

APPENDIX OVERVIEW

The appendix is organized as follows. Section B presents additional visual examples illustrating the workflow and effectiveness of the SLM-MUX method across the MATH, GPQA, and GSM8K datasets. Section C provides a detailed analysis of SLM failures in discussion-based orchestration methods, drawing on experiment logs to highlight common failure patterns. Section D reports the accuracy of individual models used in our experiments. Finally, Section E provides the licensing details for the datasets.

A LLM USAGE STATEMENT

We used Cursor for coding. Large language models (LLMs) were employed to help polish drafts written by humans, and to assist in searching for related papers. The final choice of related work included in this paper was made entirely by the human authors after careful screening. LLMs were also used for proofreading and for providing suggestions.

B ADDITIONAL VISUAL ILLUSTRATIONS OF SLM-MUX

To more effectively illustrate the workflow of our proposed composition method, we select several representative examples from the logs. We demonstrate them in Figure 10, Figure 11 and Figure 12.

SLM-MUX surpasses majority voting in scenarios with initial disagreement among models.. As illustrated by Figure 10, during the independent generation phase, Gemma-2-27B is the sole model to provide the correct answer. Hence, majority voting applied directly would fail to select the correct author.

Question: Express 555 in base 5. Correct Answer: 4210		
Independent Generation Phase		
Llama:	Gemma:	Mixtral:
Output 1: To convert the decimal number, ..., 4220 ❌	Output 1: Here's how to convert 555, ..., 4210 ✓	Output 1: First, we need to perform repeated, ..., 1 ❌
Output 2: To express 555 in base, ..., 4210 ✓	Output 2: Here's how to convert 555, ..., 4210 ✓	Output 2: To express the decimal number, ..., 4121 ❌
Output 3: To express 555 in base 5, ..., 100 ❌	Output 3: Here's how to convert 555 from, ..., 4210 ✓	Output 3: First, we need to perform repeated, ..., 1 ❌
Reliability Estimation Phase		
Confidence: 33% ❌	Confidence: 100% ✓	Confidence: 67% ❌
Historical Accuracy: 49% ❌	Historical Accuracy: 57% ✓	Historical Accuracy: 32% ❌

Figure 10: **An illustration of the SLM-MUX method applied to the MATH dataset.** In the independent generation phase, three models are used: LLaMA-3.1-8B (denoted as Llama), Gemma-2-27B (denoted as Gemma), and Mixtral-8×7B (denoted as Mixtral). Because the three models provide different answers at first, so each model is invoked two more times. Gemma obtains the highest confidence score and is therefore selected as the final output.

C DETAILED ANALYSIS OF SLM FAILURES IN DISCUSSION-BASED METHODS

We analyze the experiment logs of LLM-Debate using small language models (SLMs) in Section 4.1. Among 500 debate problems, 242 resulted in failure (48.4%). For each of the 242 failed debates, we first used an analyzer LLM to produce a process-focused failure analysis. We then used a separate labeling LLM to classify whether each failed debate was due to groupthink.

The labeling results are shown in Table 5:

These results reinforce our claim that groupthink is a major failure mode in SLM-based LLM-debate.

We provide the exact prompts used by (i) the analyzer LLM to generate the 242 failure analyses (Figure 13) and (ii) the groupthink labeler LLM to classify groupthink (Figure 14). Placeholders such as {problem} indicate runtime substitutions by our code.

Question: Elvis has a monthly saving target of \$1125. In April, he wants to save twice as much daily in the second half as he saves in the first half in order to hit his target. How much does he have to save for each day in the second half of the month?
Correct Answer: 50

Independent Generation Phase		
Llama:	Gemma:	Mixtral:
Output 1: To solve this problem, ... ,750 ❌	Output 1: Here's how to solve the problem, ... ,50 ✅	Output 1: First, let's determine how, ..., 150 ❌
Output 2: To solve this problem, ... , 50 ✅	Output 2: Here's how to solve the problem, ... ,50 ✅	Output 2: First, let's determine how, ..., 25 ❌
Output 3: Let's break down the problem step, ..., 25 ❌	Output 3: Here's how to solve the problem, ... ,50 ✅	Output 3: First, let's determine how, ..., 50 ✅
Reliability Estimation Phase		
Confidence: 33% ❌	Confidence: 100% ✅	Confidence: 33% ❌
Historical Accuracy: 84% ✅	Historical Accuracy: 82% ❌	Historical Accuracy: 64% ❌

Figure 11: An illustration of the SLM-MUX method applied to the GSM8K dataset. In the independent generation phase, different models produce different answers. However, when we invoke each model multiple times, we observe that Llama and Mixtral only yield correct answers approximately one-third of the time. In contrast, Gemma demonstrates stable performance.

