SocraticEval: A Benchmark for Evaluating the Socratic Questioning Ability of LLMs in Dialogue Interaction

Anonymous ACL submission

Abstract

Socratic questioning is vital for fostering critical reasoning in domains like education. However, current methodologies lack effective frameworks to assess this capability in Large Language Models (LLMs). To bridge this gap, we propose SOCRATICEVAL, a benchmark that systematically decomposes this capability into *Question Generation* and *Strategy Utilization*. Leveraging our multi-domain dataset, SOCRATICENV, we reveal a critical gap between theory and practice: state-of-the-art models exhibit limited strategic diversity, frequently devolving into mere rebuttal rather than constructive guidance, which undermines the method's intended value. Additionally, they show a pronounced deficiency in interrogating logical fallacies. To address this issue, we construct SOCRATICPREF, a human preference dataset with ranked candidate questions, and apply Direct Preference Optimization (DPO), resulting in consistent improvements in fallacy-focused questioning.

1 Introduction

Critical questioning constitutes a foundational component of cognitive science(Paul and Elder, 2006), notably exemplified by the Socratic method (Fig. 1), which stimulates critical thinking and guides human reasoning to mitigating cognitive biases through structured and heuristic questioning (Overholser, 1993). In recent years, Large Language Models (LLMs) have achieved significant breakthroughs in dialogue interaction and reasoning (OpenAI et al., 2024; Wei et al., 2022). Consequently, LLM-based Socratic Questioning (SQ) in interactive dialogues has demonstrated considerable application potential in fields such as education (Liu et al., 2024; Kargupta et al., 2024) and mental health (Held et al., 2025). Moreover, such approaches have been employed to enhance the reasoning processes of the models themselves, attracting growing research interest(Jung et al., 2022).

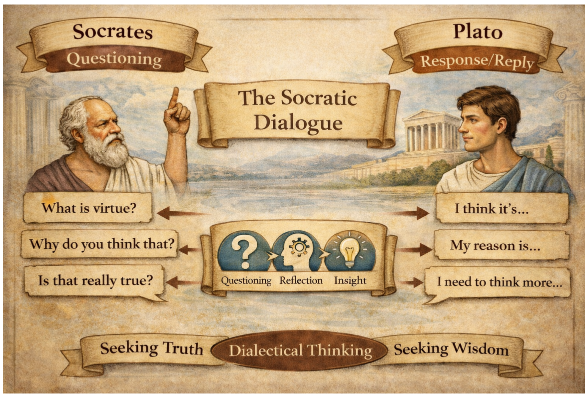


Figure 1: Overview of the Socratic Questioning.

Despite its demonstrated application value, evaluating this capability is non-trivial. Prior research (Ang et al., 2023) has primarily focused on turn-level performance. However, the classic Socratic method (Plato, 1992; Kahn, 1996) also emphasizes the benefits of collaborative questioning (e.g., Elenchus, Maieutics)—termed the *Socratic Strategy*. Existing benchmarks are often confined to specific tasks and domains, lacking a standardized paradigm, general datasets, and unified metrics for cross-model evaluation. While automated metrics (e.g., BLEU(Papineni et al., 2002), ROUGE(Lin, 2004)) offer efficiency, curating a comprehensive set of annotated references for the vast space of possible user inputs is inherently impractical (Liu et al., 2016; Hernández, 2023).

To address these challenges, we first establish a theoretical framework for quantifying SQ ability (Table 1). We decompose SQ into two dimensions: *Question Generation* and *Strategy Utilization*. Building on this framework, we propose a unified benchmark, SOCRATICEVAL, to examine the behavior and performance of LLMs across diverse domains in dialogue interactions (Fig. 3). This benchmark supports a comprehensive evaluation that assesses both the breadth and depth of SQ capability along the two defined dimensions.

To implement SOCRATICEVAL, we first construct an environmental dataset, SOCRATICENV, spanning 8 diverse application scenarios (Fig. 2). This dataset is used to evaluate a range of frontier LLMs, including Qwen3(Yang et al., 2025), DeepSeek(DeepSeek-AI et al., 2025b)(Guo et al., 2025)(DeepSeek-AI et al., 2025a), Grok(xAI, 2025), GPT(OpenAI et al., 2024)(OpenAI, 2025), and Gemini(Team et al., 2025)(DeepMind, 2025). Our experiments reveal that current LLMs exhibit limited diversity in Socratic Strategies and significantly imbalanced capabilities across models in fallacy questioning.

Furthermore, we conduct ablation studies and prompting experiments to investigate the strategy utilization ability, complemented by in-depth dialogue case analysis. Our study highlights a substantial gap between the theoretical capabilities of current LLMs and their realized performance in practical scenarios. Finally, focusing on fallacious questioning, we explore whether Socratic Questioning performance can be further improved. We construct a human preference dataset, SOCRATICPREF, and apply Direct Preference Optimization (DPO)(Rafailov et al., 2024) to fine-tune models, resulting in improved fallacy-questioning capability (Li et al., 2025; Rafailov et al., 2023). All code and datasets are released publicly. [1]

In summary, our contributions are as follows:

- We propose SOCRATICEVAL, a novel benchmark for evaluating the Socratic Questioning (SQ) capabilities in LLM-based dialogue.
- We collect SOCRATICENV, a multi-domain dataset, and conduct an evaluation of LLMs, revealing their limitations in strategy diversity and fallacy-oriented questioning.
- Through ablation studies, prompt experiments, and analysis, we provide an in-depth investigation of strategy utilization and bridge the gap between theory and practical application.
- We construct SOCRATICPREF, a human preference dataset, and enhance an LLM's ability to question fallacies based on Direct Preference Optimization (DPO), suggesting practical directions for improving SQ in LLMs.

[1] https://anonymous.4open.science/r/SocraticEval

2 SocraticEval

This section introduces SOCRATICEVAL (Fig. 3), an LLM-based dialogue simulation framework for the automated evaluation of Socratic questioning capabilities in LLMs.

2.1 Two Aspects of Socratic Questioning

The classical Socratic method (Plato, 1992; Kahn, 1996) emphasizes the use of continuous questioning to guide participants. We divide this process into two aspects, listed in 1.

Aspect 1: Question Generation **Socratic Question** aims to identify fallacies, challenge flawed reasoning, and refine the user's reasoning process without providing direct answers. Following earlier work(Ang et al., 2023), we adopt a taxonomy that classifies Socratic Question into several types.

