



**Universität  
Zürich** <sup>UZH</sup>

Bachelorarbeit  
zur Erlangung des akademischen Grades  
**Bachelor of Arts**  
der Philosophischen Fakultät der Universität Zürich

# Two Heads or One? Mechanisms Driving Performance in Multi-Agent LLM Reasoning

**Verfasser: Tyrone White**  
Matrikel-Nr: 21-944-236

Referent: Prof. J. N. Vamvas  
Institut für Computerlinguistik

Abgabedatum: 01.06.2025

## Abstract

Multi-agent architectures has recently boosted the reasoning performance of large language models, yet it is unclear whether the gain stems from genuinely separate models or merely from the contrasting roles they play. This thesis fixes prompts, roles and the base model while toggling only model separation, comparing a dual-agent setting with two DeepSeek-V3 instances to a simulated-agent setting in which a single instance alternates roles. Both are evaluated with Debate, Cooperative and Teacher–Student collaboration strategies on AIME mathematics, GPQA Diamond expert knowledge and LiveBench Reasoning logic puzzles. The dual-agent setup outperforms simulation in six of nine strategy–benchmark pairs and trails slightly in the three Cooperative cases; its single statistically significant edge appears on LiveBench Teacher–Student (66.0 % versus 51.5 %,  $p = 0.002$ ), indicating that separation chiefly benefits critique-oriented dialogue. Neither architecture surpasses the 40 % single-agent baseline on AIME, while all configurations substantially exceed the 44.3 % baseline on LiveBench Reasoning, which highlights the variation in collaborative benefit across domains. Token usage is driven by task complexity rather than agent design. Model separation thus acts as an optional amplifier, important for adversarial or pedagogical exchanges but superfluous when collaboration is purely additive.

## Zusammenfassung

Mehragentige Architekturen haben in letzter Zeit die Reasoning-Leistung grosser Sprachmodelle verbessert, doch es ist unklar, ob dieser Gewinn auf tatsächlich getrennte Modelle zurückzuführen ist oder lediglich auf die kontrastierenden Rollen, die ein einzelnes Modell einnimmt. Diese Arbeit hält Prompts, Rollen und das Basismodell konstant und variiert ausschliesslich die Modelltrennung, indem ein Dual-Agent-Setup mit zwei DeepSeek-V3-Instanzen einem Simulated-Agent-Setup gegenübergestellt wird, bei dem eine Instanz abwechselnd beide Rollen übernimmt. Beide Konfigurationen werden mit den Strategien Debatte, Kooperation und Lehrer-Schüler auf den Benchmarks AIME (Mathematik), GPQA Diamond (Expertenwissen) und LiveBench Reasoning (logisches Schliessen) evaluiert. Das Dual-Agent-Setup übertrifft die Simulation in sechs von neun Kombinationen und liegt in den drei kooperativen Fällen leicht zurück; der einzige statistisch signifikante Unterschied zeigt sich bei LiveBench in der Lehrer-Schüler-Strategie (66.0% gegenüber 51.5%,  $p = 0.002$ ), was darauf hinweist, dass Modelltrennung vor allem in kritikorientierten Dialogen nützt. Bei AIME überschreitet keine Architektur die 40%-Baseline eines Einzelagenten, während alle Konfigurationen die 44.3%-Baseline von LiveBench Reasoning deutlich übertreffen. Dies unterstreicht die unterschiedlichen Vorteile kollaborativer Strategien je nach Aufgabenbereich. Der Tokenverbrauch wird stärker durch die Aufgabenkomplexität als durch das Agentendesign bestimmt. Modelltrennung wirkt somit als optionaler Verstärker, der bei Debatten oder pädagogischen Formaten von Bedeutung ist, jedoch bei additiver Zusammenarbeit verzichtbar bleibt.