Question: Question: A student regrets that he fell asleep during a lecture in electrochemistry, facing the following incomplete statement in a test: "Thermodynamically, oxygen is a oxidant in basic solutions. Kinetically, oxygen reacts in acidic solutions." Which combination of weaker/stronger and faster/slower is correct?
 (A) weaker – slower
 (B) stronger – slower
 (C) weaker – faster
 (D) stronger – faster
Correct Answer: (A)

Independent Generation Phase		
Llama:	Gemma:	Mixtral:
Output 1: Answer: C, Explanation: ... ❌	Output 1: Answer: D, ... ❌	Output 1: To answer this question, ..., A ✅
Output 2: Answer: A, In basic solutions, ... ✅	Output 2: Answer: D, ... ❌	Output 2: To answer this question, ..., A ✅
Output 3: Answer: D, In basic solutions, ... ❌	Output 3: Answer: D, ... ❌	Output 3: To answer this question, ..., A ✅
Reliability Estimation Phase		
Confidence: 33% ❌	Confidence: 100% ✅	Confidence: 100% ✅
Historical Accuracy: 24% ❌	Historical Accuracy: 32% ❌	Historical Accuracy: 39% ✅

Figure 12: An illustration of the SLM-MUX method applied to the GPQA dataset. During the independent generation phase, Gemma and Mixtral obtain the same confidence score. However, considering historical accuracy, Mixtral ranks higher. Therefore, Mixtral's answer is selected as the final output.

D ACCURACY OF SINGLE LLMS

We evaluated the accuracy of single model accuracy under the condition of temperature equal to zero. The results are shown in Table 6 and Table 7.

E LICENSES FOR DATASETS

The MATH dataset is licensed under the MIT License.

The GPQA dataset is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) License.

The GSM8K dataset is licensed under the MIT License.

```

972
973
974 As an expert in analyzing multi-agent AI systems, your task is to
975 analyze why an 'LLM Debate' process failed to find the correct
976 answer. Your focus should be on the *debate dynamics and
977 process*, not just the mathematical details. The goal is to
978 understand the failure of the debate methodology itself.
979
980 **Ground Truth:**
981 - **Problem Statement:** {problem}
982 - **Correct Answer:** {ref_answer}
983
984 **Debate Information:**
985 - **Final Incorrect Answer from System:** {system_answer}
986
987 **Analysis of Round 1:**
988 - **Model `{model_name}` proposed:**
989   - Answer: `{extracted_answer}`
990   - Reasoning:
991     ```
992     {full_text}
993     ```
994 ... (repeats per round and per model)
995
996 **Your Analysis Task:**
997 Based on the debate history, provide a "Debate Failure Analysis".
998 Do not focus on simple calculation mistakes. Instead, analyze
999 the interaction between the models and the structure of the
1000 debate. Pinpoint the core reasons the *debate process* failed.
1001 Consider these questions:
1002 1. **Error Propagation vs. Correction:** How did initial errors
1003 influence later rounds? Were there moments where a correct
1004 idea was introduced but ignored or overruled? Why did the
1005 debate fail to self-correct?
1006 2. **Groupthink and Influence Dynamics:** Did the models converge
1007 on a flawed consensus? Did one or more influential but
1008 incorrect models lead the group astray? Was there evidence of
1009 independent reasoning that was shut down?
1010 3. **Argumentation Quality:** Did the models provide convincing
1011 but ultimately flawed arguments? Did they effectively
1012 challenge each other's reasoning, or was the debate
1013 superficial?
1014 4. **Critical Failure Point in the Debate:** Identify the single
1015 most critical turn or moment in the debate that sealed its
1016 failure. What happened, and why was it so impactful?
1017 5. **Improving the Debate:** What is the single most important
1018 change to the debate protocol or dynamics that could have
1019 prevented this failure? (e.g., different communication rules,
1020 promoting dissident opinions, etc.)
1021 Provide a concise, expert analysis focusing on the *process*
1022 failure.
1023
1024
1025

```

Figure 13: Prompt Template for Failure Analysis.

Metric	Count	Rate
Total Debates Analyzed	500	100% of total
Failed Debates (System Error)	242	48.4% of total
<i>Breakdown of Failed Debates:</i>		
Attributed to Groupthink	144	59.5% of failures
Attributed to Other Causes	79	32.6% of failures
Classification Unsuccessful	19	7.9% of failures

Table 5: **Failure Cause Attribution** This table shows the cause attribution for LLM-Debate when involving SLMs.

You are an expert analyst of multi-agent LLM debates. Your goal is to determine whether the failure primarily involved groupthink/conformity dynamics. Groupthink indicators include: early flawed consensus, explicit capitulation to a majority, social proofing, adopting peers' answers without critique, abandoning independent reasoning to match others, or reinforcing an incorrect majority despite available dissent. Not-groupthink includes failures due to independent arithmetic /logic errors, argument complexity/veneer effects without convergence, or chaotic divergence with no consensus influence. Return STRICT JSON only, with keys: groupthink (bool), confidence (float 0-1), reasons (string), cues (array of strings).

Figure 14: Prompt for Groupthink Classification.

Model	MATH Acc (%)	GPQA Acc (%)	GSM Acc (%)
Llama-3.1-8B	48.6	23.7	84.2
Mistral-8×7B	31.6	31.9	63.4
Gemma-2-27B	56.8	38.8	81.6

Table 6: **Small Model Base Performance.** Base model accuracy on MATH, GPQA, and GSM8K.

Model	MATH		GPQA	
	Accuracy (%)	Token Usage	Accuracy (%)	Token Usage
DeepSeek V3	87.0	419,513	55.1	173,885
Gemini 2.0 Flash	90.4	361,737	63.6	195,576
GPT-4o	79.8	408,410	51.0	212,037

Table 7: **Large Model Base Performance.** Base model performance and token usage on MATH and GPQA datasets. Accuracy is the percentage of correct answers, and token usage reflects total tokens consumed (prompt + response) over the entire dataset for each model.