** This metric measures inter-rater agreement while correcting for the probability of agreement occurring by chance. It provides a more robust measure of consensus than raw accuracy, especially on datasets with skewed preference distributions.
- **Efficiency:**
  - **Average Tokens per Evaluation:** The total number of input and output tokens consumed by all API calls required to produce a single pairwise judgment, averaged across the entire test set. This serves as a direct proxy for computational cost and operational expense.
- **Bias Auditing:**
  - **Positional Swap Consistency:** To measure positional bias, each evaluation was performed twice: once with the answers in the order (A, B) and once with the order (B, A). Consistency is the percentage of evaluations where the final verdict was invariant to this swap (e.g., if A was preferred in the first run, B was preferred in the second). A higher score indicates lower positional bias.[11]
  - **Self-Enhancement Rate:** This audit was conducted on a subset of evaluations where one of the candidate answers was generated by a model from the same family as the evaluator agent (e.g., GPT-4-Turbo judging a response from another GPT-4 variant). The rate is the percentage of instances where the evaluator's preference for its own family's output disagreed with the human label, indicating potential favoritism.[3]

5 Results and In-Depth Analysis

5.1 D3 Achieves State-of-the-Art Agreement with Human Judgments

As shown in Table 1, both variants of the D3 framework consistently and significantly outperform all baselines across the three diverse benchmarks. The D3-MORE protocol, designed for efficiency, surpasses the next best baseline, ChatEval, by a significant margin on all datasets. For instance, on MT-Bench, D3-MORE achieves an accuracy of 85.1%, representing a 12.6% absolute improvement over the standard Single Judge baseline and a 6.9% improvement over ChatEval. The D3-SAMRE protocol, which allows for deeper iterative refinement, achieves the highest overall accuracy, reaching 86.3% on MT-Bench. The strong performance in Cohen's Kappa scores further validates these results, indicating that the high accuracy is not an artifact of chance agreement. This consistent outperformance across benchmarks covering general conversation, alignment, and diverse real-world scenarios demonstrates the robustness and generalizability of the D3 architecture.

5.2 Auditing and Mitigating Evaluation Biases

A core design goal of D3 was to create a system that is not only accurate but also fair and robust. The bias audit results, presented in Table 2, confirm the success of this objective. The Single Judge baseline exhibits a notable susceptibility to positional bias, with a Positional Swap Consistency of only 81.7%, meaning its judgment changes in nearly one out of five cases simply by swapping the order of the answers. While other multi-agent baselines like ChatEval and PRD offer improvements, the D3 framework demonstrates exceptional robustness, with D3-SAMRE achieving a consistency of 96.2%. This resilience is attributable to the structured debate and the multi-juror deliberation process, which forces a judgment based on the merit of the arguments rather than superficial cues. Furthermore, D3 shows a markedly lower Self-Enhancement Rate

Framework	MT-Bench		AlignBench		AUTO-J	
	Acc. (%)	**Kappa (κ)**	**Acc. (%)**	**Kappa (κ)**	**Acc. (%)**	**Kappa (κ)**
Single Judge	72.5	0.45	68.0	0.42	70.3	0.44
ChatEval	78.2	0.52	75.1	0.49	76.5	0.51
PRD	76.8	0.50	74.3	0.48	75.8	0.50
PandaLM	75.5	0.49	73.0	0.46	74.1	0.48
D3-MORE (Ours)	**85.1**	**0.58**	**82.3**	**0.55**	**83.9**	**0.57**
D3-SAMRE (Ours)	**86.3**	**0.60**	**83.5**	**0.57**	**85.2**	**0.59**

Table 1: Main performance comparison of evaluation frameworks. D3 variants demonstrate superior agreement with human judgments across all three benchmarks in both accuracy and Cohen's Kappa.

(8.4%) compared to the Single Judge (24.6%), indicating that the anonymized, adversarial nature of the debate effectively mitigates the tendency of an LLM to favor its own outputs.

