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

Anonymous ACL submission

Abstract

Socratic questioning is essential for fostering critical thinking in domains such as education. However, existing evaluation methods lack a principled framework for systematically assessing this capability in Large Language Models (LLMs). To address this gap, we introduce SOCRATICEVAL, a benchmark for evaluating LLMs' Socratic questioning ability grounded in human preferences. Leveraging a human preference dataset, SOCRATICPREF, our analysis reveals a substantial discrepancy: state-of-the-art LLMs still lag significantly behind human performance in Socratic questioning. Furthermore, we conduct in-depth analyses and experiments to explore how the Socratic questioning capabilities of LLMs can be better leveraged and strengthened in real-world applications.

1 Introduction

With the widespread adoption of Large Language Models (LLMs), access to knowledge has become increasingly effortless, bringing the role of human cognition into sharper focus(Graesser et al., 2005), and raising concerns that model biases or misinformation may shape it(Huang et al., 2025; Bender et al., 2021). Socratic Questioning (SQ) (Fig.1) is a foundational technique in cognitive science (Paul and Elder, 2006) that promotes critical thinking through structured, heuristic questioning, thereby helping mitigate cognitive biases(Overholser, 1993). Current research(Wei et al., 2022; Madaan et al., 2023) is shifting from merely enhancing LLMs' ability to answer questions toward fostering deeper cognitive engagement. As a result, LLM-based SQ is gaining increasing attention, both as a research direction and for its promising applications (Liu et al., 2024; Kargupta et al., 2024; Held et al., 2025). Fields such as education and mental health urgently need dialogue models that promote reflection through questioning, rather than passive acceptance or indoctrination.

Figure 1: An example of the Socratic Questioning and compare two different strategies under one topic.

Despite its demonstrated value, evaluating this capability remains non-trivial. Prior work (Ang et al., 2023) has focused primarily on turn-level performance. However, the classic Socratic method (Plato, 1992) also emphasizes the benefits of collaborative questioning (e.g., Elenchus) – termed the *Socratic Strategy*. Existing benchmarks(Suzgun et al., 2023) 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, they fail to capture the nuanced and subjective aspects of effective Socratic questioning (Liu et al., 2016; Hernández, 2023).

To address these challenges, we first establish a theoretical framework for quantifying SQ abil-

ity (Table 1). We decompose SQ into two dimensions: *Question Generation* and *Strategy Utilization*. Building on this framework, we propose a unified benchmark, SOCRATICEVAL, to systematically examine the behavior and performance of LLMs across diverse domains in multi-turn dialogue settings (Fig. 2). To evaluate the alignment of LLM with human preferences on SQ, we first construct a human-annotated preference dataset SOCRATICPREF. This dataset is used to evaluate a range of frontier LLMs based on Bradley-Terry model, including Qwen3(Yang et al., 2025), DeepSeek(DeepSeek-AI et al., 2025), Grok(xAI, 2025), GPT(OpenAI, 2025), and Gemini(DeepMind, 2025). Our experiments reveal that current LLMs exhibit a significant gap compared to human preferences.

Furthermore, we conduct ablation studies and prompting experiments to investigate the strategy utilization ability, complemented by in-depth dialogue case analysis. Our findings provide practical insights into how SQ capabilities of LLMs can be better leveraged in real-world applications. Finally, we explore whether SQ performance can be further improved through preference alignment. We used questioning fallacies (Li et al., 2025; Rafailov et al., 2023) as a representative task to verify the effectiveness of preference alignment (Rafailov et al., 2023) based on reinforcement learning. 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 construct SOCRATICRREF, a human preference dataset, and conduct an evaluation of LLMs, revealing their gaps between human preferences.
- 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 use questioning fallacies as a representative task to verify preference alignment methods, suggesting practical directions for improving SQ in LLMs.

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

2 Socratic Questioning Task

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 Table 1.

Aspect 1: Question Generation **Socratic Questions** aim 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 Questions into several types.

Aspect 2: Strategy Utilization Classical SQ theory (Plato, 1992) centers on two core strategies: *Elenchus* (Reflective Strategy) and *Maieutics* (Guiding Strategy), respectively aimed at revealing contradictions and stimulating reasoning. Examples of these two strategies are shown in Fig. 1.

2.2 Task Definition

We evaluate the SQ ability of LLMs in dialogue interactions. This setup 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 assess model performance along three dimensions: **Effectiveness**, **Diversity**, and **Conciseness**. Effectiveness measures how well generated questions align with human preferences. Diversity captures the model's capacity to employ varied questioning types or strategies throughout the dialogue. Additionally, we observe that overly long or complex questions may impose a high cognitive load, potentially reducing their accessibility and usability. To account for this, we introduce Conciseness, which evaluates question length and structural simplicity.

3 SocraticEval Framework

This section introduces SOCRATICEVAL (Fig. 2), an LLM-based dialogue simulation framework for automatically evaluating SQ capabilities.

3.1 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: at the n-th turn, Socrates

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.
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.

raises a question Q_{n-1} based on Plato's previous argument and the dialogue history H. Formally,

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

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:

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

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 include claim, assumption, evidence, and reasoning components.

3.2 Fallacy-based Dialogue Generation

To further increase task 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). Examples are shown in Table 2. For each fallacious instance, Plato selects the most suitable rhetorical appeal to obscure the underlying fallacy and reinforce the surface-level plausibility of the claim.