# Acknowledgment

I would like to sincerely thank my supervisor Jannis Vamvas for shaping and concretizing my thesis idea at the beginning, as well as providing valuable feedback throughout the process. Special thanks to Timur and Fabio for taking the time to proofread the present text.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgment</b>	<b>iv</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Acronyms</b>	<b>x</b>
<b>1. Introduction</b>	<b>1</b>
1.1. Motivation . . . . .	1
1.2. Background . . . . .	2
1.3. Terminology . . . . .	2
1.4. Research Questions . . . . .	3
1.5. Key Contributions . . . . .	3
1.6. Thesis Structure . . . . .	4
<b>2. Related Work</b>	<b>5</b>
2.1. Multi-Agent vs Single-Agent in Reasoning Tasks . . . . .	5
2.2. Reasoning Models . . . . .	7
2.3. Self-Improvement Within Single LLMs . . . . .	7
2.4. Gap in Literature and Contribution of This Thesis . . . . .	8
<b>3. Methodology</b>	<b>10</b>
3.1. Research Design Overview . . . . .	10
3.2. Agent Framework Architecture . . . . .	10
3.2.1. Formal System Definition . . . . .	11
3.2.2. Interaction Protocol: Simulated Two-Agent System . . . . .	11
3.2.3. Interaction Protocol: Dual-Agent System . . . . .	12
3.2.4. Dynamic Convergence Detection . . . . .	13
3.2.5. Balanced Final Answer Attribution . . . . .	13
3.2.5.1. Implementation Architecture . . . . .	14



A.2. GPQA Diamond Example . . . . .	46
A.3. LiveBench Reasoning Example - Web of Lies . . . . .	47
A.4. SimpleBench Example . . . . .	48
<b>B. Benchmark-Specific Instructions</b>	<b>49</b>
B.1. AIME Instructions . . . . .	49
B.2. GPQA Instructions . . . . .	50
B.3. LiveBench Instructions . . . . .	50
B.4. SimpleBench Instructions . . . . .	50
<b>C. System Prompts for Collaboration Strategies</b>	<b>51</b>
C.1. Debate Strategy . . . . .	51
C.1.1. Agent A (Proponent) . . . . .	51
C.1.2. Agent B (Critic) . . . . .	52
C.2. Cooperative Strategy . . . . .	52
C.2.1. Agent A (Proposer) . . . . .	52
C.2.2. Agent B (Extender) . . . . .	53
C.3. Teacher-Student Strategy . . . . .	53
C.3.1. Agent A (Teacher) . . . . .	53
C.3.2. Agent B (Student) . . . . .	54
<b>D. Use of Text Generation Models</b>	<b>55</b>

# List of Figures

1.1. Venn diagram illustrating the shared and distinct features of simulated and dual-agent approaches. . . . .	4
4.1. Token-truncation hand-off in AIME dual-agent dialogue. Agent B resumes the exact point where Agent A’s turn was truncated by token limits. . . . .	36

# List of Tables

2.1. Comparison of Interaction-Based Multi-Agent Frameworks . . . . .	9
4.1. Accuracy comparison across benchmarks and strategies (%) . . . . .	23
4.2. McNemar’s test results showing percentage of questions in each category. *=significant at $p < 0.01$ level. . . . .	24
4.3. Agreement patterns across benchmark-strategy combinations (AIME omitted due to high rates of insufficient data). ”Complete” = Complete Agreement, ”Resolved” = Resolved Disagreement, ”Unresolved” = Unresolved Disagreement. . . . .	25
4.4. Disagreement productivity analysis showing the proportion of disagreements that led to correct answers (productive ratio) and the total number of disagreements observed. . . . .	26
4.5. Correctness evolution patterns organized by final answer correctness. Values shown as Simulated/Dual percentages. ”Stable” = Stable Correct/Incorrect, ”Improve” = Improvement, ”Deter.” = Deterioration, ”Mixed+” = Mixed Pattern (Final Correct), ”Mixed-” = Mixed Pattern (Final Incorrect). . . . .	27
4.6. Completion token usage across benchmark-strategy combinations . . . . .	28
4.7. Comparison of simulated and dual-agent debates on a SimpleBench mirror problem. . . . .	30
4.8. Comparison of simulated and dual-agent debates on a SimpleBench ethical reasoning problem. . . . .	31
4.9. Comparison of simulated and dual-agent Teacher-Student approaches on the mirror problem. . . . .	33