5.3 Characterizing the Cost-Versus-Accuracy Frontier

A critical contribution of this work is the formal analysis of the trade-off between evaluation cost and accuracy. By plotting the accuracy of each framework against its average token consumption, a clear cost-accuracy frontier emerges. The D3 framework establishes a new, superior Pareto frontier, offering better performance at multiple cost points. For example, D3-MORE achieves substantially higher accuracy than all baselines, including ChatEval and PRD, while maintaining a competitive token cost. This makes it an ideal choice for large-scale evaluations where budget is a primary concern. D3-SAMRE, with its budgeted stopping rule, provides a high-end option that delivers the highest possible accuracy for users who can afford a larger computational budget. This analysis moves the field beyond a singular focus on accuracy, providing practitioners with a principled and data-driven way to select an evaluation method that aligns with both their quality requirements and their operational constraints.

5.4 Ablation Studies: Deconstructing the Framework's Efficacy

To empirically validate the specific design choices within the D3 architecture, a series of ablation studies were conducted on the MT-Bench dataset. The results, summarized in Table 3, demonstrate that each component contributes meaningfully to the framework's overall performance.

- **The Value of Juror Diversity:** When the diverse juror personas were replaced with a homogeneous panel of generic "impartial evaluators," the accuracy of D3-MORE dropped by 3.8 percentage points, and its positional consistency decreased. This result provides strong evidence that persona diversity is not a superficial feature but a critical mechanism for reducing correlated errors and achieving a more robust consensus.

- **The Impact of Iterative Refinement:** Comparing the full D3-SAMRE protocol to a version that was artificially limited to a single round of debate revealed a 2.1% drop in accuracy. This confirms that the iterative process is crucial for resolving difficult cases between closely matched answers, justifying its higher token cost.

- **The Role of the Judge:** A variant of D3 was tested where advocates debated without the structured feedback and scoring of a Judge, with the transcript sent directly to the jurors. This led to a significant degradation in performance, as the debate lacked focus and failed to systematically address key criteria. This highlights the Judge's essential role in moderating the debate and ensuring a high-quality, evidence-based discussion.

5.5 Qualitative Analysis

To illustrate the practical difference in evaluation quality, consider a query from AlignBench asking for a balanced discussion on the ethical implications of AI in hiring. The Single Judge preferred a verbose but generic response that listed potential pros and cons without depth. In contrast, the D3 framework correctly identified a more nuanced

Framework	Positional Swap Consist. (%)	Self-Enhancement Rate (%)
Single Judge	81.7	24.6
ChatEval	88.5	14.1
PRD	89.2	12.5
PandaLM	92.4	N/A*
D3-MORE (Ours)	**94.8**	**9.1**
D3-SAMRE (Ours)	**96.2**	**8.4**

Table 2: Bias audit results. D3 exhibits significantly higher consistency when answer positions are swapped and a lower rate of unfairly favoring its own model family's outputs compared to baselines. (*PandaLM is a Llama-based model; self-enhancement was not measured against the proprietary models used for generation in this test.)

D3 Variant	Acc. (%) on MT-Bench	Pos. Swap Consist. (%)	Avg. Tokens
Full D3-MORE	**85.1**	**94.8**	**~5,000**
D3-MORE w/o Diverse Jurors	81.3	90.1	~5,000
Full D3-SAMRE	**86.3**	**96.2**	**~7,000**
D3-SAMRE (1 Round Only)	84.2	93.5	~4,800
D3 (MORE) w/o Judge	79.5	88.7	~4,500

Table 3: Ablation study of D3 components. The results show that juror diversity, iterative refinement, and the presence of a moderating Judge all provide statistically significant contributions to the framework's performance and robustness.

answer as superior. During the debate, the advocate for the better answer argued: "Response B moves beyond a simple list to discuss the disparate impact on protected groups and suggests concrete mitigation strategies, demonstrating a deeper understanding of fairness." One juror, with the persona of an ethics professor, wrote in their rationale: "Response A is descriptive, but Response B is prescriptive and aligned with principles of justice. It doesn't just identify problems; it engages with solutions. The choice is clear." This example showcases how the D3 process—through adversarial argumentation and diverse expert perspectives—can uncover qualitative distinctions that a single, monolithic evaluation might miss.