3.3 Environment Dataset

Debate naturally encourages critical thinking through structured discourse, making it suitable as an environment for task. Following ORCHID(Zhao et al., 2023), we collect open-domain debate topics from prominent Chinese debate competitions (Appendix A) covering 8 distinct scenarios (Fig. A.1). For each curated statement, we use an LLM to generate a three-turn demonstration. We refer to this topic dataset as SOCRATICENV, which contains 975 topics and each framed with opposing stances. We partition SOCRATICENV into test and training sets; the test set contains 300 topics and is used for evaluation, while the training set is used for subsequent model improvement.

3.4 Human Preference Dataset

Based on the test set, we constructed a human preference dataset, SOCRATICPREF, to evaluate the alignment of LLM-generated dialogues with human preferences. Following the same experimental setup, we first used an LLM to generate three distinct dialogues for each Socratic strategy. Annotators then selected the dialogue they considered the best among the three and refined it iteratively according to a given scale until it fully aligned with their preferences. Finally, a second annota-

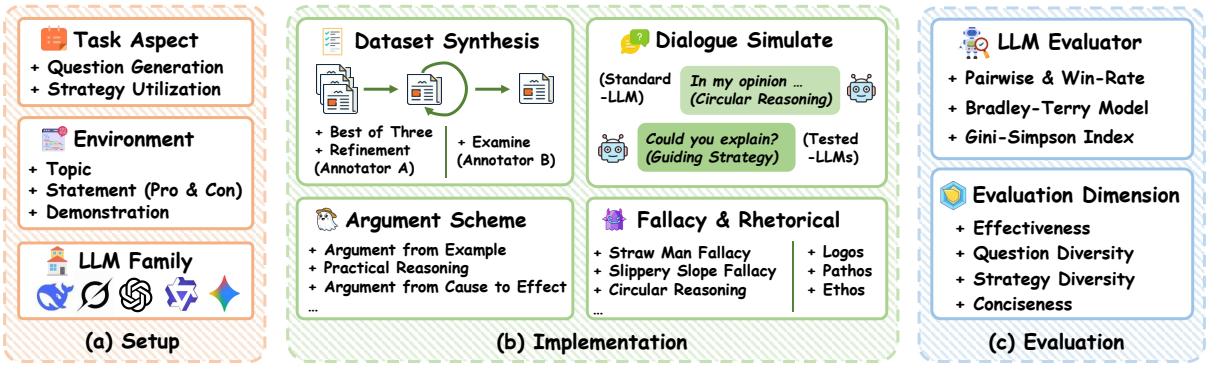


Figure 2: Overall framework of SOCRATICEVAL, illustrating the task setup, dataset synthesis, dialogue simulation, and evaluation pipeline.

tor(Appendix H) reviewed each selected dialogue to ensure it also met their preference criteria. For each of the two positions on a given topic, we constructed dialogues using two distinct Socratic strategies, resulting in a total of 1,200 dialogues.

3.5 LLM Evaluator

To measure diversity, we use an LLM to classify each question by strategy type and then compute the Gini-Simpson index. However, according to annotator feedback, evaluating effectiveness is highly subjective. Directly using the LLM-as-a-Judge method to generate an effectiveness score often yields unreliable results.

Inspired by Chatbot Arena (Chiang et al., 2024), we employ pairwise preference comparisons and win rates to evaluate effectiveness. Specifically, for a given dialogue position, we have a human preference dialogue and a set of test dialogues. We construct a ranking pool of dialogues via equal-probability random sampling without replacement and independently repeat this sampling procedure to increase the coverage of pairwise comparisons. Then, we use an LLM to rank the data in the ranking pool, and all rankings are decomposed into pairwise comparisons. Finally, we use the Bradley-Terry (BT) model (Zhang et al., 2024) to compute BT scores and win rates.

4 Experiment

4.1 Experimental Setting

In our experiments, each dialogue comprises 15 utterances: 8 responses from Plato and 7 Socratic questions. To balance difficulty and dialogue progression, we set a fallacy rate of 0.4, defined as the proportion of responses that contain fallacies.

We evaluate a diverse set of LLMs, including Qwen3, DeepSeek, Grok, GPT, and Gemini, and compare three configurations: zero-shot, few-shot, and thinking. Additionally, we employ DeepSeek-V3.2 to synthesize the dataset, simulate Plato's role in dialogue, and serve as the judge model during evaluation.

4.2 Main Results

Table 3 presents the experimental results. **Compared to human preference dialogues, LLMs still exhibit a substantial performance gap in conducting Socratic questioning.**

GPT-5.1, configured with few-shot and thinking, achieves the highest Effectiveness score and is the only model that surpasses human preference. The best-performing models among the other configurations are DeepSeek-V3.1 and Qwen3-235B-A22B, which lag behind human performance by 16.8 and 14.4 points, respectively.

Models exhibit a notable deficiency in Question Diversity. Almost all models score below 60 in this metric, indicating that their dialogues typically contain only one or two question types. In contrast, the human preference dialogues achieve a diversity score of 91.9, suggesting that, on average, each such dialogue includes four or more distinct question types.