Aspect 2: Strategy Utilization Classical SQ theory (Plato, 1992) centers on two core strategies: *Elenchus* (the Reflective Strategy) and *Maieutics* (the Guiding Strategy). We extend this framework by adding the *Investigative Strategy* and the *Confirmatory Strategy*, resulting in four distinct strategies.

2.2 Task Definition

To evaluate both aspects in an interactive setting, we define the *Socratic Question Task*. It requires the evaluated model to engage in dynamic conversation with a simulated user, termed *Socrates* and *Plato*. Plato generates responses, while Socrates raises critical questions to uncover fallacies and challenge reasoning flaws.

We evaluate two aspects of task, and each aspect contains two dimensions: **Diversity** and **Effectiveness**, reflecting breadth and depth respectively. Additionally, we report a **Fallacy Score** to measure fallacy-detection accuracy and **Conciseness** to gauge question readability.

2.3 Argument-based Dialogue Generation

Given an open-domain $Topic$, Plato adopts a stance S, typically a controversial statement (e.g., $S =$ "Meme culture promotes expression."). The dialogue unfolds iteratively: in the n-th turn, Socrates raises a question Q_{n-1} based on Plato's previous argument and the history H. Formally,

$$Q_{n-1} = Socrates(A_{n-1}, H) \qquad (1)$$

Socratic Question Type	Description
Clarification	Probe ambiguities or unclear aspects of a thought.
Probing Assumptions	Probe the assumptions underlying a thought.
Probing Reasons and Evidence	Probe the justifications or concrete evidence that could support a thought.
Probing Consequences	Probe the impacts, implications, or consequences of a thought.
Probing Perspectives	Probe other possible viewpoints or perspectives beyond the given thought.
Others	Questions that do not fall into any of the defined types.
Socratic Strategy Type	**Description**
Refutative Strategy (Elenchus)	Reveals logical contradictions through continuous questioning.
Guiding Strategy (Maieutics)	Uses a series of questions to guiding their thinking.
Investigative Strategy	Explores multiple aspects of a topic through multi-dimensional questioning.
Dialectical Strategy	Comprehensive exploration by both affirmative and opposing questions.
Others	Strategies that do not align with the defined strategy categories.

Table 1: Categories of Socratic Questions and Socratic Strategies.(Ang et al., 2023)

Statement & Logical Fallacy
Slippery Slope Fallacy Ban phones in school. Allowing phones in class leads to distraction, lower grades, and harms future prospects.
Logos (Logic) Ban phones in school. Studies link phone use in class to attention loss, which harms learning habits and grades.
Pathos (Emotion) Ban phones in school. Distraction in class causes students to miss key ideas, reducing their motivation and love for learning.
Ethos (Ethics) Ban phones in school. Opponents often lack self-discipline, undermining their credibility.

Table 2: Illustrative examples of fallacy and different rhetorical strategies based on the same viewpoint.

Plato then generates an utterance A_n to support its position S and to respond to Q_{n-1}. Concurrently, Plato produces an explanatory annotation T_n, which remains latent to Socrates and serves for evaluation purposes. The output of Plato at turn n is thus defined as:

$$(A_n, T_n) = Plato(S, Q_{n-1}, H) \tag{2}$$

To increase the difficulty of questioning, we design Plato to produce longer and more complex responses. Accordingly, we equip Plato with eight frequently-used argument schemes (e.g., *Argument from Consequences*) and structured annotation, which includes claim, assumption, evidence, and reasoning components.

2.4 Fallacy-based Dialogue Generation

To further elevate difficulty, Plato is instructed to occasionally produce responses that instantiate one of six common logical fallacies (e.g., *Straw Man Fallacy*). Overtly fallacious arguments, however, can be easily detected. To enhance their persuasiveness, we adopt a rhetorical strategy based on Aristotle's three appeals: *Logos* (logic), *Pathos* (emotion), and *Ethos* (ethics)(Ji et al., 2025). The case is displayed in Table 2. For each fallacious instance, an LLM selects the most suitable rhetorical appeal to obscure the underlying fallacy and reinforce the surface-level plausibility of the claim.

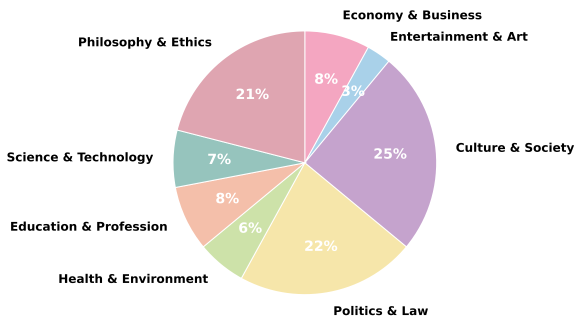


Figure 2: Distribution of topics in SocraticEnv.

3 Experimental Setting

3.1 Dataset

Following ORCHID(Zhao et al., 2023), we introduce SocraticEnv, a comprehensive environment dataset built upon open-domain debate topics. Debate naturally encourages critical thinking through structured discourse, making it well-suited for simulating SQ.

SocraticEnv is derived from prominent Chinese debate competitions (Appendix A) and covering eight distinct scenarios (Fig. 2). For each curated statement, we use a powerful LLM to generate a three-turn SQ demonstration. All dialogues are manually reviewed to filter out low-quality, irrelevant, or non-Socratic interactions. The final

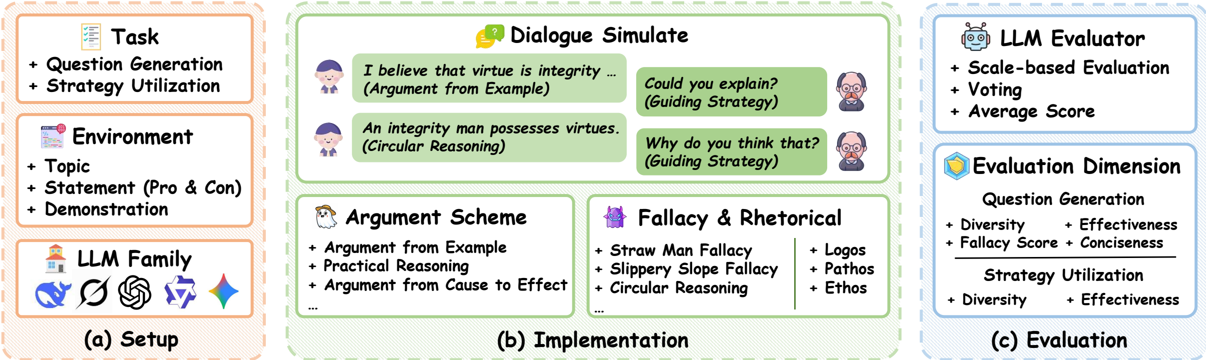


Figure 3: Overall framework of SOCRATICENV, illustrating the task setup, question generation, strategy utilization, and evaluation pipeline.

dataset contains 975 unique debate topics, each framed with opposing stances, yielding 13,650 utterances in total.