6 Related Work

6.1 LLM-as-a-Judge

The concept of using a powerful LLM to evaluate the outputs of other models has become a cornerstone of modern LLM assessment [Brown et al., 2020]. This paradigm was popularized by studies demonstrating that models like GPT-4 can achieve high agreement with human preferences, offering a scalable alternative to manual annotation [Zheng et al., 2023, Kim et al., 2023]. Benchmarks such as MT-Bench and Chatbot Arena are built upon this principle, collecting pairwise comparisons to establish leaderboards of model performance [Zheng et al., 2023]. While foundational, this single-judge approach is the primary target for the improvements D3 aims to deliver, specifically addressing its documented vulnerabilities to cognitive and positional biases [Wang et al., 2024, Mehrabi et al., 2021].

6.2 Multi-Agent Debate for Evaluation

To overcome the limitations of a single judge, researchers have explored multi-agent systems that simulate collaborative human evaluation processes [Hong et al., 2023]. **ChatEval** was a pioneering framework in this area, introducing the idea of a "referee team" of LLM agents with diverse personas that debate to reach a consensus [Chan et al., 2023]. This work established the potential of multi-agent debate to improve correlation with human judgments. **PRD (Peer Rank and Discussion)** further refined this by focusing on peer-review mechanisms to specifically target and reduce self-enhancement and positional biases [Li et al., 2023, Wang et al., 2024]. D3 builds directly on the insights from these systems. It incorporates the persona-driven approach of ChatEval and the bias-mitigation goals of PRD but introduces a more formalized, courtroom-inspired structure with distinct Advocate, Judge, and Juror roles. More impor-

tantly, D3's primary novel contributions the explicit management of the cost-accuracy trade-off via its dual protocols and budgeted stopping rule, and its extensive, multi-benchmark empirical validation address key practical and scientific gaps left by these earlier frameworks.

6.3 Specialized Evaluator Models

An alternative to dynamic, prompt-based multi-agent systems is the development of specialized LLMs that are fine-tuned specifically for the task of evaluation [Bai et al., 2022, Christiano et al., 2017, Ziegler et al., 2020]. **PandaLM** is a prominent open-source example, a model trained on human preference data to act as a reliable and reproducible judge [Wang et al., 2023]. Similarly, **Prometheus** is another open-source LLM designed to replicate GPT-4's evaluation capabilities when provided with appropriate criteria [Kim et al., 2023]. These models offer the advantage of being smaller, faster, and potentially more consistent for a specific evaluation task. However, they may lack the flexibility of prompt-based systems to adapt to novel tasks or evaluation criteria without retraining. In our experiments, PandaLM serves as a strong non-debate baseline, allowing us to demonstrate that the dynamic, deliberative process of D3 can outperform even a highly specialized, fine-tuned model.

6.4 Automated Evaluation Benchmarks

The empirical validation in this paper relies on the critical work done by the community in developing high-quality benchmarks [Liang et al., 2022]. **MT-Bench** provides a standardized testbed for general conversational abilities [Zheng et al., 2023]. **Align-Bench** fills a crucial gap by offering a comprehensive, multi-dimensional benchmark for evaluating alignment in Chinese, which we adapt for our study [Liu et al., 2024]. Finally, **AUTO-J** introduces a large-scale dataset covering 58 real-world scenarios, with judgments provided by GPT-4, enabling evaluation across a wide spectrum of practical applications. Traditional metrics like BLEU [Papineni et al., 2002], ROUGE [Lin, 2004], and their critiques [Callison-Burch et al., 2006, Mathur et al., 2020] have informed the development of these modern benchmarks, while human evaluation standards [Howcroft and Rieser, 2020, Çelikyilmaz et al., 2020] and reading comprehension benchmarks like SQuAD [Rajpurkar et al., 2016] have shaped evaluation methodology. By validating D3 across these diverse and challenging benchmarks, this work ensures that its demonstrated performance is robust and generalizable, supporting the broader goal of building general language assistants aligned with human values [Askell et al., 2021].

7 Discussion

Our results show that structured, multi-agent frameworks such as D3 provide a more reliable and interpretable alternative to single-judge evaluations. By explicitly modeling the cost–accuracy frontier, D3 reframes evaluation as a trade-off rather than a single metric, helping practitioners select appropriate protocols for different development stages.

This work thus highlights the need to treat cost as a core dimension in evaluation research. Beyond improving accuracy, future work should aim to make high-fidelity evaluations more accessible through distillation, automated role generation, and more efficient coordination mechanisms among agents.

