JudgeBoard: Benchmarking and Enhancing Small Language Models for Reasoning Evaluation

Zhenyu Bi[1], Gaurav Srivastava[1], Yang Li[2], Swastik Roy[3], Meng Lu[1], Morteza Ziyadi[3], Xuan Wang[1]

[1]Virginia Tech
[2]College of William and Mary
[3]Amazon AGI
{zhenyub,gks,menglu}@vt.edu, yli102@wm.edu,{roswasti, mziyadi}@amazon.com

Abstract

While small language models (SLMs) have shown promise on various reasoning tasks, their ability to judge the correctness of answers remains unclear compared to large language models (LLMs). Prior work on LLM-as-a-judge frameworks typically relies on comparing candidate answers against ground-truth labels or other candidate answers using predefined metrics like entailment. However, this approach is inherently indirect and difficult to fully automate, offering limited support for fine-grained and scalable evaluation of reasoning outputs. In this work, we propose JudgeBoard, a novel evaluation pipeline that directly queries models to assess the correctness of candidate answers without requiring extra answer comparisons. We focus on two core reasoning domains: mathematical reasoning and science/commonsense reasoning, and construct task-specific evaluation leaderboards using both accuracy-based ranking and an Elo-based rating system across five benchmark datasets, enabling consistent model comparison as judges rather than comparators. To improve judgment performance in lightweight models, we propose MAJ (Multi-Agent Judging), a novel multi-agent evaluation framework that leverages multiple interacting SLMs with distinct reasoning profiles to approximate LLM-level judgment accuracy through collaborative deliberation. Experimental results reveal a significant performance gap between SLMs and LLMs in isolated judging tasks. However, our MAJ framework substantially improves the reliability and consistency of SLMs. On the MATH dataset, MAJ using smaller-sized models as backbones performs comparatively well or even better than their larger-sized counterparts. Our findings highlight that multi-agent SLM systems can potentially match or exceed LLM performance in judgment tasks, with implications for scalable and efficient assessment.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in a wide range of natural language processing tasks, including reasoning, question answering, and evaluation (Achiam et al. 2023; Dubey et al. 2024; Liu et al. 2024a; Yang et al. 2025). A growing body of work has explored the use of LLMs not only as generators of content but also as evaluators, where models are tasked with assessing the relative quality of candidate outputs (Zheng et al. 2023; Tan et al. 2024). These approaches have been applied in various reasoning domains, often relying on indirect supervision via entailment metrics or comparisons to gold-standard labels or the other candidate outputs (llm-as-a-judge). While effective in some settings, these methods are inherently limited: they rely on comparison-based depend on predefined metrics that may not capture nuanced reasoning errors, and they often require human-curated labels, which hinders scalability and generalization.

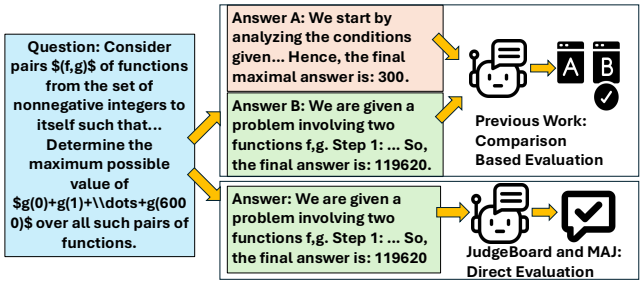

Figure 1: Comparison of JudgeBoard and MAJ with previous works. Unlike previous works that usually follow a comparison-based evaluation pipeline, JudgeBoard and MAJ focus on direct evaluation of the factual correctness of the reasoning questions.

In contrast, small language models (SLMs) offer a more efficient and accessible alternative to LLMs (Yang et al. 2025; Abdin et al. 2024; Team et al. 2024; Srivastava, Cao, and Wang 2025; Srivastava et al. 2025a,c), particularly in resource-constrained environments. Recent studies have shown that SLMs can perform competitively on various reasoning tasks when equipped with appropriate prompting or fine-tuning strategies (Wei et al. 2022). However, their ability to judge the correctness of answers rather than generate them remains underexplored. This gap is critical, as scalable and reliable evaluation is essential for deploying language models in real-world applications.

In this work, we propose JudgeBoard, a novel evaluation pipeline that directly queries models, particularly SLMs, to assess the correctness of candidate answers without requiring extra answer comparisons. Figure 1 demonstrates the difference between our proposed pipeline and prior works. We focus on two core domains: mathematical reasoning

and science/commonsense reasoning. To support this setup, we construct task-specific evaluation leaderboards across five benchmark datasets, using both accuracy-based rankings and an Elo-style rating system to enable consistent and fine-grained comparison of models in their role as judges. The results highlight the need for using multiple evaluation metrics when doing evaluation, provide insights on the judging abilities of different SLMs, and demonstrate the significant performance gap between the best-performing Large Language Models and the best-performing Small Language Models (SLMs).