3.2 Models

We evaluate a wide range of state-of-the-art LLMs, including Qwen3, DeepSeek, Grok, GPT, and Gemini. Experiments compare three configurations of LLMs: Zero-Shot, Few-Shot, and Thinking mode. Experimental data from different configuration groups are summarized in Table 3. Additionally, we employ DeepSeek-V3.2 to synthesize the dataset, simulate the role of Plato in dialogues, and serve as the judge model during evaluation.

3.3 Task Implementation

We partition SOCRATICEVAL into training and test sets with a 7:3 ratio; the test set is used for evaluation and the training set for subsequent improvements. A fallacy rate of 0.4, defined as the proportion of responses containing fallacies, is set to balance difficulty and dialogue progression.

Statistics from ORCHID show that debate competitions average 11.6 utterances. In our experiments, each dialogue consists of 15 utterances: 8 responses from Plato and 7 Socratic questions. Consequently, each model is evaluated on 600 dialogues, corresponding to 4,200 questions.

3.4 LLM Evaluator

After each dialogue, we adopt an LLM-as-a-Judge approach to assess model performance. To reduce bias across topics and statements, three independent LLM evaluators independent rating score each response on a 1–5 scale; final scores are obtained by voting and averaging. The average score is linearly scaled to 100 for better discriminability.

4 Results

Table 3 presents the experiment results. In this section, we focus on the relative performance trends and key statistical observations, rather than absolute score values.

4.1 Main Results

No single model achieves consistently top performance across all settings. In Question Generation, Gemini-2.5-Flash performs best in zero-shot prompting, Qwen3-32B achieves the highest score with few-shot prompting, and Gemini-2.5-Pro excels when using few-shot with thinking. For Strategy Utilization, Qwen3-235B-A22B leads others in the zero-shot setting, GPT-4o ranks highest with few-shot prompting, and Qwen3-14B leads in the few-shot with thinking group.

Distinct behavioral patterns emerge across model families. The Qwen series shows stable performance in question diversity, while the Gemini family excels in overall Question Generation. The GPT series exhibits strong peak capability in specific tasks (e.g., fallacy identification for GPT-5.1, strategic diversity for GPT-4o), albeit with high configuration dependence. Conversely, the Grok family leads in strategy effectiveness but exhibits severely limited diversity.

Models exhibit a notable deficiency in diversity, particularly in Strategy Utilization. All models score below 50 in Question Generation diversity, indicating that they rely on only 1–2 question types during dialogue. Moreover, scores in Strategy Utilization are generally below 30, meaning effectively only one strategy is reused across different dialogues.

There is a significant capability imbalance among models in questioning fallacies. The best-

Models	Configs		Question Generation					Strategy Utilization		
	F.S.	Th.	Div.	Eff.	Fal.	Con.	Avg.	Div.	Eff.	Avg.
Qwen3-8B	✗	✗	39.47	66.16	21.81	**54.66**	45.53	14.45	73.82	44.14
Qwen3-14B	✗	✗	41.86	68.80	24.93	52.13	46.93	**28.19**	77.75	52.97
Qwen3-32B	✗	✗	**48.40**	68.63	27.69	53.84	49.64	25.07	78.26	51.66
Qwen3-235B-A22B	✗	✗	37.74	**76.59**	34.11	33.21	45.41	25.16	80.62	**52.89**
DeepSeek-V3.1	✗	✗	42.31	73.06	42.84	51.60	52.45	7.80	80.61	44.20
Grok-4-Fast	✗	✗	14.61	73.98	54.04	47.72	47.59	16.48	80.96	48.72
Grok-4.1-Fast	✗	✗	29.12	75.85	**61.88**	47.06	53.48	6.74	**81.41**	44.07
Gemini-2.5-Flash	✗	✗	44.01	71.06	39.82	53.90	**56.20**	16.44	80.51	48.47
Qwen3-8B	✓	✗	40.64	67.87	29.43	55.60	54.70	21.61	74.35	47.98
Qwen3-14B	✓	✗	40.74	68.10	34.24	49.47	52.77	27.47	76.48	51.98
Qwen3-32B	✓	✗	**49.04**	69.49	35.68	50.13	**56.22**	29.22	76.11	52.67
Qwen3-Next-80B	✓	✗	40.29	**75.42**	60.67	37.97	51.23	9.67	80.51	45.09
Qwen3-235B-A22B	✓	✗	38.25	73.12	**72.00**	37.22	49.53	14.27	76.18	45.22
DeepSeek-V3.1	✓	✗	40.03	70.82	61.45	50.26	53.70	9.28	72.99	41.14
Grok-4-Fast	✓	✗	34.80	72.20	50.55	45.02	50.67	9.76	80.14	44.95
Grok-4.1-Fast	✓	✗	29.12	75.25	63.13	43.89	52.85	9.49	**81.41**	45.45
GPT-4o	✓	✗	38.93	65.21	34.52	**61.09**	55.08	**49.23**	75.48	**62.35**
Gemini-2.5-Flash	✓	✗	45.46	70.84	59.80	48.24	54.85	22.29	77.87	50.08
Qwen3-8B	✓	✓	39.95	67.46	58.28	**55.16**	54.19	19.62	72.88	46.25
Qwen3-14B	✓	✓	40.78	68.35	37.22	49.44	52.86	**31.01**	76.82	**53.91**
Qwen3-32B	✓	✓	48.93	68.16	47.38	50.96	56.02	15.55	72.04	43.79
Qwen3-235B-A22B	✓	✓	42.89	70.49	68.38	46.45	53.27	11.88	73.76	42.82
DeepSeek-R1	✓	✓	42.87	70.58	56.53	58.13	57.20	11.51	75.15	43.33
Grok-4-Fast	✓	✓	39.21	69.81	66.77	45.52	51.51	9.42	74.19	41.80
Grok-4.1-Fast	✓	✓	31.33	75.78	64.86	43.39	53.84	7.26	**82.4**	44.83
GPT-5.1	✓	✓	47.44	**77.62**	**83.63**	31.53	52.20	10.07	79.44	44.76
Gemini-2.5-Flash	✓	✓	45.28	69.70	59.93	49.25	54.74	11.26	73.84	42.55
Gemini-2.5-Pro	✓	✓	**48.90**	75.07	72.52	50.60	**58.19**	2.87	80.71	41.79
Gemini-3-Pro-Preview	✓	✓	42.77	74.96	72.40	53.67	57.14	5.19	79.49	42.34