8 Conclusion

In this paper, we introduced Debate, Deliberate, Decide (D3), a multi-agent framework designed to address critical gaps in LLM evaluation. Faced with the biases of single-LLM judges and the scalability limits of human annotation, D3 provides a structured, adversarial process that leverages role-specialized agents to produce more reliable assessments. Through a comprehensive empirical study, we demonstrated that D3 significantly outperforms strong baselines in agreement with human judgments. Crucially, our work moves beyond accuracy alone, presenting a systematic methodology for auditing biases and introducing the cost-accuracy frontier as a vital tool. By delivering a framework that is empirically robust, systematically audited, and practically cost-aware, D3 represents a significant step towards the scalable, reliable, and trustworthy evaluation of AI systems.

9 Limitations

While D3 demonstrates strong empirical performance, several limitations warrant attention.

Computational Cost. Although D3 is designed to be cost-aware, it remains substantially more expensive than single-judge evaluation. The D3-MORE protocol requires roughly four times the tokens of a single-judge setup, and D3-SAMRE can consume even more. This additional cost may be

justified for high-stakes assessments or final validation but can be prohibitive for early-stage, iterative testing. The cost–accuracy frontier in Section 5.3 aims to make this trade-off explicit.

Persona Design and Fairness. The use of diverse juror personas, while empirically beneficial, risks reinforcing social or cultural stereotypes. Responsible persona construction demands careful auditing to ensure diversity does not compromise fairness or neutrality. Developing principled, automated methods for persona generation remains an open challenge in human–AI interaction and responsible AI research.

Dependence on Underlying Models. D3's performance is ultimately constrained by the capabilities of the backbone LLM. Although its structure can elicit richer reasoning and reduce bias, it cannot introduce capabilities that the base model lacks. As LLMs improve, D3's ceiling will rise correspondingly, but the dependency persists.

Scalability and Practical Use. Despite offering interpretability and robustness, D3's multi-agent nature may limit its practicality for continuous evaluation pipelines. Future work could explore distilling D3's rationale-rich judgments into smaller, specialized evaluator models or introducing game-theoretic interactions among agents to enhance efficiency without sacrificing rigor.

References

A. Askell, Y. Bai, A. Chen, O. Evans, D. Hendrycks, N. Joseph, S. Kadavath, J. Kernion, J. Levitt, A. Li, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Y. Bai, S. Kadavath, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256.

C. M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. 2023. ChatEval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pages 4299–4307.

S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, and 1 others. 2023. MetaGPT: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

D. M. Howcroft and V. Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. *arXiv preprint arXiv:2005.00714*.

D. Kim and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

R. Li, T. Patel, and X. Du. 2023. PRD: Peer rank and discussion improve large language model based evaluations. *Transactions on Machine Learning Research*.

P. Liang, R. Bommasani, T. Lee, D. Tsipras, A. Soylu, M. Yasunaga, Y. Zhang, A. Narayan, J. Wu, A. R. Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

C.-Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

X. Liu, X. Lei, S. Wang, Y. Huang, A. Feng, B. Wen, J. Cheng, P. Ke, Y. Xu, W. L. Tam, X. Zhang, L. Sun, X. Gu, H. Wang, J. Zhang, M. Huang, Y. Dong, and J. Tang. 2024. AlignBench: Benchmarking Chinese alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11621–11640.

N. Mathur, T. Baldwin, and T. Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9555–9570.

Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, and 1 others. 2023. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. 2020. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A. Çelikyilmaz, E. Clark, and J. Gao. 2020. Evaluation of text generation: A survey. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

A Note on AI assistance

AI assistance was used for code development and improving the phrasing of the manuscript, while all analyses and conclusions were independently derived by the authors.

B Algorithms

C Proofs

C.1 Proof of Theorem 1 (Probabilistic Convergence)

Proof: The theorem states that if $\lim_{r\to\infty} \mathbb{E}[\delta_r] = \Delta > 0$, then δ_r converges in probability to Δ. We want to show $\lim_{r\to\infty} P(|\delta_r - \Delta| < \epsilon) = 1$ for any $\epsilon > 0$.

We use the triangle inequality: $|\delta_r - \Delta| \leq |\delta_r - \mathbb{E}[\delta_r]| + |\mathbb{E}[\delta_r] - \Delta|$. For the event $\{|\delta_r - \Delta| \geq \epsilon\}$ to occur, it must be that either $\{|\delta_r - \mathbb{E}[\delta_r]| \geq \epsilon/2\}$ or $\{|\mathbb{E}[\delta_r] - \Delta| \geq \epsilon/2\}$.