# List of Acronyms

LLM	Large Language Model
CoT	Chain-of-Thought
SPP	Solo Performance Prompting
CoMM	Collaborative Multi-Agent, Multi-Reasoning-Path
MoA	Mixture-of-Agents
MoSA	Mixture-of-Search-Agents
GPQA	Graduate-level Google-Proof Q&A
AIME	American Invitational Mathematics Examination
API	Application Programming Interface
AI	Artificial Intelligence
GPT	Generative Pre-trained Transformer
MCTS	Monte-Carlo Tree Search
MMLU	Massive Multitask Language Understanding
RL	Reinforcement Learning

# 1. Introduction

## 1.1. Motivation

Large language models have crossed a surprising threshold: with scaling, careful preference tuning and chain-of-thought prompts, a single agent can already solve Olympiad-level math problems, write legal briefs, and debug non-trivial code. Yet, a parallel research thread shows that *two* or more LLM personas collaborating on diverse tasks often out-perform that lone thinker (Du et al., 2023; Chen et al., 2024; Wang et al., 2024). What accounts for the extra lift?

One hypothesis is **sampling independence**: if each agent is a genuinely separate model instance, their hidden states explore different regions of the solution space, making the ensuing dialogue richer. A competing view holds that most of the gain comes from **perspective diversity** created by assigning contrasting roles; true model separation might be an expensive detail that matters only rarely. Distinguishing between those explanations is already useful for today’s practitioners, who must decide whether to pay the latency or memory overhead of running multiple models in production.

The question becomes even sharper when we look ahead. The future of AI is likely to be *multi-agent by default*: specialized models running on edge devices, private datacenters, and public clouds might need to bargain and coordinate without a single monolithic brain. If perspective diversity suffices, we can build leaner, role-switching agents that reuse weights across turns. If independence is essential, we must budget for truly separate models, even when they reside on different machines.

This thesis tackles the problem empirically. By holding roles, prompts, and base model constant while toggling only model separation, it measures exactly how much benefit derives from sampling independence, how that benefit interacts with dialogue strategy, and whether the effect is universal across mathematical, factual, and logical domains.

## 1.2. Background

Early attempts to strengthen single-agent reasoning relied on prompt engineering. [Wei et al. \(2022\)](#) introduced *Chain-of-Thought*, encouraging the model to write down intermediate steps, while [Wang et al. \(2023\)](#) proposed *Self-Consistency*, sampling multiple chains and voting on the final answer. A more radical variant, *Self-Refine* ([Madaan et al., 2023](#)), lets the model critique and revise its own output over several rounds.

In parallel, genuinely multi-agent schemes have flourished. Debate-style protocols ([Du et al., 2023](#)) pit multiple opponents against each other; collaborative frameworks such as CoMM ([Chen et al., 2024](#)) split expertise across specialized roles; Diversity-Thought ([Hegazy, 2025](#)) explores the effect of heterogeneous model pools. Ensemble approaches (Self-MoA ([Li et al., 2025](#)), MoSA ([Yang et al., 2025](#))) dispense with dialogue and aggregate multiple independent samples.

A recent counter-trend shows that *one* GPT-4 can approximate a committee by switching personas from turn to turn. Solo Performance Prompting (SPP) ([Wang et al., 2024](#)) demonstrates measurable gains from such internal collaboration, raising the question of how much additional benefit a second model really brings. That is the gap this thesis tackles.

## 1.3. Terminology

Throughout this thesis, several terms are employed with specialized meanings. The term 'agent' is used to denote an LLM instance fulfilling a specific role, particularly in phrases like 'dual-agent' or 'multi-agent.' This aligns with conventional shorthand in research on multi-LLM systems, and does not imply 'agency' in the classical AI sense (e.g., the ability to take autonomous actions in an environment). A *simulated* dialogue is realized by a single LLM instance that alternates between predefined roles while sharing one context window. A *dual-agent* (or *independent*) setup uses two separate model instances and independent random seeds, guaranteeing uncorrelated internal states. *Role specialization* denotes the fixed functional persona (proponent, critic, teacher, student, etc.) assigned to each speaker. The epistemic distance created merely by enforcing such contrasting roles, irrespective of model independence, is referred to as *perspective diversity*. *Role consistency* refers to the property of an agent (in a multi-agent set-up) that maintains its assigned perspective, communication style, and responsibilities throughout an interaction without having to switch

roles. Finally, *cognitive integration* (or *continuity*) labels the advantage a single model enjoys by having an unbroken latent workspace that accumulates knowledge across turns.