To address the limitations of individual SLMs in judgment tasks, we introduce MAJ (Multi-Agent Judging), a novel Elo-based framework that leverages multiple interacting SLMs to approximate the judgment performance of LLMs. Inspired by recent advances in multi-agent collaboration and self-consistency techniques (Zhuge et al. 2024; Guo et al. 2024; Wu et al. 2024), MAJ enables lightweight models to collectively reason about answer correctness, improving both reliability and robustness. Our experiments reveal a substantial performance gap between isolated SLMs and LLMs in judgment tasks, but also show that MAJ significantly narrows this gap. On the MATH dataset (Lightman et al. 2023), our MAJ framework using Qwen3-14B model as the backbone outperform the best-performing LLM by an average of 2% in judging accuracy across all categories. Our findings suggest that with the right collaborative framework, SLMs can rival or even surpass LLMs in evaluative reasoning; this finding paves the way for more efficient and democratized model assessment.

2 Related Works

Language Models as Judges LLMs-as-a-judge frameworks have emerged as an alternative to human annotators and traditional metrics. Tan et al. evaluated 11 LLMs across 20 NLP tasks, finding variable reliability depending on task type and data source. Tang, Duan, and Cai surveyed the paradigm across functionality, methodology, applications, meta-evaluation, and limitations, highlighting interpretability benefits while cautioning against biases in proprietary models. Practical implementations include MT-Bench and Chatbot Arena (Zheng et al. 2023), which demonstrate that LLMs can approximate human evaluators with proper calibration.

LLM Reasoning Prompting Prompting techniques have enhanced LLM reasoning capabilities. Chain-of-Thought (CoT) (Wei et al. 2022) breaks reasoning into sequential steps, while self-consistency (Wang et al. 2022) aggregates multiple reasoning paths. Tree-of-Thought (ToT) (Yao et al. 2023) maintains structured intermediate steps, and Graph-of-Thought (Besta et al. 2024) uses directed graphs to revisit and combine reasoning paths. Additional X-of-Thought variants address specific tasks (Chen et al. 2023b; Sel et al. 2024; Bi et al. 2024). Recent reasoning benchmarks (Srivastava et al. 2025b,c) also evaluate and compare the effect of different reasoning prompting methods on LLMs.

Multi-Agent Reasoning Multi-agent LLM frameworks have shown improved reasoning over single-agent systems (Wu et al. 2024; Chen et al. 2023a; Lu et al. 2024). Research on multi-agent debate dynamics (Wang et al. 2023, 2024; Pezeshkpour et al. 2024) reveals that most interaction protocols are manually defined (Wu et al. 2023; Chan et al. 2023) or follow simple formats like majority voting and summarization (Chen, Saha, and Bansal 2023; Liang et al. 2024; Chan et al. 2023). GPTSwarm and OptAgent (Zhuge et al. 2024; Bi et al. 2025) models multi-agent systems as graph networks, enabling algorithmic optimization of interaction patterns and agent-level prompts.

Small Language Models Small Language Models (SLMs) offer efficient alternatives for reasoning and evaluation with reduced computational costs. Recent instruction-tuned SLMs with fewer than 3 billion parameters achieve strong reasoning capabilities through high-quality training data (Gunasekar et al. 2023; Abdin et al. 2024; Bai et al. 2023; Team et al. 2024). Recent work (Srivastava, Cao, and Wang 2025; Srivastava et al. 2025a) explored methods to enhance SLMs' reasoning abilities through targeted training and architectural improvements. Bansal et al. found that SLMs can effectively judge reasoning quality with appropriate prompting, while Liu et al. identified key factors enabling effective reasoning, including training data quality and model architecture. Efficient reasoning frameworks (Shao et al. 2024) demonstrate that training on mathematical and logical reasoning data enables SLMs to compete with larger models.

3 JudgeBoard

3.1 Pipeline Overview

We propose a novel evaluation pipeline, **JudgeBoard**, that evaluates the models' judging ability by querying models to assess the correctness of candidate answers without requiring extra answer comparisons. This approach treats models as direct evaluators, enabling scalable and flexible assessment across diverse reasoning tasks. While our judging protocol is direct, meta-evaluation of judge quality necessarily requires gold-standard labels for ground truth. The overall pipeline of JudgeBoard is demonstrated in Figure 2. JudgeBoard consists of four stages:

- **Candidate Answer Collection**: Student model is provided with a set of reasoning questions and is prompted to provide answers.
- **Judgment Collection**: A set of judge models independently evaluate the correctness of each candidate answer.
- **Pairwise Competitions**: For the same question-answer pair, evaluations from different judges are compared in a pairwise fashion. We then calculate the Elo score based on the competition results.
- **Leaderboard Construction**: We compute two types of rankings: Accuracy-based ranking and Elo-style-based ranking, to assess and compare model performance across tasks and subcategories.