To establish a baseline for Strategy Diversity, we conducted experiments using two strategies sampled with equal probability from SOCRATICPREF, which yielded a score of 100. **Among the evaluated models, GPT-4o with few-shot achieves the highest score, with a distribution of approximately 70% Refutative and 30% Guiding strategies.** The next best is Qwen3-14B, which attains scores of 48.5 and 52.4, corresponding to an 85%–15% split in favor of the Refutative strategy. Scores for the remaining models are generally

Models	F.S.	Th.	Eff.	Q.D.	S.D.	Con.	Avg.
Qwen3-8B	✗	✗	37.8±1.7	51.4±2.2	18.0±6.0	**54.66±0.3**	40.5±1.7
Qwen3-14B	✗	✗	48.8±1.8	53.7±2.2	**48.5±8.2**	52.13±0.3	**50.8±2.2**
Qwen3-32B	✗	✗	51.0±1.8	**60.4±2.0**	38.0±8.2	53.84±0.3	**50.8±2.2**
Qwen3-235B-A22B	✗	✗	59.2±2.1	48.6±2.4	33.7±7.7	33.21±0.5	43.7±2.1
DeepSeek-V3.1	✗	✗	**60.1±2.1**	53.3±2.3	8.7±4.6	51.60±0.4	43.4±1.4
Grok-4-Fast	✗	✗	20.9±1.9	17.8±2.5	34.9±13.6	47.72±0.7	30.3±3.5
Grok-4.1-Fast	✗	✗	44.9±2.4	35.6±2.7	8.1±5.4	47.06±0.5	33.9±1.6
Gemini-2.5-Flash	✗	✗	51.3±2.0	55.8±2.2	21.7±7.3	53.90±0.4	45.7±2.0
Human Preference	–	–	**76.9±2.1**	**91.9±0.3**	**100.0±0.2**	**55.69±0.3**	**81.1±0.5**
Qwen3-8B	✓	✗	42.9±2.2	52.7±2.2	31.4±1.4	55.60±0.4	45.7±0.9
Qwen3-14B	✓	✗	38.5±2.1	52.6±2.1	50.3±7.8	49.47±0.3	47.7±2.1
Qwen3-32B	✓	✗	52.9±2.1	**62.3±1.9**	50.4±8.0	50.13±0.3	53.9±2.1
Qwen3-Next-80B	✓	✗	59.2±2.3	51.8±2.3	11.1±5.0	37.97±0.4	40.0±1.5
Qwen3-235B-A22B	✓	✗	**62.0±2.0**	49.8±2.1	17.5±6.2	37.2±0.4	41.6±1.7
DeepSeek-V3.1	✓	✗	54.6±2.2	51.6±2.2	12.5±5.1	50.26±0.4	42.2±1.5
Grok-4-Fast	✓	✗	50.7±2.2	45.1±2.4	11.8±5.4	45.02±0.4	38.2±1.6
Grok-4.1-Fast	✓	✗	25.5±2.1	38.0±3.0	12.6±6.6	43.89±0.5	30.0±1.9
GPT-4o	✓	✗	42.9±2.0	58.4±2.1	**88.8±5.2**	**61.09±0.4**	**62.8±1.5**
Gemini-2.5-Flash	✓	✗	51.7±2.1	58.0±1.9	33.2±8.0	48.24±0.3	47.8±2.1
Human Preference	–	–	**76.4±2.3**	**91.9±0.3**	**100.0±0.2**	**55.69±0.3**	**81.0±0.6**
Qwen3-8B	✓	✓	49.3±2.0	52.1±2.1	28.9±6.9	55.16±0.3	46.4±1.9
Qwen3-14B	✓	✓	50.5±2.1	52.8±2.1	**52.4±8.2**	49.44±0.3	**51.3±2.2**
Qwen3-32B	✓	✓	59.0±2.1	**62.1±1.7**	23.9±6.7	50.96±0.3	49.0±1.8
Qwen3-235B-A22B	✓	✓	33.9±3.0	55.6±2.8	16.8±7.5	46.45±0.4	38.2±2.1
DeepSeek-R1	✓	✓	49.4±2.9	54.6±2.2	17.4±6.7	**58.13±0.5**	44.9±1.9
Grok-4-Fast	✓	✓	60.6±2.2	50.7±2.2	12.3±5.3	45.52±0.3	42.3±1.5
Grok-4.1-Fast	✓	✓	19.5±2.3	40.8±3.9	9.8±7.6	43.39±0.7	28.4±2.2
GPT-5.1	✓	✓	**79.1±2.3**	60.3±2.0	14.7±6.4	31.53±0.4	46.4±1.8
Gemini-2.5-Flash	✓	✓	59.4±2.1	57.7±1.9	15.6±5.9	49.25±0.3	45.5±1.6
Gemini-2.5-Pro	✓	✓	28.7±2.4	61.8±2.3	2.8±3.2	50.60±0.4	36.0±1.2
Gemini-3-Pro-Preview	✓	✓	59.8±2.1	55.2±2.0	5.3±3.6	53.67±0.4	43.5±1.2
Human Preference	–	–	**76.9±2.5**	**91.9±0.3**	**100.0±0.2**	**55.69±0.3**	**81.1±0.6**

Table 3: **The Results of Experiments.** Among them, **F.S.** represents **Few-Shot**, **Th.** represents **Thinking**, **Eff.** represents **Effectiveness**, **Q.D.** represents **Question Diversity**, **S.D.** represents **Strategy Diversity**, **Con.** represents **Conciseness**, and **Avg.** represents **Average**. The best results in each column is highlighted in **bold**.

below 30, indicating that, in practice, only the Refutative strategy is reused across different topics.

Furthermore, most models achieve conciseness scores close to those of human preferences, with GPT-4o and DeepSeek-R1 even surpassing human-level.