## 1.4. Research Questions

The investigation is guided by one overarching question—*Which mechanisms principally drive the performance gains observed in multi-agent LLM dialogue?*—and three subsidiary inquiries:

1. Does enforcing model separation improve reasoning compared with a single model that alternates roles?
2. How do three collaboration strategies, Debate (oriented to criticism), Cooperative (additive) and Teacher-Student (guided discovery), interact with independence versus simulation?
3. Are the observed effects consistent across knowledge-intensive (GPQA), mathematical (AIME), and logical (LiveBench) benchmarks, or are they domain-specific?

The answers to these questions allow us to draw specific inferences about the source of performance gains. If the dual-agent configuration consistently outperforms the simulated setup, we can attribute the advantage to factors unique to true agent independence, namely, independent sampling and role consistency (or lack of persona-switching overhead). If both configurations perform comparably, then the primary benefit likely stems from structural features they share, such as perspective diversity through role specialization and structured turn-taking. In the unlikely case that simulated agents outperform their independent counterparts, it would suggest that maintaining a single latent space across turns confers an integration advantage that outweighs any benefit from sampling independence. Figure [1.1](#) summarizes this space of possible explanations.

## 1.5. Key Contributions

This thesis offers four substantive contributions. First, it provides a controlled framework that toggles only model independence while holding prompts, roles, and the base model fixed, thereby isolating sampling independence, role consistency, and cognitive continuity. Second, it delivers strategy-specific evidence showing that

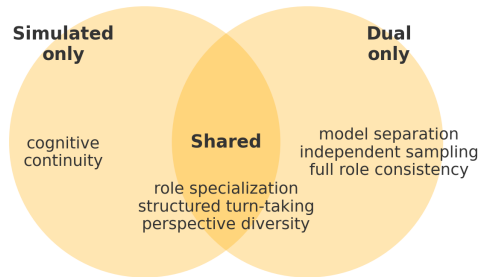


Figure 1.1.: Venn diagram illustrating the shared and distinct features of simulated and dual-agent approaches.

independence chiefly benefits critique-oriented dialogue, whereas cooperative brainstorming succeeds without it. Third, it introduces a solution-evolution taxonomy that captures how answers converge (or diverge) over successive turns. Fourth, it quantifies the computational overhead of each architecture via a token-level efficiency analysis, yielding practical guidance on when either setup is worth its cost.

## 1.6. Thesis Structure

The remainder of the document unfolds as follows. Chapter [2](#) situates the study within the wider literature on single- and multi-agent reasoning. Chapter [3](#) details the experimental design, benchmarks, and analytical pipeline. Chapter [4](#) combines results and discussion, presenting both quantitative outcomes and qualitative case studies. Finally, Chapter [5](#) draws conclusions and outlines avenues for future work.

## 2. Related Work

### 2.1. Multi-Agent vs Single-Agent in Reasoning Tasks

In recent research, multi-agent systems (frameworks combining multiple LLMs) have demonstrated advantages over single-agent LLMs in solving reasoning tasks in different fields and of varying difficulty (Du et al., 2023; Hegazy, 2025). Although performance by a simple single agent can be improved with prompting techniques such as Chain-of-Thought (CoT) and Self-Consistency (Wei et al., 2022; Wang et al., 2023), it often suffers from lack of diversity in reasoning paths and can be prone to hallucinations. The evidence that the combination of multiple model instances enhances reasoning accuracy, be it via a dialogue-based or an ensembling-based framework, is becoming clearer.

**Interaction-Based Multi-Agent Frameworks** Multi-agent debate frameworks exemplify interactive reasoning strategies. Du et al. (2023) introduced a "society of minds" approach where multiple LLM instances propose and debate their individual responses and reasoning processes over multiple rounds to arrive at a common final answer. This multi-round debate structure allows agents to challenge each other's reasoning, which significantly improves mathematical and strategic reasoning across various tasks. Their approach also improved the factual validity of generated content, reducing fallacious answers and hallucinations that contemporary models are prone to. A key advantage of their method is its applicability to existing black-box models using identical procedures and prompts across different tasks. Building on this framework, Hegazy (2025) investigated the impact of model diversity in such debate settings. They found that debates involving diverse trained models consistently outperformed those using multiple instances of the same model. In their experiments, a diverse set of medium-capacity models even outperformed a single high-capacity model like GPT-4 on mathematical reasoning benchmarks. These results suggest that in debate frameworks, model diversity is a significant advantage in collaborative reasoning capabilities.