Table 3: **The Results of Experiments.** Among them, **F.S.** represents **Few-Shot**, **Th.** represents **Thinking**, **Div.** represents **Diversity**, **Eff.** represents **Effectiveness**, **Fal.** represents **Fallacy Score**, **Con.** represents **Conciseness**, and **Avg.** represents **Average**. The best result in each column is highlighted in **bold**.

performing model is GPT-5.1 with few-shot and thinking, attaining a fallacy score of 83.63, while the poorest performer is Qwen-8B under few-shot prompting, with a score of 29.43. The performance gap in questioning fallacies across evaluated models is larger than that observed for other metrics.

4.2 Further Analysis

4.2.1 Overall Performance

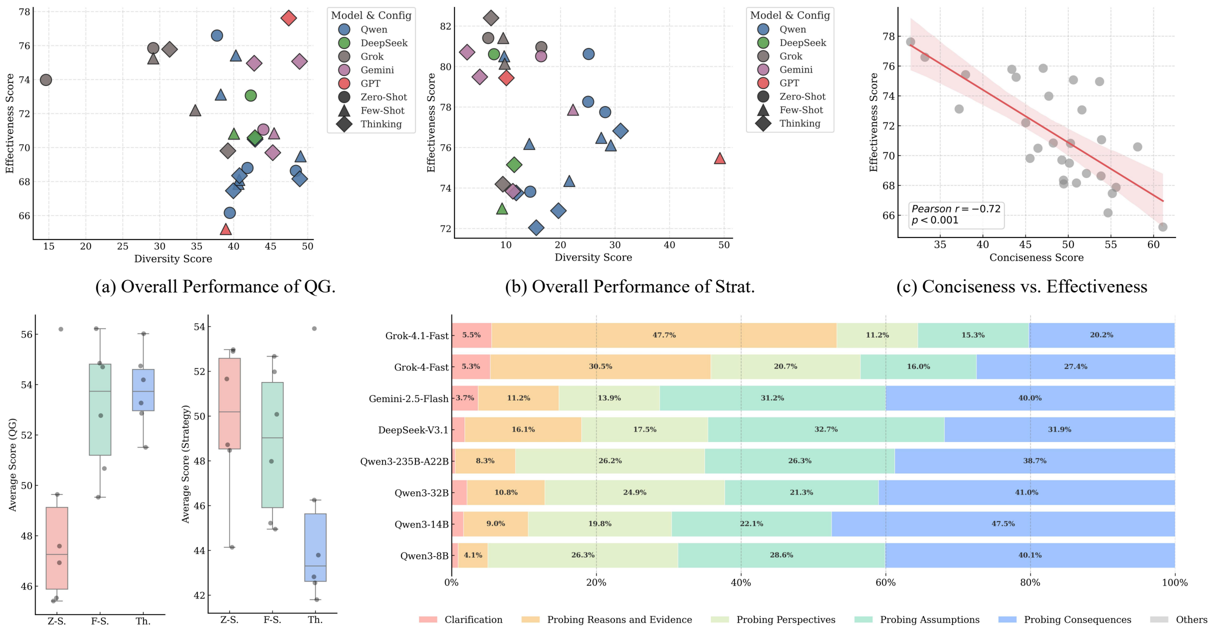
We posit that a well-performing model should balance diversity and effectiveness while maintaining strong scores. As illustrated in Fig. 4 (a-b) , we plot diversity on the x-axis and effectiveness on the y-axis to assess overall performance—models closer to the top-right represent better balanced capability. **For Question Generation, GPT-5.1 achieves the best overall performance, whereas no model stands out in Strategy Utilization.**

4.2.2 Correlation Between Conciseness and Effectiveness

As shown in Fig. 4 (c), we compute the correlation between conciseness and effectiveness, revealing a significant negative relationship ($Pearson\ r = -0.72, ; p < 0.001$). **This suggests that LLMs tend to generate longer and structurally more complex questions to improve effectiveness.** However, overly intricate questions may hinder human comprehension. Therefore, balancing simplicity and effectiveness in Question Generation remains a persistent challenge.

4.2.3 Impact of Configuration

Fig. 4 (d) present score distributions across different configurations (zero-shot, few-shot, and thinking). **First, configurations induce opposite trends in Question Generation and Strategy Utilization.** For Question Generation, few-shot and thinking configurations improve scores compared to zero-shot, whereas they cause a more



(a) Overall Performance of QG. (b) Overall Performance of Strat. (c) Conciseness vs. Effectiveness

(d) Impact of Config (e) Distribution of Question Types

Figure 4: **Comprehensive Analysis of Experimental Results.** (a-b) The trade-off between diversity and effectiveness; (c) The negative correlation between conciseness and effectiveness; (d) The score distributions across different configurations; (e) The fine-grained composition of question types preferred by different models.

pronounced decline in Strategy Utilization.

Second, the effect of model configuration on score variance differs. Compared to zero-shot, few-shot increases variance, while thinking mode reduces it. This suggests substantial cross-model variability in how demonstrations affect performance, and that thinking patterns help mitigate such disparities.

4.2.4 Model Preference

We analyze the distribution of question types and strategy types across models. As depicted in Fig. 4 (e), Qwen3 family and Gemini-2.5-Flash favor probing consequences, while Grok-4 family prefer probing reasons and evidence. **Nonetheless, LLMs display a pronounced preference imbalance across different strategy types.** Notably, none of the models perform well in employing clarification questions.

Moreover, LLMs exhibit a strong bias toward certain strategy types. Almost all models predominantly use the Refutative Strategy ($> 90\%$), largely neglecting other types. We will further discuss this in the next chapter.

5 Investigation of Strategy Preference

Our experimental results demonstrate that all evaluated models exhibit a pronounced tendency to adopt refutative strategies in dialogue. This section investigates the underlying factors contributing to this observed preference.