4.3 Further Analysis

Overall Performance As illustrated in Fig. 3 (a–c), we plot diversity and conciseness on the x-axis and effectiveness on the y-axis to assess overall performance, with models positioned closer to the top-right corner indicating a more balanced capability. **Although a few models perform well in a single dimension, their overall performance still lags behind that of human preference baseline.**

Impact of Configuration Figure. 3 (d) shows the score distributions across different configurations (zero-shot, few-shot, and thinking). **First, few-shot prompting and thinking mode tend to enhance task performance.** Under different configurations, the average scores for Effectiveness and Question Diversity exhibit an upward trend. **Second, the impact of model configuration on score variance varies across settings.** Compared to zero-shot, few-shot reduces the variance in Effectiveness, whereas thinking mode increases it.

Preference for Question Type We analyze the distribution of question types and strategy types across models. As depicted in Fig. 3 (e), the Qwen3 family and Gemini-2.5-Flash favor Probing Consequences, while the Grok-4 family prefers 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 effectively.

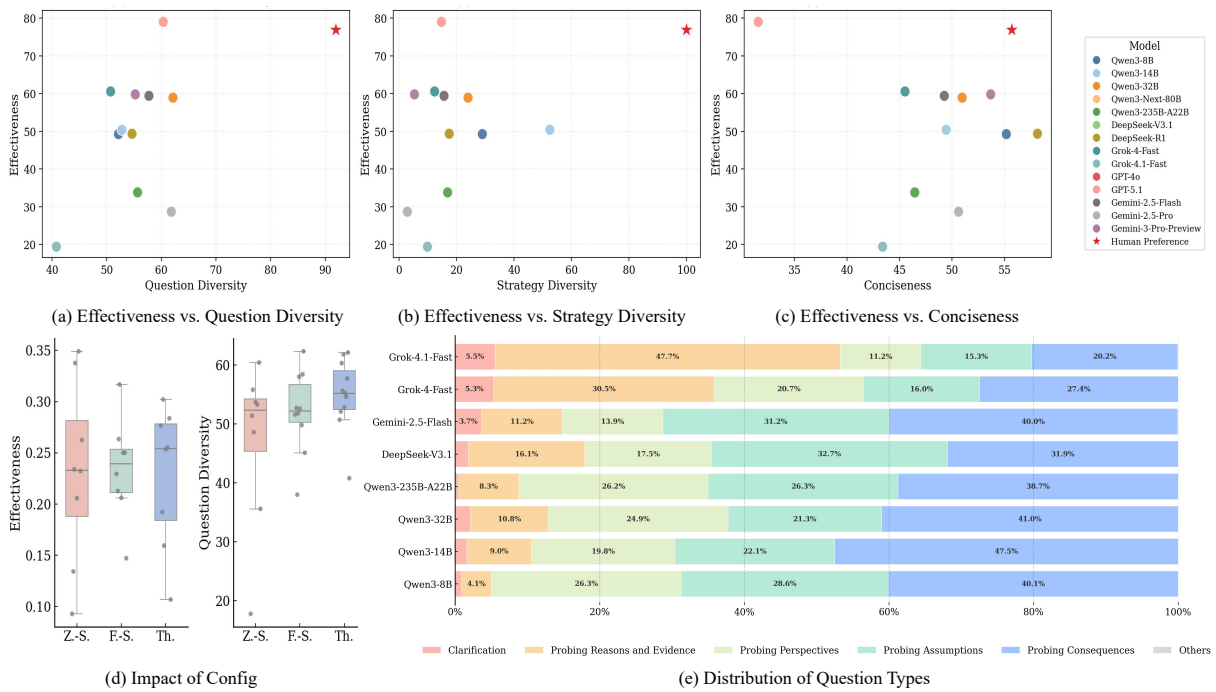


Figure 3: **Comprehensive Analysis of Experimental Results.** (a-c) The overall performance; (d) The score distributions across different configurations; (e) The fine-grained composition of question types.

4.4 Supplementary Experiments

In our experiments, results are obtained using an LLM-as-a-Judge framework based on annotations produced by Plato. To ensure the reliability of our results, we conducted supplementary experiments. First, human annotators verified annotation accuracy on 200 randomly sampled dialogues (100 argument responses, 100 fallacious responses), achieving accuracies of 0.98 and 0.96, respectively. We then assessed model and prompt consistency using Cohen's Kappa (Cohen and J., 1960). Replacing the judge model with Qwen3-32B and modifying prompt instructions yielded Kappa scores of 0.89 (model consistency) and 0.94 (prompt consistency), indicating high evaluation reliability.

Finally, we measured agreement between human evaluators and LLM using Fleiss' Kappa (Fleiss, 1971). Three annotators independently annotated 100 dialogues for pairwise preferences, and 50 dialogues with 50 embedded questions for classification. The resulting Kappa scores were 0.81 (pairwise preference), 0.75 (question classification), and 0.77 (strategy classification), demonstrating substantial agreement with human judgments.

5 Investigation of Strategy Utilization

The main results show that all evaluated models exhibit a pronounced tendency to adopt refutative strategies. We hypothesize that this pattern arises from artifacts in Plato's response behavior and the dialogue setup, rather than reflecting an intrinsic property of the evaluated models.