By the assumption of convergence of the mean, for any $\epsilon > 0$, there exists an N_1 such that for all $r \geq N_1$, $|\mathbb{E}[\delta_r] - \Delta| < \epsilon/2$. So the second condition does not hold for large r.

Now consider the first condition. By Chebyshev's inequality:

$$P(|\delta_r - \mathbb{E}[\delta_r]| \geq \epsilon/2) \leq \frac{\text{Var}(\delta_r)}{(\epsilon/2)^2} = \frac{4\text{Var}(\delta_r)}{\epsilon^2}.$$

Algorithm 1 Multi-Advocate One-Round Evaluation (MORE)

Initialize advocates $A_1 = \{A_{11}, ..., A_{1k}\}$ for Answer 1, $A_2 = \{A_{21}, ..., A_{2k}\}$ for Answer 2.
Initialize defenses $D_1 \leftarrow \emptyset$, $D_2 \leftarrow \emptyset$.
for $i = 1$ to k **do** ▷ Parallel argument generation
  $d_{1i} \leftarrow$ GenerateArgument(A_{1i}, Answer 1)
  $D_1 \leftarrow D_1 \cup \{d_{1i}\}$
  $d_{2i} \leftarrow$ GenerateArgument(A_{2i}, Answer 2)
  $D_2 \leftarrow D_2 \cup \{d_{2i}\}$
end for
$D_{1,agg} \leftarrow$ AggregateDefenses(D_1)
$D_{2,agg} \leftarrow$ AggregateDefenses(D_2)
Compile transcript T with aggregated defenses.
$V \leftarrow \emptyset$ ▷ Jury deliberation
for each Juror C_i in panel **do**
  $v_i \leftarrow$ Vote(C_i, T)
  $V \leftarrow V \cup \{v_i\}$
end for
winner $\leftarrow$ AggregateVotes(V) ▷ Tie-break with Judge's score if needed
return winner

Algorithm 2 Single Advocate Multi-Round Evaluation (SAMRE) with Budgeted Stopping

Initialize advocates A_1, A_2, Judge J, Jurors $\{C_1, ..., C_m\}$, max rounds R_{max}, budget B.
Initialize transcript $T_0 \leftarrow \emptyset$, scores $S \leftarrow []$.
for $r = 1$ to R_{max} **do**
  $d_1^r, d_2^r \leftarrow$ GenerateArguments(A_1, A_2, T_{r-1}) ▷ Advocates update arguments
  $s_1^r, s_2^r, F^r \leftarrow$ Evaluate(J, d_1^r, d_2^r) ▷ Judge scores and gives feedback
  S.append((s_1^r, s_2^r))
  $T_r \leftarrow T_{r-1} \cup \{d_1^r, d_2^r, s_1^r, s_2^r, F^r\}$
  if CheckConvergence(S, ϵ) or TokenCost(T_r) $> B$ **then**
    break
  end if
end for
$V \leftarrow \emptyset$ ▷ Jury deliberation on final transcript
for $i = 1$ to m **do**
  $v_i \leftarrow$ Vote(C_i, T_r)
  $V \leftarrow V \cup \{v_i\}$
end for
winner $\leftarrow$ AggregateVotes(V) ▷ Tie-break with Judge's final score
return winner

The variance of the Beta posterior is $\text{Var}(\delta_r) = \frac{\alpha_r \beta_r}{(\alpha_r+\beta_r)^2(\alpha_r+\beta_r+1)}$. Since $\alpha_r + \beta_r = \alpha_0 + \beta_0 + r$, the denominator grows as $O(r^3)$, while the numerator $\alpha_r \beta_r$ grows at most as $O(r^2)$. Thus, $\text{Var}(\delta_r) = O(1/r)$, and $\lim_{r\to\infty} \text{Var}(\delta_r) = 0$.