3.2 Model-as-Judge Protocol

To operationalize the Model-as-Judge pipeline, we design a structured prompting protocol that ensures consistency and

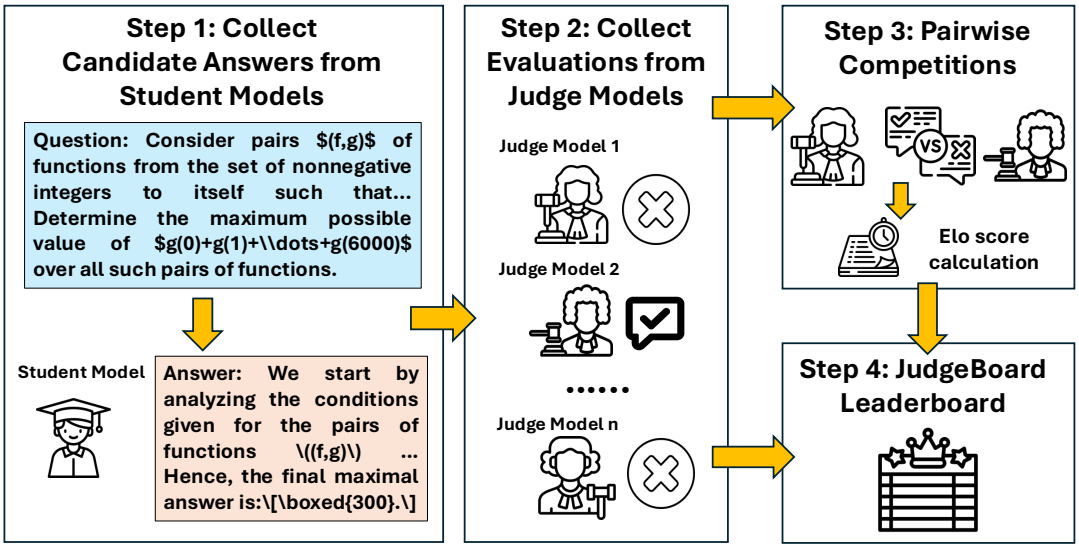

Figure 2: Overview of the JudgeBoard Pipeline

objectivity in model evaluations. Each judge model is presented with the original question, the candidate answer generated by the student model, and a prompt instructing the judge to determine whether the answer is correct or incorrect and output its reasoning. Prompts are carefully engineered to minimize ambiguity and bias, and to encourage factual, step-by-step reasoning.

Additionally, we ensure similar but different reasoning by assigning the agents with the same baseline reasoning prompt but different agent profiles in system prompts (see Appendix A for detailed prompts). The agent profiles were manually crafted to reflect common reasoning strategies found in human problem-solving, such as deductive reasoning, logical reasoning, and robust reasoning.

3.3 Pairwise Evaluation and Elo Rating Calculation

After collecting judgments, we conduct pairwise comparisons between candidate answers. For each question, judgments of the answers are evaluated against each other. We determine outcomes based on agreement with gold-standard labels: a judge wins if its judgment matches the gold label while its comparator does not. To quantitatively assess the relative performance of judge models, we adopt an Elo-style rating system inspired by the ChatbotArena.

The Elo system provides additional value beyond simple accuracy metrics through three key mechanisms. First, it accounts for question difficulty by awarding judges more credit for correct judgments on questions where other judges failed, effectively weighting performance by task complexity. Second, it measures consistency across diverse question types, rewarding models that reliably agree with gold labels regardless of domain or difficulty rather than those that perform well only on specific subsets. Finally, the dynamic rating system better distinguishes models with similar raw accuracy by capturing the relative strength demonstrated through head-to-head comparisons, revealing nuanced performance differences that aggregate accuracy scores may obscure. This enables a more granular evaluation of model capabilities across reasoning tasks.

Rating Update Mechanism Let R_i and R_j denote the current Elo ratings of models i and j, respectively. The expected score for model i is computed using the standard logistic function: $E_i = \frac{1}{1+10^{(R_j-R_i)/400}}$. Following the outcome of the match, the rating of model i is updated by $R_i' = R_i + K(S_i - E_i)$, where R_i is the rating of model i, $S_i \in \{1, 0.5, 0\}$ represents the actual match outcome (win, draw, loss), and $K = 10$ is a constant controlling the magnitude of rating updates. All models are initialized with a uniform baseline rating. To ensure robustness, each model participates in a large number of comparisons across diverse questions and judge profiles. The final Elo score reflects the model's overall performance relative to its peers.

3.4 Leaderboard Construction

We construct task-specific leaderboards using two complementary metrics:

Accuracy-Based Ranking Measures agreement with the gold-label given by the original dataset across three dimensions: Overall Accuracy (correctness across all questions), Student Wrong (SW) Accuracy (ability to identify incorrect answers), and Student Right (SR) Accuracy (ability to validate correct answers). These metrics reveal not only judging performance but also potential sycophacy toward agreement.

Elo-Style Rating System Models are evaluated in pairwise comparisons, and ratings are updated based on relative performance. This enables fine-grained differentiation and comparison even in ambiguous cases.