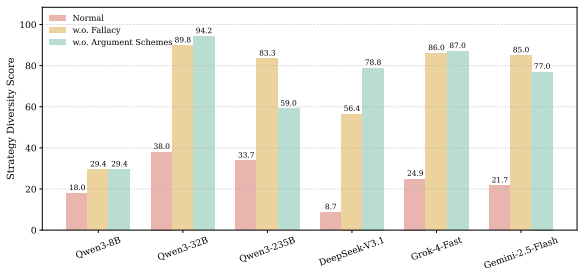


Figure 4: Strategy diversity in different ablation settings.

5.1 Influence of the Response Behavior

In SOCRATICEVAL, the participant model (Plato) generates responses based on predefined argument schemes and fallacies. To assess the impact of this response behavior 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. 4, 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 strat-

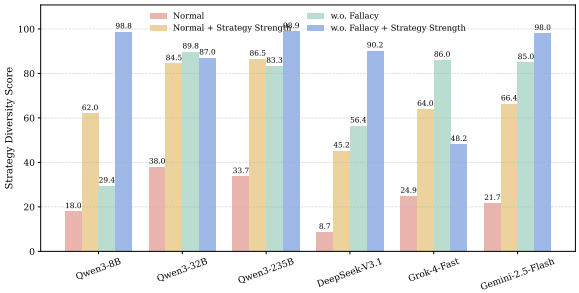


Figure 5: Effect of Strategy-Enhanced Prompting.

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.

egy variation.

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 provides detailed strategy introduction and examples, and explicitly instructs the model to plan its strategy before generating questions. We evaluate this method across different response mechanisms. The results, presented in Fig. 5, 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 dialogue histories 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 capabilities in question formulation, global planning, and dynamic adaptation.**

6 Towards Better Socratic Questioning

Finally, we explore a more complex task: how to further enhance the performance of LLM-based SQ in real-world application scenarios.

6.1 Preference Alignment

The core challenge of this task lies in aligning with diverse and complex human preferences, as there is no single optimal SQ strategy applicable to all scenarios. For instance, in a courtroom debate, we expect the LLM to precisely identify fallacies and provide rebuttals; in educational settings, however, we hope the model offers guidance rather than refutation.

We address this challenge through preference alignment using reinforcement learning. To simulate the real-world task, we first design a simplified task—questioning fallacies—whose preferences are easily defined and validated experimentally. Given a fallacious response, an LLM can generate numerous plausible follow-up questions, and we aim to favor those that directly target the fallacy. To this end, we construct a pairwise preference dataset, SOCRATICFALLACY, and train the LLM to question fallacies using Direct Preference Optimization (DPO) based on this dataset.

6.2 Pairwise Preference Dataset

The SOCRATICFALLACY 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. This dataset is compatible with reinforcement learning algorithms beyond DPO.

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.

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. Finally, the human annotators perform a sampling check to ensure it meets the preferences.

6.3 Results

Leveraging the SOCRATICENV training set, we construct SOCRATICFALLACY, which comprises 9,578 ranked preference pairs. For implementation, we use Qwen3-14B to synthesize questions and employ DeepSeek-V3.2 as the judge model. The base model, Qwen3-8B, is fully fine-tuned on a single NVIDIA H100 (80GB) GPU. Detailed training configurations are provided in Appendix E.

We define the proportion of questions that target fallacies as the effectiveness for this task. The results are presented in Table 5. We compare the model under two prompt settings: normal and fallacy-strength. The former instructs the model to generate SQ in a standard manner, while the latter emphasizes fallacy identification through enhanced instructions and demonstrations. Our model shows consistent performance in both normal (+18.89) and fallacy-strength (+19.88) settings.

7 Related Work

7.1 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 (Calvo Figueras and Agerri, 2024). Significant attention has also been directed toward Socratic Question Generation, 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 (Chang, 2023), fine-tuning (Liu et al., 2024), and RL method (Dinucu-Jianu et al., 2025). 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).

7.2 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., 2023). 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.

8 Conclusion

We introduce SOCRATICEVAL, a benchmark for evaluating LLM-based SQ grounded in human preferences. Using the human preference dataset SOCRATICPREF, our analysis uncovers a significant gap: SOTA LLMs still fall considerably short of human-level SQ performance. Further investigation attributes this limitation to underlying response mechanisms and demonstrates that strategy-enhanced prompting can alleviate the diversity deficit. Finally, we explore the potential of preference alignment to further improve SQ capabilities. Taking questioning fallacies as a representative task, we construct SOCRATICFALLACY and show that DPO yields substantial performance gains. This work establishes a principled evaluation framework and provides actionable insights for advancing SQ abilities in 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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Blanca Calvo Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 105–116, Miami, FL, USA. Association for Computational Linguistics.

Edward Y. Chang. 2023. Prompting large language models with the socratic method. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0351–0360.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Cohen and J. 1960. A coefficient of agreement for nominal scales. *Educational Psychological Measurement*, 20(1):37–46.

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

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. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. 2025. From problem-solving to teaching problem-solving: Aligning LLMs with pedagogy using reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 272–292, Suzhou, China. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

A.C. Graesser, P. Chipman, B.C. Haynes, and A. Olney. 2005. Autotutor: an intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618.

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

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

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. 43(2).

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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20482–20503, Suzhou, China. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.

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, Barcelona, Spain. Association for Computational Linguistics.

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*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 37, pages 85693–85721. Curran Associates, Inc.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 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 5602–5621, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

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

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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.

Yifan Zhang, Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. 2024. Beyond bradley-terry models: A general preference model for language model alignment. In *International Conference on Machine Learning*.