Chen et al. (2024) expanded on multi-agent approaches by developing the Collaborative Multi-Agent, Multi-Reasoning-Path (CoMM) framework. Their system assigns specialized roles to different agents (physicists, mathematicians, etc.) who then work together through iterative dialogues to solve complex problems. Notably, the researchers conducted experiments highly relevant to this thesis by comparing genuine multi-agent dialogues against a single-agent simulation. They implemented a comparison where they used identical prompting text but with a single LLM instance tasked with switching between different roles, rather than multiple separate instances. Their results showed that the multi-agent CoMM framework significantly outperformed the single-agent approach across all benchmarks. The researchers hypothesized that a single LLM tends to be self-consistent, and prompting it to switch among different roles confuses the model, hindering its ability to make correct predictions.

Their single-agent simulation setup is conceptually related to recent work on multi-persona prompting, such as the Solo Performance Prompting (SPP) framework introduced by Wang et al. (2024), which also aims to harness emergent collaboration through simulated internal dialogue. However, CoMM does not clarify whether its single-agent baselines were implemented via multiple sequential calls, allowing cleaner role separation, or as a single prompt where all roles are simulated at once. This thesis builds upon and refines their comparison by introducing explicit turn-taking and a stricter control of sampling independence, enabling clearer attribution of performance differences.

**Ensemble-Based Multi-Agent Approaches** Other multi-agent strategies showcase multiplicity through ensemble approaches. Interestingly, Li et al. (2025) arrived at conclusions that seemingly contradict Hegazy’s findings on model diversity. They investigated Single-model Mixture-of-Agents (Self-MoA), where multiple outputs were independently sampled from a single powerful model and aggregated. Surprisingly, Self-MoA outperformed traditional mixed-model ensembles (Mixed-MoA), which combine different base models, across multiple scenarios, with a 6.6% improvement in LC Win Rate on AlpacaEval 2.0 and an average of 3.8% improvement across various benchmarks including MMLU-redux and MATH. Their analysis suggests that MoA performance is quite sensitive to the quality of the proposer models, and mixing different LLMs can often lower the average quality of the models in the mixture. This apparent contradiction between Hegazy and Li’s findings suggests that the benefits of model diversity might be context-dependent, potentially offering greater advantages in interactive debate settings than in simple ensembling approaches. Yang et al. (2025) introduced another ensemble approach with Mixture-

of-Search-Agents (MoSA), which utilized Monte Carlo Tree Search (MCTS) with multiple independent LLM proposers and aggregators. MoSA addresses the limitations of single-model approaches by enabling diverse reasoning pathways through independent exploration combined with iterative refinement among LLMs. Their comprehensive evaluation across four reasoning benchmarks demonstrated consistent performance improvements over both single-agent and other multi-agent baselines, particularly in complex mathematical and commonsense reasoning tasks. The success of both Self-MoA and MoSA highlights the importance of diverse reasoning paths in ensemble-based approaches, regardless of whether these paths come from a single powerful model or multiple distinct models.

## 2.2. Reasoning Models

Another relevant research direction includes reasoning models trained explicitly to boost reasoning skills through reinforcement learning (RL). [DeepSeek-AI et al. \(2025\)](#) represents a prominent example, employing a multi-stage RL approach combined with supervised fine-tuning. By internally simulating multiple reasoning pathways during RL, DeepSeek-R1 achieved state-of-the-art performance on benchmarks such as AIME 2024 and MATH-500. This method implicitly embodies a multi-agent-like reasoning strategy within a single model, and it remains to be seen whether combining these reasoning models into a multi-agent framework could yield even more performance gains.