Model	Algebra			Number Theory			Counting and Probability		
	Overall Accuracy	SW Accuracy	Elo	Overall Accuracy	SW Accuracy	Elo	Overall Accuracy	SW Accuracy	Elo
Large Language Models (>14B)									
Qwen3_30B_A3B	0.7789	0.570	1070.8	0.9077	0.825	1110.3	0.8889	0.793	1089.3
Qwen3_32B	0.7632	0.516	1068.8	0.8846	0.778	1100.6	0.8888	0.782	1091.4
Gemma3-27b	0.6947	0.398	1038.8	0.8077	0.603	1069.2	0.7944	0.621	1054.4
Qwen2.5-72b	0.6368	0.269	1017.2	0.6846	0.365	1020.9	0.7167	0.471	1026.4
Llama3-70b	0.6053	0.215	1005.5	0.6462	0.270	1006.0	0.7056	0.414	1022.5
Llama4-17B	0.5895	0.183	999.6	0.7000	0.381	1026.9	0.6778	0.356	1012.6
Mixtral-8x22B	0.5579	0.097	987.9	0.5462	0.095	973.2	0.5667	0.126	977.0
Mixtral-8x7B	0.5316	0.054	978.2	0.5692	0.111	979.2	0.5389	0.069	967.0
Small Language Models (<14B)									
Gemma3_12b	0.6737	0.366	1030.9	0.7923	0.571	1063.1	0.7611	0.575	1042.4
Qwen3_14B	0.6684	0.344	1030.9	0.6923	0.365	1020.9	0.7111	0.506	1024.5
Phi4	0.6737	0.344	1030.9	0.7077	0.413	1032.9	0.7056	0.425	1022.5
Gemma3_4b	0.6526	0.323	1021.1	0.7461	0.556	1038.8	0.7444	0.540	1036.4
Qwen3_4B	0.6421	0.280	1019.2	0.7384	0.492	1041.9	0.7444	0.540	1036.4
Qwen3_8B	0.5737	0.129	976.2	0.6308	0.238	958.2	0.6667	0.356	988.9
Llama3.1_8B	0.5211	0.054	970.3	0.5385	0.093	961.2	0.5667	0.138	967.0
Mistral_7B	0.5211	0.032	978.2	0.5308	0.048	964.2	0.5555	0.081	971.0
Llama3.2_3B	0.5158	0.032	964.4	0.5308	0.032	958.2	0.5389	0.069	963.0
Qwen3_1.7B	0.5789	0.151	995.7	0.5615	0.111	958.2	0.5944	0.161	965.0
Qwen3_0.6B	0.5211	0.022	974.2	0.5462	0.064	967.2	0.5555	0.092	969.0
Qwen3_4B_Reasoning	0.6474	0.323	1023.1	0.7000	0.413	1026.9	0.7000	0.437	1020.5
Qwen3_14B_Reasoning	0.6474	0.301	1019.2	0.7077	0.397	1032.9	0.7222	0.471	1030.4
Qwen3_8B_Reasoning	0.5895	0.204	1001.6	0.6923	0.413	1023.9	0.6389	0.333	998.8

Table 1: JudgeBoard Leaderboard Results for Base Language Models. "Overall Accuracy" represents the overall judging accuracy of the judge models; "SW Accuracy" represents the judging accuracy of the judge models when student model have made a wrong answer; "Elo" represents the elo score of the juge model. This comprehensive evaluation compares the performance of large language models (>14B parameters) and small language models (<14B parameters) across three mathematical domains: Algebra, Number Theory, and Counting & Probability.

4 Multi-Agent Judging with Profiling

To enhance the robustness and performance of Small Language Model evaluation, we introduce a multi-agent judge framework in which multiple SLMs independently assess the correctness of candidate answers. Each model, referred to as an agent, operates autonomously and contributes to a collective judgment through structured debate and interaction. This setup allows us not only to mitigate the biases of individual models but also to analyze their evaluative behavior through profiling.

4.1 Agent Configuration and Profiling

Each agent is instantiated from a distinct language model instance, ensuring diversity in reasoning while maintaining a consistent evaluation framework. To promote varied yet comparable analytical behavior, all agents are initialized with a shared baseline reasoning prompt that outlines the general task and expectations. However, each agent is also assigned a unique system prompt, which is referred to as its profile, that guides its reasoning style and evaluative priorities (We provide the details of the prompts in Appendix A). The agent profiles were manually crafted to reflect common reasoning strategies found in human problem-solving.

During evaluation, each agent receives the original question and a candidate answer and is asked to determine whether the answer is correct. In addition to a binary judgment (correct/incorrect), agents are prompted to provide a concise natural language explanation that justifies their decision. These explanations serve two purposes: they make the reasoning process interpretable and they form the basis for subsequent inter-agent debate.

4.2 Interaction and Debate

After the agents get their initial answers, we let them discuss with each other in order to find a better answer. We implement a structured multi-turn debate protocol where in each round, agents are allowed to critique the reasoning of their peers and defend their own positions. Following the debate, each agent is given a final opportunity to revise its judgment and explanation. This post-debate revision phase allows agents to incorporate new insights or correct earlier errors. The final collective decision is then determined through

Model	Algebra		Number Theory		Counting and Probability	
	Overall Accuracy	Accuracy - Student Wrong	Overall Accuracy	Accuracy - Student Wrong	Overall Accuracy	Accuracy - Student Wrong
Multi-Agent Profiled Debate Systems						
Qwen3_14B	0.7789	0.591	0.9385	0.873	0.9056	0.851
Qwen3_8B	0.7632	0.538	0.9	0.81	0.8889	0.751
Qwen3_4B	0.7632	0.527	0.9	0.794	0.8444	0.713
Gemma3_12b	0.7	0.441	0.7231	0.413	0.7167	0.437
Gemma3_4b	0.6421	0.29	0.7231	0.444	0.7944	0.724
Phi4	0.6421	0.29	0.6692	0.333	0.7278	0.471

Table 2: JudgeBoard Leaderboard Results for Multi-Agent Jury Systems. This table presents the performance of collaborative multi-agent systems that employ profiled debate methodologies with majority voting mechanisms.