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, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

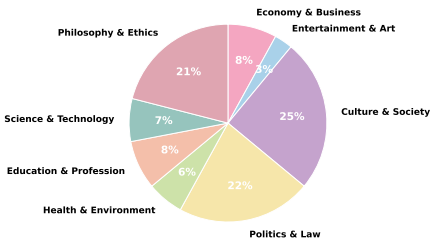


Figure A.1: Topic distribution of the evaluation dataset.

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, detailing the scale and nature of the argumentative data collected.

Figure A.1 illustrates the distribution of topics in the evaluation dataset. The proportions indicate the relative frequency of each topic category.

B Argument Schemes and Fallacy Types

This section details the theoretical framework used to categorize the logical structures within the dialogue dataset.

B.1 Argument Schemes

We utilize a set of standard argument schemes to characterize the logical structure of Plato's valid responses in the Socratic dialogue. Table 2 presents the definitions of these schemes along with illustrative examples adapted for the Socratic context.

B.2 Fallacy Types

To evaluate the robustness of Socratic questioning, we incorporate specific logical fallacies. Table 3 specifies the fallacy types used for controlled fallacy injection experiments and for evaluating the model's ability to detect invalid reasoning.

C Prompt Templates

This section presents the prompt templates used to generate and evaluate Socratic dialogues, ensuring the reproducibility of all prompt-based experiments. We designed distinct system instructions for the different roles in the framework:

- **Plato (Argument-based):** Instructs the agent to construct valid arguments based on specific argument schemes (Table 4).
- **Plato (Fallacy-based):** Instructs the agent to intentionally inject logical fallacies into the response for testing purposes (Table 5).
- **Socrates:** Guides the agent to perform critical questioning based on the dialogue history (Table 6).
- **Evaluator:** Provides instructions for the LLM judge to rate the dialogue quality based on the defined rubrics (Table 7).

D Demonstration of Socratic Questioning

To illustrate the nature of the generated data, Table 8 provides a concrete example of a multi-turn Socratic dialogue. This excerpt demonstrates how the Socratic agent probes the interlocutor's adherence to procedures, moving from specific claims to deeper conceptual definitions.

E Details of DPO Training

Table 9 summarizes the hyperparameters used for the Direct Preference Optimization (DPO) phase. These settings were selected to balance training stability with effective preference learning on NVIDIA H100 80GB GPUs.

F Training Dynamics

Figure A.2 visualizes the training dynamics and preference-learning diagnostics throughout the DPO process. The subplots track the convergence of training and evaluation loss, as well as key reward metrics (accuracy, margin, and separation), demonstrating the model's increasing alignment with the preference data.

G Pairwise comparison matrices.

Figure A.3 presents the pairwise evaluation results across all models under three prompting settings: *zero-shot*, *few-shot-thinking*, and *few-shot*. For each setting, we report both the *match-count matrix* and the corresponding *win-rate matrix*. Each cell (i, j) denotes the number of instances (or the proportion of wins) in which the output produced by model i is preferred over that produced by model j. The match-count matrices show the raw number of wins across the evaluation set, while the win-rate matrices normalize these counts to facilitate comparison across model pairs. Diagonal entries are omitted since they correspond to self-comparisons.

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.

H Human Preference Scale

To systematically evaluate the quality of the generated dialogues, we design a human preference scale that captures multiple dimensions, including format correctness, structural constraints, content validity, and overall effectiveness. Annotators are asked to assess each dialogue according to the following criteria:

1. **Format Check**: Is the dialogue in valid JSON format?
2. **Number of Turns**: Does it contain three dialogues, each with 15 turns, including 8 turns by Plato and 7 turns by Socrates?
3. **Dialogue Content**: Are all of Plato's utterances argumentative statements, and are all of Socrates' utterances questions?
4. **Dialogue Quality**: Does the naturalness and fluency of the dialogue meet your expectations?
5. **Question Consistency**: Are the types of questions in the dialogue consistent with their assigned labels?
6. **Question Effectiveness**: Does the effectiveness of the questions in the dialogue meet your expectations?
7. **Strategy Consistency**: Is the strategy exhibited in the dialogue consistent with the predefined strategy type?
8. **Strategy Effectiveness**: Does the effectiveness of the adopted strategy in the dialogue meet your expectations?

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.

Plato: Response based on Argument Scheme
Profile You are a skilled debater in various types of dialogue. You are participating in a unique dialogeu called Socratic Dialogue. The Socratic Dialogue consists of a series of reponses and questions, including Plato and Socratic. Plato presents a series of responses to support his position. Socratic posed critical questions, questioned Plato's response, and pushed for further development of the dialogue. The role you play is Plato. The position you support is {{position}}. The demonstrations of this debate is {{demonstration}}. Requirements 1. You need to first read the history of debates refer to the given example, if any, and raise new response based on your position and Socratic's current questioning. 2. You only need to provide an response based on argument schemes, just one paragraph, without further breaking it down. 3. The argument you provide only needs to reflect the conclusion and premise, while the scheme is implicit. 4. Please try to ensure that your argument is free from logical fallacies. 5. Then, you need to provide another paragraph called Tips to indicate what scheme you are using and explain the assumptions, evidence, and reasoning in the argument in a simple sentence. Argument Schemes:{{Types of Argument Schemes}} Attention: Please do not provide any other content. Please provide the output in JSON format. Output format: {"Plato": {response}, "Tips": {"scheme": {scheme}, "explanation":{explanation}}} Dialogue History: {{history}}