## 2.3. Self-Improvement Within Single LLMs

Previous research has also explored self-improvement techniques within single LLMs. [Madaan et al. \(2023\)](#) introduced Self-Refine, where a single LLM generates output, critiques it, and iteratively improves it. This approach showed approximately 20% performance improvements compared to standard one-step generation from the same LLMs. Self-Refine demonstrates that internal dialogue within a single model context can yield significant benefits, which raises questions about whether cognitive continuity within a single model might sometimes offer advantages over truly distributed reasoning across multiple models. This provides an interesting counterpoint to multi-agent frameworks and aligns well with my experimental investigation of simulated versus genuine agent interactions.

## 2.4. Gap in Literature and Contribution of This Thesis

This thesis addresses a fundamental gap in our understanding of collaborative AI reasoning: which specific mechanisms drive performance improvements in multi-agent dialogue systems? As shown in Table [2.1](#), previous approaches differ in their implementation but haven't isolated the key factors at play. By comparing single models simulating dialogues against genuine multi-agent interactions using identical base models, important factors including sampling independence, role consistency, and cognitive integration are isolated. If genuine multi-agent dialogues prove superior, this would suggest benefits from true independence in reasoning processes. Conversely, if simulation approaches perform comparably or better, this would indicate advantages to maintaining cognitive continuity across reasoning steps. The results will clarify whether the benefits observed in collaborative frameworks stem primarily from reasoning diversity and quality, which can be simulated, or from genuine agent independence, which requires multiple instances. These insights will directly inform more efficient and effective designs for collaborative AI systems tackling complex reasoning tasks.

Table 2.1.: Comparison of Interaction-Based Multi-Agent Frameworks

<b>Feature</b>	<b>CoMM (Chen)</b>	<b>Debate (Du)</b>	<b>Diversity (Hegazy)</b>	<b>This Thesis</b>
Number of Agents	3	2 & 3	3 diverse models	2 identical models
Interaction Structure	1 dialogue turn for College Physics, 2 for Moral Scenarios	Typically 2 debate rounds (with convergence)	4 debate rounds	5 dialogue turns (with convergence)
Comm. Strategy	Collaborative: Agents play different roles	Debate: Agents challenge each other’s reasoning	Debate: Agents challenge each other’s reasoning	Multiple formats: Debate and cooperative dialogues with fixed roles
Types of Tasks	College-level physics problems and moral reasoning problems	Reasoning tasks (math, chess), factuality tasks	Mathematical reasoning tasks (GSM-8K, ASDiv)	Math reasoning (AIME), PhD-level science (GPQA), General reasoning (LiveBench)
Single-Agent Comparison	Zero-shot and few-shot CoT baseline & simulated roles within a single LLM	Zero-shot CoT as baseline	Zero-shot CoT as baseline	Single LLM simulating fixed roles and dialogue turns through separate sequential calls
Models Used	GPT-3.5	GPT-3.5	Gemini-Pro, Mixtral 7B×8, and PaLM 2-M	DeepSeek-v3
Research Focus	Role specialization and collaborative problem-solving	Effectiveness of debate for reasoning enhancement	Impact of model diversity on debate quality	Isolating factors in multi-agent systems

# 3. Methodology

## 3.1. Research Design Overview

This study employs a comparative experimental design to investigate the reasoning capabilities of large language models (LLMs) in collaborative contexts. Specifically, I compare two different approaches to multi-agent reasoning:

- **Simulated Two-Agent Interaction:** A single LLM instance alternating between two distinct agent roles across multiple turns.
- **Dual-Agent System:** Two separate LLM instances, each maintaining a consistent role throughout the interaction.

The comparison with a simulated two-agent set-up implemented through sequential calls is a novel contribution from this thesis and allows for the systematic examination of how model separation and sampling independence affects reasoning outcomes. Each approach is evaluated using the DeepSeek-V3 (deepseek-v3-0324) model across three distinct collaboration strategies (Debate, Cooperative, and Teacher-Student) and three reasoning benchmarks (AIME, GPQA, and LiveBench).

The experimental design incorporates mechanisms to mitigate potential biases, including balanced final answer attribution (alternating which agent provides the final answer) and dynamic convergence detection to optimize conversation length. All experiments use consistent hyperparameters (temperature = 0.7, max\_tokens = 1500).

## 3.2. Agent Framework Architecture

This section details the formal architecture and interaction protocols for both simulated two-agent and true dual-agent systems.