Model	Overall Accuracy	SW Accuracy	Elo Score
Large Language Models (>14B)			
Gemma3_27b	0.880	0.430	1021.1
Qwen2.5_72b	0.870	0.500	1038.7
Llama3.3_70b	0.850	0.210	992.0
Llama4_17B	0.850	0.320	997.8
Mixtral_8x22B	0.850	0.210	992.0
Mixtral_8x7B	0.820	0.110	983.1
Small Language Models (<14B)			
Gemma3_12b	0.850	0.360	1015.3
Qwen3_14B	0.870	0.540	1032.8
Phi4	0.850	0.320	1009.5
Gemma3_4b	0.810	0.070	968.7
Qwen3_4B	0.840	0.320	1003.6
Qwen3_8B	0.850	0.460	1027.0
Llama3.1_8B	0.780	0.390	980.4
Llama3.2_3B	0.650	0.360	968.7
Qwen3_1.7B	0.810	0.320	1009.5
Qwen3_0.6B	0.790	0.140	974.5

Table 3: JudgeBoard Leaderboard Results for Base Language Models on ARC-Challenge Dataset. ARC-Challenge contains grade school science questions requiring reasoning.

majority voting across the revised judgments. In cases of a tie, we either defer to a designated tie-breaker agent with a meta-evaluator profile or flag the instance for manual review, depending on the experimental setting.

5 Experimental Setup

5.1 Dataset and Tasks

We experiment on two downstream tasks: math reasoning and science reasoning. All experiments were tested on publicly available datasets. For the math reasoning task, we use four datasets: GSM8K (Cobbe et al. 2021), which contains basic arithmetic questions; GSM-PLUS (Li et al. 2024), which contains adversarial arithmetic questions; MATH (Hendrycks et al. 2021), which contains high-school level competition questions; and OmniMATH (Gao et al. 2024), which contains olympiad-level questions. For the science reasoning task, we use two datasets: ARC-Challenge (Clark et al. 2018), which contains basic science questions; and GPQA (Rein et al. 2024), which contains undergraduate-level physics, biology, and chemistry questions. For each and every dataset, we construct the leaderboard based on its pre-defined categories. For GSM8K, GSM-PLUS, ARC-Challenge, and MATH, we randomly select 300 questions for the student model to answer, and extract out the same number of correct and wrongly answered questions. For GPQA and OmniMATH, we randomly select 100 questions from each category for the student model to answer. Due to space constraints, we present the leaderboard on three subcategories of the MATH dataset in our main content, and the rest of the leaderboard in the appendix, which can be found in our Arxiv version.

5.2 Model and Implementation

For the student model, we use the GPT-series models for generating candidate answers. For easier datasets like GSM8K, we use the GPT-3.5 model (Brown et al. 2020); for hard datasets like MATH and OmniMATH, we prompt the GPT4o-mini model (OpenAI 2023). We use direct API calling when prompting the GPT-3.5 and GPT4o-mini models. For the judge models, we experiment with all open-source SLM models, including Qwen3 series (Yang et al. 2025), LLama3 series (Dubey et al. 2024), Mixtral series (Jiang et al. 2023), Gemma3 series (Team et al. 2024), and Phi4 series (Abdin et al. 2024). All experiments using the open-sourced models were run on 8 NVIDIA-A40 GPUs with 48GB memory each. with For all models, we set the temperature to 0.5 and the top-k to 1.0.

5.3 Evaluation Metrics

For the JudgeBoard leaderboard, model performance is measured by four metrics: Overall Accuracy, which is the overall judging accuracy of the judge model; Student Wrong Accuracy, which is the judging accuracy when the student model makes a wrong answer; and Elo score, which is introduced in Section 3.3.