Table 4: Prompt of Plato (Argument-based)

Plato: Response based on Fallacy
Profile You are a skilled debater in various types of dialogue. You are participating in a unique dialogeu called Socratic Dialogue. The Socratic Dialogue consists of a series of reponses and questions, including Plato and Socratic. Plato presents a series of responses to support his position. Socratic posed critical questions, questioned Plato's response, and pushed for further development of the dialogue. The role you play is Plato. The position you support is {{position}}. The demonstrations of this debate is {{demonstration}}. Requirements 1. You need to first read the history of debates refer to the given example, if any, and raise new response based on your position and Socratic's current questioning. 2. You only need to provide an response based on argument schemes, just one paragraph, without further breaking it down. 3. The argument you provide only needs to reflect the conclusion and premise, while the scheme is implicit. 4. Please try to ensure that your argument is free from logical fallacies. 5. Then, you need to provide another paragraph called Tips to indicate what scheme you are using and explain the assumptions, evidence, and reasoning in the argument in a simple sentence. Argument Schemes:{{Types of Fallacy}} Attention: Please do not provide any other content. Please provide the output in JSON format. Output format: {"Plato": {response}, "Tips": {"scheme": {scheme}, "explanation":{explanation}}} Dialogue History: {{history}}

Table 5: Prompt of Plato (Fallacy-based)

Socrates
Profile You are a skilled debater in various types of dialogue. You are participating in a unique dialogeu called Socratic Dialogue. The Socratic Dialogue consists of a series of reponses and questions, including Plato and Socratic. Plato presents a series of responses to support his position. Socratic posed critical questions, questioned Plato's response, and pushed for further development of the dialogue. The role you play is Socrates. The position you support is {{position}}. The demonstrations of this debate is {{demonstration}}. Requirements You need to first read the history of debates, refer to the given example, if any, and raise Socratic questions based on Plato's current response, which is the last dialogue in history. Attention: Please do not provide any other content. Please provide the output in JSON format. Output format: {"Socrates": {question}} Dialogue History: {{history}}

Table 6: Prompt of Socrates

Evaluator
Profile You are a referee skilled in various debate dialogues, judging a special dialogue. This is a Socratic dialogue between Plato and Socrates. Plato argues around his own position. Socrates engages in critical questioning, questioning Plato's arguments and helping to construct a more comprehensive system of argumentation. After each response, there is a tip, which points out the core argument schemes or logical fallacies in Plato's argument. Requirements Please carefully read the given dialogues and sort them in descending order based on the given evaluation criteria. Please do not consider the word count or complexity of the problem. Evaluation Dimension: Socrates' strategy is divided into two types, and their effectiveness is defined as follows: 1. Refutative Strategy(Elenchus): Its effectiveness lies in its ability to use interconnected inquiries to force the interlocutor into an unavoidable logical contradiction while defending their position, thereby destabilizing their original stance. 2. Guiding Strategy(Maieutics): Its effectiveness lies in its ability to gradually inspire the interlocutor through a step-by-step heuristic process, helping them clarify vague notions through their own reasoning and ultimately autonomously construct clearer, more profound insights. Attention Please provide the output in JSON format!!! Please do not provide any other content!!! Dialogue History: {"Dialogue1": {}, "Dialogue2: {}", ...}

Table 7: Prompt of Evaluator

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

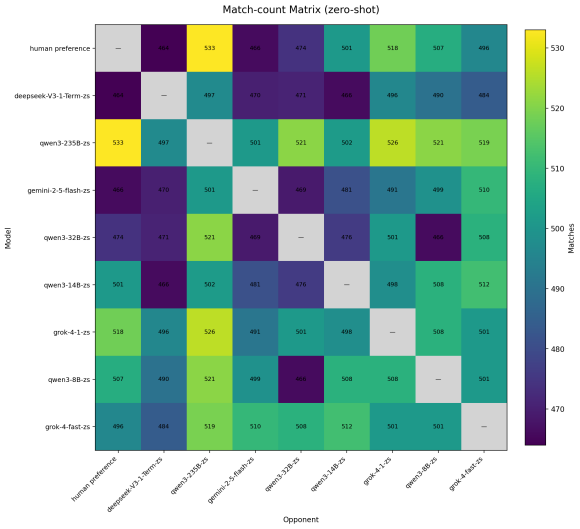
Category	**Setting**	**Category**	**Setting**
Batch size (per device)	4	Learning rae	5×10^{-7}
Evaluation batch size	4	LR scheduler	Cosine
Gradient accumulation steps	4 (effective batch size = 16)	Warmup ratio	0.1
Optimizer	AdamW (torch fused)	Training epochs	3
Precision	BF16 + TF32	Gradient checkpointing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

Table 9: Hyperparameters used for DPO training.