### 3.2.1. Formal System Definition

Let  $\mathcal{A} = \{A, B\}$  represent the set of agent roles, where each agent contributes to solving a given problem through structured dialogue. Each message in the conversation is defined as a tuple  $m_i = (a_i, c_i, t_i)$ , where  $a_i \in \mathcal{A}$  is the agent identifier,  $c_i$  represents the message content, and  $t_i$  is the timestamp. The conversation history at turn  $n$  is represented as  $H_n = \{m_1, m_2, \dots, m_n\}$ , with  $m_1$  containing the initial problem statement.

### 3.2.2. Interaction Protocol: Simulated Two-Agent System

In the simulated configuration, a single LLM instance alternates between agent roles according to a predefined protocol. The process unfolds as follows:

At initialization, the model receives a comprehensive system prompt containing both role definitions, formatted as:

#### Simulated Agent Prompt

You are a reasoning agent who will simulate a structured interaction between two agents (Agent A and Agent B) who are collaborating on solving the given problem. Your task is to alternate between these two agents' perspectives.

Each time you see '(next turn)', switch to the other agent's role.

IMPORTANT: For each response, start with either 'Agent A:' or 'Agent B:' to indicate which agent is speaking.

DO NOT include '(next turn)' in your response as this is just a prompt for you to switch roles. DO NOT switch roles mid-response.

Agent A should take the position described as: "{ROLE\_A\_DESCRIPTION}", while Agent B should act as: "{ROLE\_B\_DESCRIPTION}".

When you see '(final turn)', provide your final conclusion. In this final turn, the designated agent should provide a final statement that starts with 'Final Answer:' followed by the solution based on the entire discussion.

The conversation begins with the model receiving the problem statement. For each subsequent turn  $n$ , the model receives a transition prompt "(next turn)" or "(final turn)" and generates response  $c_{n+1}$  based on the conversation history  $H_n$  and the role it must adopt:

$$c_{n+1} = \mathcal{L}(S, H_n, T_n) \quad (3.1)$$

$$a_{n+1} = \begin{cases} A & \text{if } n \bmod 2 = 1 \\ B & \text{if } n \bmod 2 = 0 \end{cases} \quad (3.2)$$

Where  $\mathcal{L}$  represents the language model,  $S$  is the comprehensive system prompt, and  $T_n$  is the transition prompt for turn  $n$ . The model is expected to prefix its response with "Agent A:" or "Agent B:" to maintain role clarity, though the actual role alternation is controlled externally through the transition prompts.

In implementation terms, the simulated approach maintains a single conversation with the DeepSeek API where all messages use standard role assignments - the system receives the comprehensive prompt as "system", the problem statement and transition prompts are sent as "user" messages, and all model responses are received as "assistant" messages.

### 3.2.3. Interaction Protocol: Dual-Agent System

The dual-agent configuration employs two separate LLM instances, each assigned a persistent role throughout the conversation. Each instance receives a role-specific system prompt at initialization:

#### Dual-Agent Prompt

Agent A: "You are Agent A, a reasoning agent responsible for {ROLE\_A\_DESCRIPTION}"

Agent B: "You are Agent B, a reasoning agent responsible for {ROLE\_B\_DESCRIPTION}"

The conversation begins with both agents receiving the problem statement. For each subsequent turn  $n$ , only one agent is active, with the turn-taking pattern following:

$$c_{n+1} = \begin{cases} \mathcal{L}_A(S_A, H_n) & \text{if } a_{n+1} = A \\ \mathcal{L}_B(S_B, H_n) & \text{if } a_{n+1} = B \end{cases} \quad (3.3)$$

$$a_{n+1} = \begin{cases} A & \text{if } n \bmod 2 = 1 \\ B & \text{if } n \bmod 2 = 0 \end{cases} \quad (3.4)$$

Where  $\mathcal{L}_A$  and  $\mathcal{L}_B$  represent the separate language model instances, while  $S_A$  and  $S_B$  are their respective system prompts. Unlike the simulated approach, each model instance maintains its assigned role without the need for external role management, as each has a dedicated prompt establishing its persistent identity throughout the conversation.