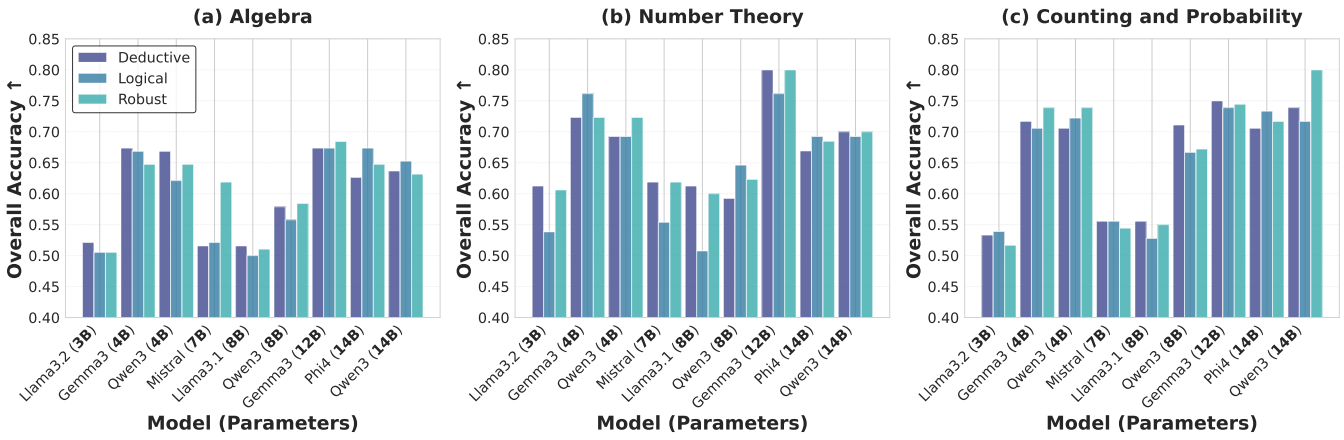

Figure 3: Overall accuracy for using the **Deductive Reasoner (DR), Logical Reasoner (LR),** and **Robust Reasoner (RR)** profiles on (a) Algebra, (b) Number Theory, and (c) Counting and Probability tasks.

6 Results

6.1 Main Results

We present the results of the JudgeBoard leaderboard on the MATH dataset across three categories in Table 1, and on the ARC dataset in Table 3. Due to space constraints, full result tables on other datasets can be found in our Arxiv version. The results for the MAJ framework is presented in Table 2. Due to space constraints, the rest of the results are presented in the Technical Appendix.

Large vs. Small Models Large Language Models (LLMs), which have a size of greater than 14B, generally outperform small ones across all metrics. However, the newer Small Language Models (SLMs), such as the smaller versions of Gemma3 and Qwen3, outperform the older LLMs, such as Llama3 and Qwen 2.5. The performance of Mixtral models, which is a Mixture-of-Expert-based LLM, is very unsatisfactory and barely outperforms its basic version, which is Mistral-7B.

Performance on Wrong Candidate Answers On the MATH dataset, we find that even the worst performing SLMs have an overall accuracy of over 50%. In other words, for models that are performing poorly overall, we find that they are much more likely to agree with what the student model is saying, regardless of their answer correctness. This phenomenon can be observed across datasets: On more complex datasets such as GPQA and OmniMATH, we find that better-performing models tend to have a more balanced judging accuracy, which means they are able to tell whether the potential answer is correct or wrong at a similar level. On simpler datasets like GSM8K and ARC, the worst performing models would have very high model correct accuracy and very low model wrong accuracy. This finding highlights the importance of using multiple evaluation metrics when evaluating the judging ability of the models, as metrics like overall accuracy cannot accurately reflect the true judging ability of the models.

Subject-Wise Performance All models tend to perform better at judging Number Theory questions and Counting and Probability questions, and worse at Algebra questions. This phenomenon suggests that the current Language Models are more adapted to handling arithmetic and modular reasoning, and are inept at symbolic manipulation capabilities.

Multi-Agent Judge (MAJ) Framework We present the results of our MAJ framework in Table 2. We set up different profiles on the same model family and do a multi-agent debate among the profiled agents. The results on the Qwen3 model families demonstrate significant potential of our MAJ framework on SLMs. With the help of MAJ, smaller-sized Qwen3 models (4b, 8b, and 14b versions) could perform comparatively well or even better than the larger-sized Qwen3 models (30B and 32B versions) across all tasks and datasets. Our MAJ framework using Qwen3-14B model as the backbone could outperform the best performing Qwen3-30B-A3B model by an average of 2% in judging accuracy across different categories. Also, larger-sized Qwen3 models demonstrate better debate and interaction ability. Even though the judging performance of Qwen3-8B model is significantly below that of Qwen3-4B model, the MAJ performance when using Qwen3-8B model still outperforms that of using Qwen3-4B model. However, the performance of MAJ on Phi4 and Gemma3 is inconsistent across categories, signaling that these two models are more sensitive to prompts when engaged in debating.

6.2 Ablation Studies

Reasoning and Non-Reasoning Variants For the Qwen3 model families, we conduct additional experiments on both the reasoning and non-reasoning modes of the models, and present our results in Table 1. Using the reasoning mode of the Qwen3 model families does not yield consistent gains across different categories. For the Qwen3-4B version, reasoning mode works best on the Algebra category but drags down the performance in other categories; For the Qwen3-8B and the Qwen3-14B versions, reasoning mode works