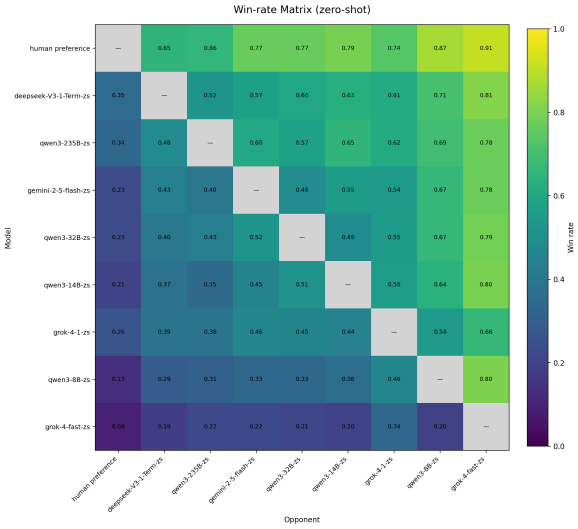
(a) Train loss (b) Eval loss (c) Reward accuracy

(d) Reward margin (e) Reward (chosen) (f) Reward (rejected)

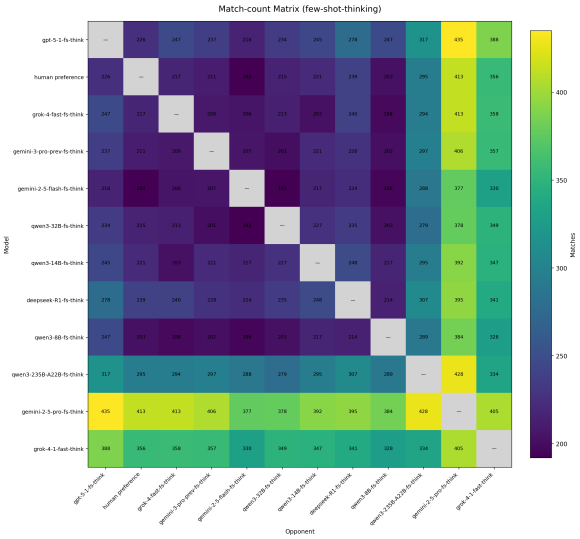
Figure A.2: **Training dynamics and preference-learning diagnostics.** (a–b) Optimization convergence on training and evaluation sets. (c–f) Preference-learning signals, including reward accuracy, margin, and the reward separation between chosen and rejected responses.



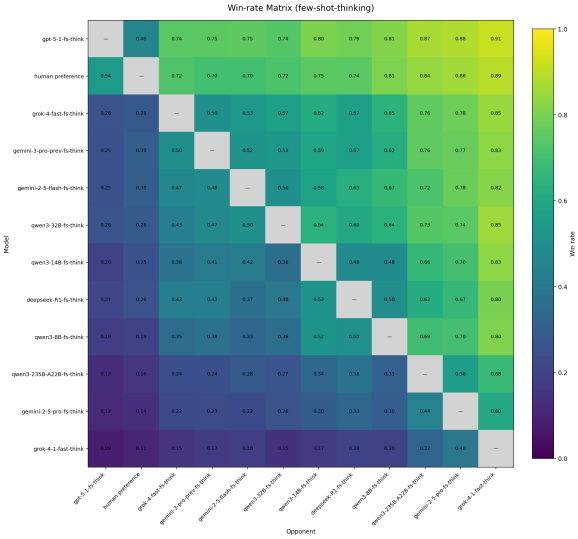
(a) Match-count Matrix (zero-shot)



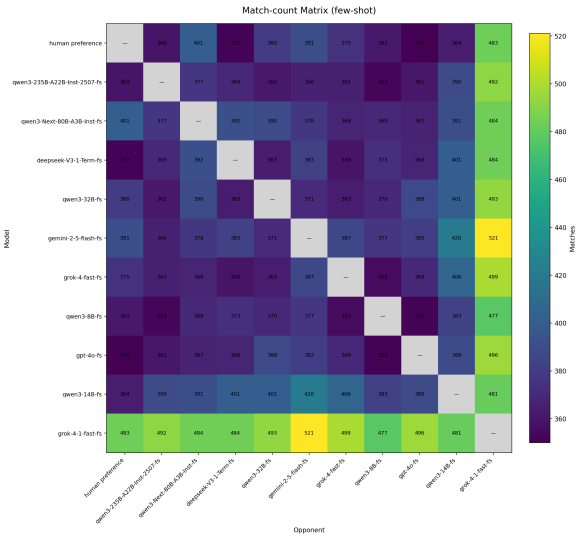
(b) Win-rate Matrix (zero-shot)



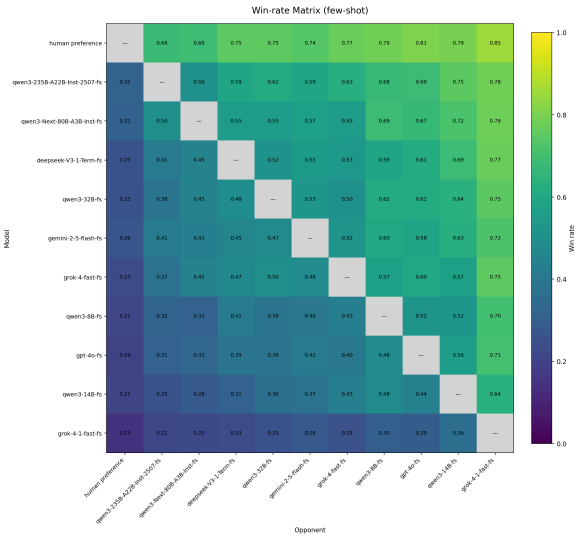
(c) Match-count Matrix (few-shot-thinking)



(d) Win-rate Matrix (few-shot-thinking)



(e) Match-count Matrix (few-shot)



(f) Win-rate Matrix (few-shot)

Figure A.3: Comparison of match-count and win-rate matrices under zero-shot, few-shot-thinking, and few-shot settings.