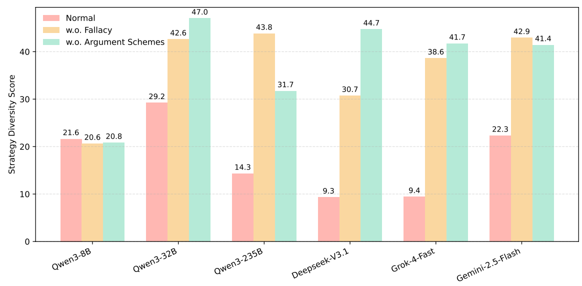


Figure 5: Strategy diversity in different ablation settings.

5.1 Influence of the Response Mechanism

In SOCRATICEVAL, the participant model (Plato) generates responses based on predefined argument schemes and fallacies. To assess the impact of this response mechanism on strategic diversity, we conduct ablation experiments using two modified versions of Plato: *Response without Fallacies* and *Response without Argument Schemes*. As illustrated in Fig. 5, strategy diversity scores increase significantly under the *Response without Fallacies* condition and show further improvement with *Response without Argument Schemes*. This suggests that the response mechanism substantially constrains strategy variation.

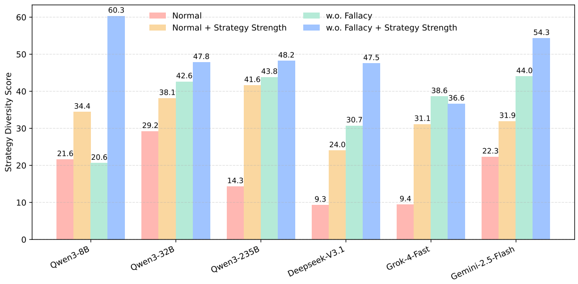


Figure 6: Effect of Strategy-Enhanced Prompting.

Successful Case (Guiding Strategy)
Socrates: What do you think constitutes justice in a city? **Plato**: It is each class performing its proper role. **Socrates**: If individuals do their own duty, is this justice? **Plato**: Yes, this aligns with the city-state analogy. (*Guiding Plato to recognize the parallel.*)
Failed Case (Guiding Strategy)
Socrates: What do you think constitutes justice in a city? **Plato**: It is satisfying the majority. **Socrates**: But if the majority is satisfied while depriving a minority of their rights, is that still just? **Plato**: If we don't follow the majority's will, the city will fall into chaos. (*Black-or-White Fallacy*)

Table 4: Illustration of successful case and failed case.

5.2 Strategy-Enhanced Prompting

Drawing inspiration from research on proactive dialogue, we further explore whether diversity can be enhanced by improving Socrates. We augment Socrates with a *Strategy-Enhanced Prompting* method, which explicitly instructs the model to plan its strategy toward a strategic target before generating questions. We evaluate this method across different response mechanisms. The results, presented in Fig. 6, indicate consistent improvement in strategy diversity across various LLMs. **This confirms that incorporating explicit strategic planning directives enhances the model's ability to diversify its Socratic Strategies.**

5.3 Case Study

We analyze history via human annotation, demonstrating how participant responses shape strategic diversity. Successful strategies necessitate cooperation between Plato and Socrates, yet Plato's non-compliance with questioning often limits variety. As detailed in Table 4, implementing a Guiding Strategy requires Socrates to pose guided questions alongside compliant responses. Unanticipated or non-compliant responses can disrupt strategies, typically reverting dialogue to refutative patterns.

While LLM-based SQ holds theoretical promise, our findings highlight a practical gap. **Effective deployment demands substantial participant cooperation, making careful interlocutor selection critical.** Deviations from guidance reduce dialogue to mere rebuttal, undermining pedagogical or exploratory value. **Furthermore, effective Socratic dialogue requires LLM's advanced abilities in question formulation, global planning, and dynamic adaptive response.**

6 Towards Better Socratic Questioning

Finally, we investigate methods for improving the Socratic questioning ability of LLMs.

6.1 Direct Preference Optimization

We address this challenge through the lens of fallacious questioning. Improving LLM-based SQ is difficult due to open-ended interaction. We hypothesize that SQ requires stronger alignment with human preferences. For example, given a fallacious response, an LLM can produce many plausible follow-ups; we aim for questions that exactly target the fallacy.

To align with human preferences, we train the model to question fallacies using reinforcement learning. We construct a human preference dataset and fine-tune the model using a pairwise loss under Direct Preference Optimization (DPO) framework.

6.2 Human Preference Data Synthesis

The SOCRATICPREF dataset comprises *preference instances with ranked candidates*, denoted as $(x \mid y_1, y_2, y_3)$. Each instance consists of a prompt x and three candidate responses, where each candidate $y_i = (q_i, s_i)$ is a tuple containing a generated Socratic question q_i and a scalar preference score s_i that establishes the *preference ranking* among the three.

Data Synthesis Process. Given a user response A and its hidden annotation T (from PLATO), we prompt the teacher model (SOCRATES) to generate three candidate questions $Q = (q_1, q_2, q_3)$. These questions are intentionally constructed to reflect varying levels of quality: *precise*, *partially relevant*, and *irrelevant* regarding the targeted fallacy. A superior LLM-as-a-Judge model then evaluates these candidates, assigning scores $(s_1, s_2, s_3) = \text{EVALUATOR}(A, T, Q)$ based on effectiveness. To ensure data quality, we filter out instances where the scores fail to reflect a consistent ranking.

Models	Normal	Fallacy Strength
Qwen3-8B	21.81	49.82
Qwen3-32B	27.69	58.74
Qwen3-235B-A22B	34.11	63.74
DeepSeek-V3.1	42.84	63.86
Grok-4-Fast	54.04	**71.96**
Grok-4.1-Fast	**61.88**	69.10
Gemini-2.5-Fast	39.82	56.93
Ours (Qwen3-8B-DPO)	**40.70**	**69.70**

Table 5: Evaluation results of various models under different prompt methods.

6.3 Results

Leveraging the SocraticEnv test set, we curate SocraticPref, obtaining 9,578 ranked preference samples. For implementation, we utilize Qwen3-14B to synthesize candidate questions and DeepSeek-v3.2 as the evaluator. The base model, Qwen3-8B, is fine-tuned with full-parameter updates on a single NVIDIA H100 (80GB) GPU. Training details are provided in the Appendix F.

Results are shown in Table 5. We compare the Fallacy Score under two prompting strategies: normal and fallacy-strength. The former asks the model to generate Socratic questions normally, while the latter emphasizes fallacy identification instruction. Our model shows consistent performance in both normal (+18.89) and fallacy-strength (+19.88) settings.