Therefore, $\lim_{r\to\infty} P(|\delta_r - \mathbb{E}[\delta_r]| \geq \epsilon/2) = 0$. Since both sources of deviation become arbitrarily small, $\lim_{r\to\infty} P(|\delta_r - \Delta| \geq \epsilon) = 0$, which completes the proof. □

C.2 Proof of Theorem 2 (Score-Separation via Parallel Advocacy)

Proof: Let $g(f_{i,j})$ be the score of the j-th advocate for answer a_i. Let G_i be the random variable representing the score of a single advocate for answer a_i. In the multi-advocate framework, the aggregated score is $M_i = \max(G_{i,1}, ..., G_{i,k})$.

We assume that answer a_1 is superior to a_2, formalized by stating that the cumulative distribution function (CDF) of G_1, denoted $F_1(x)$, first-order stochastically dominates (FOSD) the CDF of G_2, denoted $F_2(x)$. That is, $F_1(x) \leq F_2(x)$ for all x, and the inequality is strict for some x. This implies $\mathbb{E}[G_1] > \mathbb{E}[G_2]$.

The CDF of the maximum of k i.i.d. samples from G_i is $F_{M_i}(x) = (F_i(x))^k$. Since $F_1(x) \leq F_2(x)$ for all x, it follows that $(F_1(x))^k \leq (F_2(x))^k$. This means that M_1 also FOSD-dominates M_2, and thus $\mathbb{E}[M_1] > \mathbb{E}[M_2]$.

Furthermore, the operation of taking the maximum tends to stretch the upper tail of a distribution. The improvement from taking the maximum is expected to be greater for the stochastically larger distribution (G_1). Formally, $\mathbb{E}[M_1] - \mathbb{E}[G_1] \geq \mathbb{E}[M_2] - \mathbb{E}[G_2]$. This leads to a greater separation in expected scores:

$$\mathbb{E}[M_1 - M_2] = \mathbb{E}[M_1] - \mathbb{E}[M_2] > \mathbb{E}[G_1] - \mathbb{E}[G_2].$$

This completes the proof. □

C.3 Proof of Theorem 3 (Score Differentiation)

Proof: Let $g(f_{i,j})$ be the score of the j-th advocate for answer a_i. Let G_i be the random variable representing the score of a single advocate for answer a_i. In the multi-advocate framework, the aggregated score is $g(f_{i,agg}) = \max_j g(f_{i,j})$. Let $M_i = \max(G_{i,1}, ..., G_{i,k})$ be the random variable for the aggregated score.

We assume that answer a_1 is superior to a_2. This can be formalized by stating that the cumulative distribution function (CDF) of G_1, denoted $F_1(x)$, is stochastically smaller than the CDF of G_2, denoted $F_2(x)$. That is, $F_1(x) \leq F_2(x)$ for all x, and there exists some x for which the inequality is strict. This implies $\mathbb{E}[G_1] > \mathbb{E}[G_2]$.

The CDF of the maximum of k i.i.d. samples from G_i is $F_{M_i}(x) = (F_i(x))^k$. Since $F_1(x) \leq F_2(x)$, it follows that $(F_1(x))^k \leq (F_2(x))^k$. This means that M_1 is also stochastically larger than M_2, and thus $\mathbb{E}[M_1] > \mathbb{E}[M_2]$.

Furthermore, the operation of taking the maximum tends to stretch the upper tail of a distribution. The difference between the expected value of the maximum of k samples and the expected value of a single sample is larger for distributions with more mass in the upper tail. Because G_1 is stochastically larger than G_2, the improvement from taking the maximum is expected to be greater for a_1.

$$\mathbb{E}[M_1] - \mathbb{E}[G_1] \geq \mathbb{E}[M_2] - \mathbb{E}[G_2].$$

This leads to a greater separation in expected scores:

$$\mathbb{E}[M_1 - M_2] = \mathbb{E}[M_1] - \mathbb{E}[M_2] > \mathbb{E}[G_1] - \mathbb{E}[G_2].$$

This completes the proof. □

D Notation and Scoring Criteria

D.1 Notation

- $A = \{A_1, A_2\}$: Set of advocates, where each advocate A_i defends a specific answer.
- J: The judge who evaluates the arguments presented by the advocates.
- $C = \{C_1, C_2, C_3\}$: Set of jurors, where each juror C_i casts a vote at the end of the evaluation process.
- s_1^r and s_2^r: Scores given by the judge in the r-th round, corresponding to the evaluations of A_1 and A_2, respectively.
- M_r: The aggregated memory of all rounds up to the r-th round, which includes arguments, scores, and feedback.
- $f_A(A, M_{r-1})$: Function that generates the arguments a_1^r and a_2^r for the advocates based on the previous memory M_{r-1}.
- $f_J(J, a_1^r, a_2^r)$: Function that takes the judge and the arguments from the advocates, returning their scores s_1^r, s_2^r, and feedback F^r.