Model	Algebra	NumTheo	C&P
Deductive Reasoning (DR) Judges			
Gemma3_12b	1029.0	1066.2	1038.4
Gemma3_4b	1030.9	1032.9	1026.4
Llama3.1_8B	970.3	964.2	967.0
Llama3.2_3B	964.4	949.1	951.0
Mistral_7B	974.2	970.2	969.0
Phi4	1015.3	1014.9	1024.5
Qwen3_14B	1013.3	1020.9	1026.4
Qwen3_4B	1029.0	1020.9	1020.5
Qwen3_8B	968.3	933.8	976.6
Logical Reasoning (LR) Judges			
Gemma3_12b	1027.0	1047.9	1032.4
Gemma3_4b	1011.4	1038.8	1018.5
Llama3.1_8B	853.1	841.4	835.4
Llama3.2_3B	966.4	964.2	963.0
Mistral_7B	972.3	970.2	969.0
Phi4	1023.1	1009.0	1030.4
Qwen3_14B	1017.2	1012.0	1020.5
Qwen3_4B	980.1	997.1	1014.6
Qwen3_8B	942.6	930.7	949.0
Robust Reasoning (RR) Judges			
Gemma3_12b	1029.0	1063.1	1032.4
Gemma3_4b	993.8	1006.0	1024.5
Llama3.1_8B	962.4	952.1	961.0
Llama3.2_3B	946.6	943.0	932.7
Mistral_7B	974.2	967.2	967.0
Phi4	1021.1	1026.9	1026.4
Qwen3_14B	1050.4	1014.9	1032.4
Qwen3_4B	1013.3	1020.9	1032.4
Qwen3_8B	950.5	933.8	963.0

Table 4: JudgeBoard Leaderboard of models prompted by different profiles, in terms of Elo Scores; Numtheo represent the Number Theory category, and C&P represent the Counting and Probability category.

best on the Number Theory category and drags down the performance in other categories.

Judging Ability and Problem-Solving Ability We conduct an ablation study comparing the Problem-Solving ability and judging ability of the models and present the results in Figure 4. The results on the MATH dataset are taken directly from the models' technical reports. For the Qwen3 mode families and the Llama4 model, their judging ability is better than their problem-solving ability. The other models, including Gemma3 model families, Phi4, and LLama3-70B, have slightly worse judging ability compared with their problem-solving ability. From the graph, we can see that the Judging Ability and Problem-Solving Ability of the models are positively correlated. However, two prominent outliers exist: the Qwen3-30b-A3B model and the Qwen3-32B model demonstrate very good judging ability, but only average problem-solving ability.

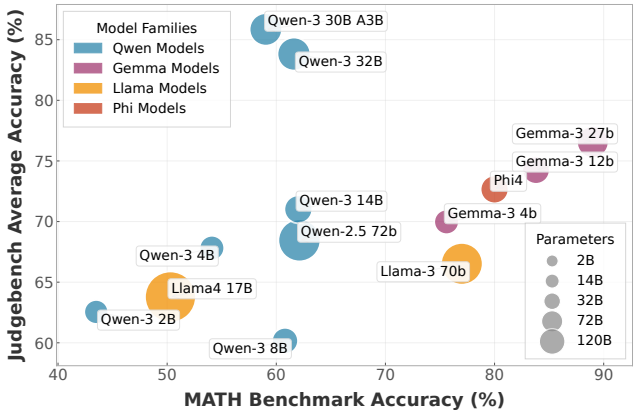

Figure 4: Performance comparison of large language models on mathematical reasoning benchmarks. Each point represents a model, with circle size proportional to parameter count (2B-120B parameters).

Effects of Profiling We present the results of profiling in terms of accuracy in different models in Figure 4 and full results in our Arxiv version. In general, adding profiles to the agents would bring much more versatility compared with the base version of the model. Out of the three manually-curated profiles, Robust Reasoner yields the most consistent gains compared with the base prompt across all categories; Logical Reasoner is the most inconsistent and would rarely give performance gains across categories. In terms of Elo scores, one outlier is the Qwen3-8B model, where it suffers from a significant decrease across all profiles and categories. Qwen3-4B and Qwen3-14B models, even though from the same model family, do not demonstrate this phenomenon.

7 Conclusion and Future Work

In this paper, we introduce JudgeBoard, a novel evaluation pipeline that enables language models to serve as direct judges, moving beyond comparison-based methods. By constructing task-specific leaderboards using both accuracy and Elo-style metrics, we provide a comprehensive framework for assessing evaluative capabilities across mathematical and science reasoning domains. Our leaderboards reveal a significant performance gap between Small Language Models (SLMs) and Large Language Models (LLMs) in judgment tasks, while highlighting the importance of using multiple evaluation metrics to identify systematic biases. To address SLM limitations, we propose the Multi-Agent Judging (MAJ) framework, demonstrating that collaborative reasoning among SLMs substantially closes the performance gap and, in some cases, even surpasses state-of-the-art LLMs. Future work could explore several promising directions: extending MAJ to broader reasoning domains; investigating adaptive agent profiling strategies that dynamically adjust reasoning styles; and exploring the integration of MAJ with fine-tuning approaches to create specialized judge models.

Ethics Statement

This research adhered to the ethical standards and best practices outlined in the AAAI Ethics Code. Language Models can sometimes produce illogical or inaccurate reasoning paths, so their outputs should be cautiously used. The outputs are only examined to understand how a model arrives at its answers and investigate why it makes certain errors. All experiments used publicly available datasets from previous works and did not involve ethical or privacy issues.