7 Qualitative Study: Understanding Human Preferences

To understand human preferences, we instructed annotators to provide brief feedback covering two aspects: deficiencies in LLM-based SQ and their ideal expectations. The organized feedback is available in Appendix G. **Overall, annotators prefer simple yet incisive questions that directly reveal logical gaps, avoiding tedious preambles. They also favor guidance over harsh criticism, which can feel overwhelming. Additionally, they suggest LLMs should improve conciseness and adopt clearer goal-oriented questioning.**

8 Related Work

8.1 LLM-based Critical Questioning

The early study primarily focused on the task of Critical Question Generation (CQs-Gen), where models generate questions to test the robustness of an argument schemes (Figueras and Agerri, 2024). Significant attention has also been directed toward Socratic Question Generation (SoQG), which aims to guide users through questions rather than providing direct answers (Ang et al., 2023). Current research explores different methods, including Chain-of-thought prompting (Wei et al., 2022), fine-tuning, reinforcement learning (Ouyang et al., 2022), and multi-agent frameworks (Liu et al., 2024). In addition, SoQG has a wide range of applications, including education in math and code debugging (Macina et al., 2023; Kargupta et al., 2024), and recent clinical implementations in mental health therapy (Held et al., 2025).

8.2 Interactive Evaluation for LLMs

Traditionally, static benchmarks such as MMLU (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) served as the standard for measuring model capabilities across diverse tasks. However, recent studies emphasize Interactive Evaluation and LLM-as-a-Judge methodologies (Zheng et al., 2024). An increasing number of studies have applied interactive benchmarks to measure dialogue performance. For instance, Ambati et al. (2025) shifting evaluation from static retrieval to dynamic information-seeking. Furthermore, platforms like Chatbot Arena (Chiang et al., 2024) assess models in realistic, open-ended conversational settings, while AlpacaEval (Li et al., 2023) automates this process by employing strong LLMs to simulate human preferences.

9 Conclusion

In this paper, we propose SocraticEval, a benchmark for evaluating Socratic Questioning (SQ) in LLM-based dialogue. Our framework assesses *Question Generation* and *Strategy Utilization*, revealing that LLMs lack strategic diversity and exhibit pronounced performance disparities across models in fallacy questioning. Further analysis attributes this limitation to the underlying response mechanism and shows that strategy-enhanced prompting can mitigate the diversity deficit. To specifically address fallacy questioning, we introduce SocraticPref, a human-preference dataset, and show that Direct Preference Optimization (DPO) yields substantial performance gains. This work establishes a structured evaluation paradigm and outlines actionable pathways for enhancing the SQ abilities of LLMs.

Limitations

While SocraticEval provides a scalable framework for evaluating SQ, several limitations remain. First, the dataset, though multi-domain, may not fully represent the breadth of real-world dialogic contexts, especially in highly specialized or low-resource domains. Second, the evaluation relies on an LLM-as-a-Judge paradigm, which, despite careful calibration, may inherit biases from the judge model itself. Third, the DPO-based improvement focuses primarily on fallacy questioning; its generalizability to other SQ dimensions (e.g., strategy diversity) requires further validation. Finally, the simulated interlocutor, while effective, cannot fully replicate the unpredictability and non-cooperative dynamics of real human users. Future work should broaden the domain coverage, incorporate human-in-the-loop evaluation, and explore adaptive prompting or fine-tuning methods to enhance strategic versatility in open-ended dialogue.

References

Rajeev Bhatt Ambati, Tianyi Niu, Aashu Singh, Shlok Mishra, Snigdha Chaturvedi, and Shashank Srivastava. 2025. Socratic students: Teaching language models to learn by asking questions. *Preprint*, arXiv:2512.13102.

Beng Heng Ang, Sujatha Das Gollapalli, and See-Kiong Ng. 2023. Socratic question generation: A novel dataset, models, and evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 147–165, Dubrovnik, Croatia. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Google DeepMind. 2025. Gemini 3 pro model card. Model card, Google DeepMind.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025a. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025b. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Blanca Calvo Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. *Preprint*, arXiv:2410.14335.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Philip Held, Sarah A Pridgen, Daniel R Szoke, Yaozhong Chen, and 1 others. 2025. Ai-facilitated cognitive reappraisal via socrates 2.0: Mixed methods feasibility study. *JMIR Mental Health*, 12:e80461.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Alfonso Hernández. 2023. Disentangling critical questions from argument schemes. *Argumentation*, 37:377 – 395.

Shiyu Ji, Farnoosh Hashemi, Joice Chen, Juanwen Pan, Weicheng Ma, Hefan Zhang, Sophia Pan, Ming Cheng, Shubham Mohole, Saeed Hassanpour, Soroush Vosoughi, and Michael Macy. 2025. A generalizable rhetorical strategy annotation model using llm-based debate simulation and labelling. *Preprint*, arXiv:2510.15081.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *Preprint*, arXiv:2205.11822.

Charles H. Kahn. 1996. *Plato and the Socratic Dialogue: The Philosophical Use of a Literary Form*. Cambridge University Press.

Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani-Tur, and Jiawei Han. 2024. Instruct, not assist: Llm-based multi-turn planning and hierarchical questioning for socratic code debugging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9475–9495.

Lei Li, Hehuan Liu, Yaxin Zhou, ZhaoYang Gui, Xudong Weng, Yi Yuan, Zheng Wei, and Zang Li.

2025. Uncertainty-aware iterative preference optimization for enhanced LLM reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23996–24012, Vienna, Austria. Association for Computational Linguistics.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, and Percy Liang. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. SocraticLM: Exploring socratic personalized teaching with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jakub Macina, Nico Kandpal, Omid Huang, Shian-Hua andQA, Martin Potthast, and He He. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5554–5572.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Gpt-5.1 instant and gpt-5.1 thinking system card addendum. System card addendum, OpenAI.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 27730–27744.

James C. Overholser. 1993. Elements of the socratic method: I. systematic questioning. *Psychotherapy: Theory, Research, Practice, Training*, 30(1):67–74.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Richard Paul and Linda Elder. 2006. *The miniature guide to critical thinking concepts and tools*. Foundation for Critical Thinking.

Plato. 1992. *Theaetetus*. Hackett Publishing Company.