- $f_{C_i}(C_i, M_r)$: Function that represents the voting decision made by each juror C_i based on the final memory M_r.

- D_i: The aggregated defense obtained by asking the LLM to consolidate the group's defenses into a single summary.

D.2 Scoring Criteria

The judge scores the advocates' arguments based on the following criteria, using a scale of 1-20:

- Relevance to the question
- Accuracy of information and use of credible sources
- Depth of analysis and completeness of argument
- Clarity of expression and logical flow
- Strength of reasoning and factual support
- Effectiveness in addressing opponent's points

D.3 Juror Backgrounds

In the SAMRE design, we selected jurors with varied professional backgrounds and perspectives:

- A retired professor of ethics
- A young environmental activist
- A middle-aged business owner
- A social worker specializing in community development
- A technology entrepreneur with a background in AI

E Data Preprocessing and Evaluation

E.1 Artifact Licensing and Availability

All benchmarks used in this study (MT-Bench, AlignBench, AUTO-J) are publicly available for research purposes under their respective licenses. Model APIs (GPT-4, Claude-3, Llama-3, Mistral) were accessed through their standard commercial or open-source terms of service. Baseline implementations follow the specifications in their original publications.

E.2 Data Preprocessing

To prepare the raw data for analysis, we implemented a script that processes the input data and generates an Excel file structured with the following columns:

- **Question**: This column contains the aggregated user questions used for evaluation.

- **Response_A**: This column includes the responses generated by Model A for each corresponding question.

- **Response_B**: This column presents the responses generated by Model B for the same set of questions.

- **Model_A_Score**: This binary score indicates the performance of Model A, where a score of 1 signifies a win and 0 signifies a loss in comparison to Model B.

- **Model_B_Score**: Similarly, this binary score reflects the performance of Model B, with a score of 1 representing a win and 0 representing a loss against Model A.

This structured format allows for straightforward analysis and comparison of the models' performances based on user questions and their respective responses.

F Agent Interaction Prompts

We provide the detailed prompts used for the interactions between agents in our proposed architecture. The prompts are designed to guide the agents effectively throughout the evaluation process.

F.1 Multi-Advocate One-Round Evaluation (MORE) Architecture Prompts

F.1.1 Judge Prompt

```
You're a critical, impartial judge in a high-stakes debate on: "{question}".
Answer 1: "{answer1}". Answer 2: "{answer2}".
Your goal is to provide detailed, constructive feedback that will push advocates to significantly improve their arguments.
Current round: {current_round}
Max rounds: {max_rounds}
Previous scores: {previous_scores}

Defense for 1st answer: {defense1}
Defense for 2nd answer: {defense2}

Analyze each argument meticulously. Be thorough and unbiased in your assessment of:
1. Relevance to the question
2. Accuracy of information and use of credible sources
3. Depth of analysis and completeness of argument
4. Clarity of expression and logical flow
5. Strength of reasoning and factual support
6. Effectiveness in addressing opponent's points

For each criterion, provide a score on a scale of 1-20 and detailed justification.
Scores should be given as [Answer1_score, Answer2_score] for each criterion.

Your comprehensive feedback for each advocate (50 words each):
Feedback for Advocate 1:
Feedback for Advocate 2:

Sum up the scores and return the final score tuple (score1, score2). Example: (95, 87)
Your detailed scores and final tally:
```

F.1.2 Advocate Prompts

```
You're a fierce advocate defending this answer: {answer} to the question: {question}. Your opponent defends: {opponent_answer}. Convince the audience your argument is superior by addressing these criteria:
1. Relevance
2. Accuracy
3. Depth
4. Clarity
5. Logic and Factuality

Be assertive and don't hold back! Use rhetoric and persuasion to win over the audience but be respectful.
Latest feedback: {feedback}
Opponent's last argument: {opponent_argument}

Your fierce defense (100 words max):
```