Acknowledgements

This work is sponsored by NSF #2442253, NAIRR Pilot with PSC Neocortex and NCSA Delta, Commonwealth Cyber Initiative, Children's National Hospital, Fralin Biomedical Research Institute (Virginia Tech), Sanghani Center for AI and Data Analytics (Virginia Tech), Virginia Tech Innovation Campus, and generous gifts from Nivida, Cisco, and the Amazon + Virginia Tech Center for Efficient and Robust Machine Learning.

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bansal, H.; Gopalakrishnan, K.; Dingliwal, S.; Bodapati, S.; Kirchhoff, K.; and Roth, D. 2024. Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale. *arXiv preprint arXiv:2212.09095*.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 17682–17690.

Bi, Z.; Hajialigol, D.; Sun, Z.; Hao, J.; and Wang, X. 2024. STOC-TOT: Stochastic Tree-of-Thought with Constrained Decoding for Complex Reasoning in Multi-Hop Question Answering. arXiv:2407.03687.

Bi, Z.; Lu, M.; Li, Y.; Roy, S.; Guan, W.; Ziyadi, M.; and Wang, X. 2025. OPTAGENT: Optimizing Multi-Agent LLM Interactions Through Verbal Reinforcement Learning for Enhanced Reasoning. arXiv:2510.18032.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; teusz Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *ArXiv*, abs/2308.07201.

Chen, G.; Dong, S.; Shu, Y.; Zhang, G.; Sesay, J.; Karlsson, B. F.; Fu, J.; and Shi, Y. 2023a. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*.

Chen, J. C.-Y.; Saha, S.; and Bansal, M. 2023. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. *ArXiv*, abs/2309.13007.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2023b. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. arXiv:2211.12588.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Gao, B.; Song, F.; Yang, Z.; Cai, Z.; Miao, Y.; Dong, Q.; Li, L.; Ma, C.; Chen, L.; Xu, R.; et al. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Del Giorno, A.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Li, Q.; Cui, L.; Zhao, X.; Kong, L.; and Bi, W. 2024. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. *arXiv preprint arXiv:2402.19255*.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2024. Encouraging Divergent

Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, W.; Zeng, W.; He, K.; Jiang, Y.; and He, J. 2024b. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Lu, M.; Ho, B.; Ren, D.; and Wang, X. 2024. TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5747–5764. Miami, Florida, USA: Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report.

Pezeshkpour, P.; Kandogan, E.; Bhutani, N.; Rahman, S.; Mitchell, T.; and Hruschka, E. R. 2024. Reasoning Capacity in Multi-Agent Systems: Limitations, Challenges and Human-Centered Solutions. *ArXiv*, abs/2402.01108.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Sel, B.; Al-Tawaha, A.; Khattar, V.; Jia, R.; and Jin, M. 2024. Algorithm of Thoughts: Enhancing Exploration of Ideas in Large Language Models. arXiv:2308.10379.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Zhang, M.; Li, Y.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Srivastava, G.; Bi, Z.; Lu, M.; and Wang, X. 2025a. DEBATE, TRAIN, EVOLVE: Self-Evolution of Language Model Reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32752–32798. Suzhou, China: Association for Computational Linguistics.

Srivastava, G.; Cao, S.; and Wang, X. 2025. ThinkSLM: Towards Reasoning in Small Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32600–32650. Suzhou, China: Association for Computational Linguistics.

Srivastava, G.; Hussain, A.; Bi, Z.; Roy, S.; Pitre, P.; Lu, M.; Ziyadi, M.; and Wang, X. 2025b. BeyondBench: Benchmark-Free Evaluation of Reasoning in Language Models. arXiv:2509.24210.

Srivastava, G.; Hussain, A.; Srinivasan, S.; and Wang, X. 2025c. Do LLMs Overthink Basic Math Reasoning? Benchmarking the Accuracy-Efficiency Tradeoff in Language Models. arXiv:2507.04023.

Tan, S.; Zhuang, S.; Montgomery, K.; Tang, W. Y.; Cuadron, A.; Wang, C.; Popa, R. A.; and Stoica, I. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Tang, X.; Duan, X.; and Cai, Z. G. 2024. Large Language Models for Automated Literature Review: An Evaluation of Reference Generation, Abstract Writing, and Review Composition. *arXiv preprint arXiv:2412.13612*.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; ran Yang, H.; Zhang, J.; Chen, Z.-Y.; Tang, J.; Chen, X.; Lin, Y.; Zhao, W. X.; Wei, Z.; and rong Wen, J. 2023. A Survey on Large Language Model based Autonomous Agents. *ArXiv*, abs/2308.11432.

Wang, Q.; Wang, Z.; Su, Y.; Tong, H.; and Song, Y. 2024. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? In *Annual Meeting of the Association for Computational Linguistics*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; Awadallah, A. H.; White, R. W.; Burger, D.; and Wang, C. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.

Zhuge, M.; Wang, W.; Kirsch, L.; Faccio, F.; Khizbullin, D.; and Schmidhuber, J. 2024. Language Agents as Optimizable Graphs. *ArXiv*, abs/2402.16823.