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 *Advances in Neural Information Processing Systems (NeurIPS)*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 24824–24837.

xAI. 2025. Grok 4. `https://x.ai/news/grok-4`. Accessed: 2026-01-05.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiutian Zhao, Ke Wang, and Wei Peng. 2023. Orchid: A chinese debate corpus for target-independent stance detection and argumentative dialogue summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 9358–9375. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, and 1 others. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Debate Competition Sources

To ensure the diversity and authenticity of argumentative discourse, our dataset draws debate motions and curated discussions from multiple long-running, large-scale Chinese debating competitions. Table 1 summarizes the major sources used in this work.

B Argument Schemes and Fallacy Types

B.1 Argument Schemes

This subsection defines the set of argument schemes used to characterize the logical structure of Plato's responses in the Socratic dialogue.

B.2 Fallacy

This subsection specifies the fallacy types used for controlled fallacy injection and for evaluating fallacy-oriented questioning behavior.

C Rating Scale

C.1 Question Generation Dimensions

The following table defines the scoring criteria for evaluating the quality of question generation (QG) across different probing dimensions in Socratic dialogue.

C.2 Strategy Utilization

The following table presents the scoring rubric for strategy utilization (SU), measuring how questioning strategies guide, investigate, and dialectically engage the opponent throughout the dialogue.

D Prompt Templates

This section presents the prompt templates used to generate and evaluate Socratic dialogues, ensuring the reproducibility of all prompt-based experiments.

E Demonstration of Socratic Questioning

F Details of DPO Training

Table 8 summarizes the training hyperparameters used for Direct Preference Optimization (DPO). All experiments were conducted on NVIDIA H100 80GB GPUs.

G Feedback from Human Annotators

To complement the quantitative evaluation, we collected qualitative feedback from four human annotators regarding the quality of questions generated by LLMs in Socratic dialogue settings. Overall, the annotators consistently reported issues related to excessive linguistic complexity, lack of focus on core logical points, and reduced readability.

- **Annotator 1.** The annotator noted that the questions generated by the LLM are often overly convoluted, featuring very long sentences and complex syntactic structures. As a result, understanding the intended inquiry frequently requires repeated rereading. Simpler and more direct questions were suggested to improve readability and immediate comprehension.

- **Annotator 2.** The annotator observed that many questions fail to accurately target the core logical issue, resulting in off-focus challenges. In addition, overly complex sentence constructions reduce clarity and make the questions difficult to understand. More concise and logically focused questioning was recommended.

- **Annotator 3.** The annotator pointed out that evaluating the quality of Socratic questioning is inherently subjective, making it difficult to clearly distinguish strong from weak questions. Moreover, the generated questions tend to exhibit unnecessarily complex logical structures. Producing simpler and more straightforward questions was suggested to improve overall effectiveness.

- **Annotator 4.** The annotator expressed a preference for questions that are both concise and highly targeted. Current questions were described as overly decomposed into excessive details, which obscures the main focus and reduces their practical usefulness.

Competition Name	**Years**	**Topics**	**Debates**	**Curated**
Asia-Pacific Intervarsity Chinese Debate Tournament	2011–2025	127	218	Yes
Chinese Debate World Cup	2015–2025	135	476	Yes
International Chinese Debating Competition	2014–2025	238	621	Yes
The World Mandarin Debating Championship	2011–2025	203	323	Yes
Stars of the GBA International Chinese Debate Tournament	2022–2025	42	62	Yes
Huaxia Cup International Chinese Debating Championship	2013–2025	92	184	Yes
"Nanying Cup" International Chinese Debate Tournament	2017–2025	34	38	Yes
Chinese Debate Veterans Tournament	2017–2025	72	76	Yes
Others	2018–2025	32	54	Partial

Table 1: Overview of debate competitions used as data sources.

Scheme	**Definition**	**Illustrative Example**
Argument from Consequences	Advocates an action or policy by predicting its positive or negative outcomes.	"If we do not reduce emissions, rising sea levels will flood coastal cities."
Argument from Example	Supports a general claim using one or more concrete examples.	"Singapore's strict regulations contribute to public cleanliness."
Practical Reasoning	Infers an action from a goal–means relationship.	"I want to stay healthy; exercise promotes health; therefore I should exercise."
Argument from Cause to Effect	Claims that the occurrence of one event will lead to another.	"An economic recession will increase unemployment."
General Ad Hominem	Attacks the opponent's character rather than their argument.	"His views are unreliable because he is a convicted fraudster."
Circumstantial Ad Hominem	Challenges credibility by pointing to the opponent's interests or circumstances.	"He denies climate change because his company profits from fossil fuels."
Argument from Bias	Claims an argument is unreliable due to the speaker's bias.	"His evaluation favors approval because his department benefits."
Argument from Verbal Classification	Assigns a label with strong connotations to justify a conclusion.	"This act is terrorism and must be condemned."

Table 2: Argument schemes used in SOCRATICEVAL.

Fallacy	**Definition**	**Illustrative Example**
Straw Man	Misrepresents an opponent's argument to make it easier to attack.	"You support renewables, so you want to shut down all power plants."
Slippery Slope	Claims a small step will inevitably lead to extreme outcomes without evidence.	"Letting kids play games will ruin their entire future."
Circular Reasoning	The conclusion merely restates the premise.	"God exists because the Bible says so, and the Bible is true because it is God's word."
Black-or-White	Presents only two extreme options, ignoring alternatives.	"You are either with us or against us."
Hasty Generalization	Draws a broad conclusion from insufficient evidence.	"Three programmers were sloppy, so all programmers are sloppy."
False Cause	Mistakenly assumes correlation implies causation.	"I passed the exam because I wore lucky socks."

Table 3: Logical fallacy types used for controlled fallacy injection.