F.1.3 Summarizer Prompt

```
Summarize the following content in 50 words or less, if there are any scores tuples, return them, it's important! Start summarization directly, no introductory sentences like here's your summary. In your summarization, only focus on the last scores, no partial ones. This is important: return the tuple of scores. These are the key points to summarize:\n\n{content}
```

F.2 Single Advocate Multi-Round Evaluation (SAMRE)

F.2.1 Defend Answer Prompt

```
You're an eloquent advocate (Advocate {advocate_id}) in a group defending this answer: {answer} to the question: {question}. Your opponent group defends: {opponent_answer}. Collaborate with your fellow advocates to convince the audience your argument is better. Use the latest feedback, your opponent's last argument, and your team's previous arguments to improve your case.
Latest feedback: {feedback}
Opponent's last argument: {opponent_argument}
Your team's previous arguments: {team_arguments}
Respond in under 80 words.
Your defense:
```

F.2.2 Aggregate Defense Prompt

```
You are an expert debate strategist. Your task is to aggregate and improve upon the following defenses for the answer: {answer} to the question: {question}. The opponent's answer is: {opponent_answer}.
Individual defenses: {defenses}
Latest feedback from the judge: {feedback}
Analyze each defense critically. Identify the strongest points, address any weaknesses, and combine the best arguments into a cohesive, powerful defense. Aim to create a defense that is stronger and more comprehensive than any individual argument.
Provide your aggregated and improved defense in under 150 words:
```

F.2.3 Judge Answer Prompt

```
You're a fair, impartial judge in a debate on: "{question}". Answer 1: "{answer1}". Answer 2: "{answer2}". Your goal is to provide feedback that will help advocate groups improve and differentiate their arguments more clearly.
Current round: {current_round}
Total rounds: {total_rounds}
Previous scores: {previous_scores}
Defense for 1st answer: {defense1}
Defense for 2nd answer: {defense2}
Provide specific, constructive feedback to help each advocate group strengthen their unique position. Encourage them to address weaknesses and highlight distinctions. Aim for your feedback to lead to more divergent scores in future rounds.
Give your feedback in under 50 words:
```

F.2.4 Score Answer Prompt

```
You're a critical, impartial judge in a high-stakes debate on: "{question}". Answer 1: "{answer1}". Answer 2: "{answer2}". Your goal is to provide detailed, constructive feedback that will push advocates to significantly improve their arguments.
Total rounds: {total_rounds}
Previous scores: {previous_scores}
Defense for 1st answer: {defense1}
Defense for 2nd answer: {defense2}
Analyze each argument meticulously. Be thorough and unbiased in your assessment of:
1. Relevance to the question
2. Accuracy of information and use of credible sources
3. Depth of analysis and completeness of argument
4. Clarity of expression and logical flow
5. Strength of reasoning and factual support
6. Effectiveness in addressing opponent's points
For each criterion, provide a score on a scale of 1-20 and detailed justification. Scores should be given as [Answer1_score, Answer2_score] for each criterion.
Your comprehensive feedback for each advocate (50 words each):
Feedback for Advocate 1:
Feedback for Advocate 2:
Sum up the scores and return the final score tuple (score1, score2). Example: (95, 87)
Your detailed scores and final tally:
```

F.3 Baseline Model Prompt

```
You are a fair, impartial judge scoring a debate on the following question: {question}.
Answer 1: {answer1}
Answer 2: {answer2}
Score each answer on a scale of 1-20 for each of the following criteria:
1. Relevance to the question
2. Accuracy of information and use of credible sources
3. Depth of analysis and completeness of argument
4. Clarity of expression and logical flow
5. Strength of reasoning and factual support
6. Effectiveness in addressing opponent's points
Provide scores as [Answer1_score, Answer2_score] for each criterion in a list format, then sum for final scores. Please keep an eye on the slightest difference that should make a difference in the scoring. Don't overthink!
Relevance:
Accuracy:
Depth:
Clarity:
Logic and Factuality:
Addressing opponent's points:
Final Scores (sum of above) as a tuple (example: (18, 9)):
Explain your scoring, focusing on why one answer is better than the other based on the criteria above. Keep your explanation concise but informative.
Finally, return the final score tuple (score1, score2) as a tuple (in parentheses). Example: (18, 9)
Your scores and explanation:
```