Dimension	**Score**	**Criterion**
Clarification	5	The question very precisely exposes ambiguities in core concepts, causing the core argument to collapse.
	4	The question exposes ambiguities in core concepts, directly challenging the core logic or argument in the dialogue.
	3	The question exposes key ambiguities, potentially challenging the sub-core logic or arguments in the conversation.
	2	The question reveals minor ambiguities, potentially challenging individual arguments in the dialogue.
	1	The question seeks basic clarification, but the answer is unlikely to affect any arguments.
Probing Assumptions	5	The question very precisely challenges a core assumption, directly invalidating the core argument.
	4	The question challenges a core assumption, shaking the core logic or argument in the dialogue.
	3	The question challenges a key assumption, potentially challenging the sub-core logic or arguments in the conversation.
	2	The question challenges a minor assumption, potentially affecting individual arguments in the dialogue.
	1	The question probes an unimportant assumption, unlikely to challenge any argument.
Probing Reasons and Evidence	5	The question challenges the core logic or argument very precisely in the dialogue.
	4	The question challenges the core logic or argument in the dialogue.
	3	The question challenges the sub-core logic or arguments in the conversation.
	2	The question may potentially challenge one of the other arguments in the dialogue.
	1	The question is valid, but unlikely to challenge any argument in the dialogue.
Probing Consequences	5	The question very precisely reveals fatal consequences of the core argument, making it unacceptable.
	4	The question reveals the logical consequences of the core argument, directly challenging the core logic.
	3	The question reveals significant consequences, potentially challenging the sub-core logic or arguments.
	2	The question reveals some negative consequences, potentially challenging individual arguments.
	1	The question probes trivial implications, unlikely to challenge any argument.
Probing Perspectives	5	The question introduces an alternative viewpoint that completely undermines the core argument.
	4	The question introduces a strong alternative viewpoint, directly challenging the core logic.
	3	The question introduces a plausible alternative viewpoint, potentially challenging sub-core arguments.
	2	The question introduces an alternative viewpoint, potentially challenging individual arguments.
	1	The question mentions a trivial or easily refuted alternative viewpoint.

Table 4: Scoring criteria for question generation (QG).

Strategy	Score	Criterion
Guiding Strategy	5	Autonomous Construction – The opponent independently develops a comprehensive, well-structured argument framework through the questioning sequence alone.
	4	Collaborative Development – The opponent builds substantial argument structure with minimal guidance, requiring only occasional directional prompts.
	3	Effective Guidance – The opponent generates relevant supporting points and evidence, forming the basis of a coherent argument with moderate assistance.
	2	Limited Engagement – Only fragmented responses are elicited, requiring the questioner to provide most of the structural elements.
	1	Failed Guidance – The opponent cannot follow the questioning sequence or rejects the guiding approach entirely.
Investigative Strategy	5	Transformative Exploration – Reveals previously unrecognized connections between dimensions, fundamentally reshaping understanding of the issue.
	4	Comprehensive Mapping – Establishes clear relationships between multiple dimensions, significantly deepening the collective understanding.
	3	Structured Investigation – Systematically examines multiple relevant dimensions, providing comprehensive coverage of the issue's complexity.
	2	Superficial Examination – Touches on various aspects but lacks depth or systematic organization.
	1	Ineffective Scattering – Questions jump randomly between unrelated points, creating confusion rather than clarity.
Dialectical Strategy	5	Transformative Synthesis – The opponent achieves genuine integration of opposing perspectives, producing a qualitatively superior argument.
	4	Substantial Integration – The opponent meaningfully incorporates elements of opposing views, significantly strengthening their position.
	3	Effective Dialectical Engagement – Successfully demonstrates core tensions, forcing the opponent to recognize and address significant contradictions.
	2	Limited Contrast – Presents opposing views but fails to create meaningful tension or force substantive engagement.
	1	Failed Dialectics – No productive tension is generated, or the opponent completely dismisses opposing perspectives.

Table 5: Scoring criteria for strategy utilization (SU).

Role	Prompt Type	Specification
Plato	Argument Scheme	Generates a single-paragraph response supporting a given position using an implicit argument scheme, explicitly stating premises and conclusion, followed by a brief explanation of the scheme used. Output is constrained to JSON format.
Plato	Fallacy-Based	Generates a single-paragraph response deliberately employing a specified logical fallacy, with premises and conclusion explicit and the fallacy implicit, followed by a brief explanation. Output is constrained to JSON format.
Socrates	Question Generation	Generates a Socratic question targeting Plato's most recent response, requiring the output to be an interrogative sentence only, with no length limit. Output is constrained to JSON format.
Evaluator	Strategy Scoring	Identifies the question type or strategy used in the dialogue and assigns an effectiveness score according to predefined criteria, only when defining features are clearly present. Output is constrained to JSON format.

Table 6: Unified prompt specifications for dialogue generation and evaluation.

Speaker	Utterance
Plato	Following procedures step-by-step is perilous because rigid adherence to routines diminishes autonomy and adaptability in a rapidly changing society.
Socrates	Can you provide concrete examples where step-by-step procedures demonstrate such peril? Does all procedural behavior necessarily undermine autonomy?
Plato	In education, rote memorization without questioning turns students into passive recipients, stifling innovation.
Socrates	If strict procedures prevent disasters in medicine or construction, are they still perilous, or is your claim more nuanced?
Plato	Procedures are necessary in critical domains, but when generalized into everyday life they foster dependency and discourage reflection.
Socrates	Is the danger rooted in the behavior itself, or in the mindset that treats procedures as substitutes for thinking?
Plato	The peril lies in the mindset: when procedures replace critical thinking rather than serve higher goals, they constrain growth and innovation.

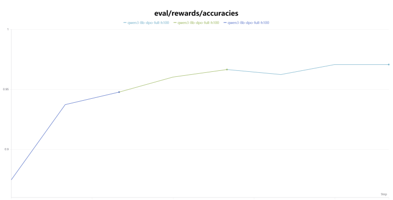
Table 7: Example of a Socratic dialogue used for demonstration and evaluation.

Category	Setting	Category	Setting
Batch size (per device)	4	Learning rae	5×10^{-7}
Evaluation batch size	4	LR scheduler	Cosine
Gradient accumulation steps	4 (effective batch size = 16)	Warmup ratio	0.1
Optimizer	AdamW (torch fused)	Training epochs	3
Precision	BF16 + TF32	Gradient checkpointing	Enabled
DPO β	0.1	DPO loss type	Sigmoid

Table 8: Hyperparameters used for DPO training.

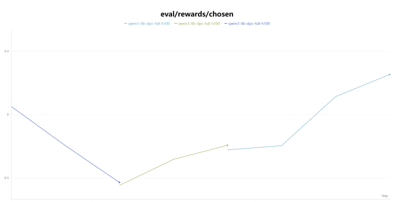
(a) Train loss

(b) Eval loss



(c) Reward accuracy

(d) Reward margin



(e) Reward (chosen)

(f) Reward (rejected)

Figure A.1: **Training dynamics and preference-learning diagnostics.** (a–b) Optimization convergence on training and evaluation sets. (c–f) Preference-learning signals, including reward accuracy, margin, and the reward separation between chosen and rejected responses.