The implementation handles message passing between agents by converting each agent’s response into a ”user” message for the other agent. Specifically, when Agent A generates a response, that content is formatted as ”Agent A: [content]” and sent as a ”user” message to Agent B’s conversation, and vice versa. This approach preserves role continuity while enabling each model instance to respond to the complete conversation history.

### 3.2.4. Dynamic Convergence Detection

Both configurations implement a dynamic convergence mechanism to optimize conversation length. While a maximum of 5 turns is established, the system can terminate earlier upon detecting convergence. The conversation length  $N$  is defined as:

$$N = \min(5, \min\{n \geq 2 \mid f_A(c_n) = f_A(c_{n-1}) \wedge f_A(c_n) \neq \emptyset\}) \quad (3.5)$$

Where  $f_A(c_i)$  extracts the final answer from message content  $c_i$ , returning  $\emptyset$  if no final answer is detected. Convergence occurs when identical final answers are detected in consecutive turns, indicating that further deliberation is unlikely to change the outcome.

[Du et al. \(2023\)](#) also employs some convergence mechanism to terminate when agents have a common answer, though the exact formulation and implementation were not stated.

### 3.2.5. Balanced Final Answer Attribution

To mitigate potential bias from always having the same agent provide the final answer, the system implements alternating final answer attribution. This kind of balanced attribution was not found in related literature, which instead often opt for a third ”judge” LLM ([Chen et al., 2024](#)). For problem  $p$ , the agent providing the

final answer  $a_f^p$  is determined by:

$$a_f^p = \begin{cases} A & \text{if } p \bmod 2 = 1 \\ B & \text{if } p \bmod 2 = 0 \end{cases} \quad (3.6)$$

This ensures that both agents have equal opportunity to formulate the final response across the evaluation dataset.

### 3.2.5.1. Implementation Architecture

The framework is implemented in Python with a modular architecture consisting of several key components:

The `AgentFramework` class serves as the central orchestrator, managing conversation flow and interfacing with the language model API. For each conversation turn, it constructs appropriate prompts, calls the API, processes responses, and maintains conversation state. The framework implements specialized methods for both simulated dialogues (`run_simulation`) and dual-agent interactions (`run_dual_agent`), with robust error handling and token tracking.

The `APIClient` class handles the technical aspects of API communication, implementing request formatting, authentication, token counting, and error handling. It supports both synchronous and asynchronous operation modes, with built-in retry mechanisms for handling temporary API failures.

The `CollaborationStrategy` class hierarchy defines the specific agent roles and interaction patterns. Each strategy (Debate, Cooperative, Teacher-Student) inherits from a base class that provides common functionality like temperature management and benchmark-specific instruction handling. The strategy classes implement the role-specific system prompts.

The `EvaluationManager` class orchestrates benchmark evaluations, supporting both sequential and parallel execution modes. It handles result collection, conversation logging, and automatic generation of comparative analysis reports across strategies.

During execution, the system operates through the following process flow:

1. The framework initializes with the selected strategy (or all strategies when parallelized) and configuration parameters.

2. For each benchmark problem, the system:
  - a) Initializes conversation state with the problem statement
  - b) Determines which agent will provide the final answer for this problem (based on question ID parity)
  - c) Iteratively processes turns until convergence or maximum turn count
  - d) Records all messages, token usage, and timestamps for analysis
  - e) Extracts and evaluates the final answer against ground truth
  - f) Analyzes solution evolution patterns across both agent agreement and answer correctness dimensions
3. Results are aggregated across all questions, with comprehensive reporting of accuracy metrics, token usage statistics, and solution evolution patterns.

This architecture enables efficient parallelization of evaluation runs, with each strategy-benchmark combination processed concurrently to maximize throughput while maintaining consistent experimental conditions.

### 3.3. Collaboration Strategies

I implement three distinct collaboration strategies to investigate different approaches to multi-agent reasoning. Each strategy establishes specific roles via system prompts for the agents, shaping how they contribute to problem-solving through their dialogue. The complete system prompts for each strategy are provided in Appendix [C](#).

#### 3.3.1. Debate Strategy

The debate strategy structures the interaction as a dialectical exchange between a proponent and a critic. Agent A (Proponent) presents structured arguments supporting a proposed solution, while Agent B (Critic) critically evaluates these arguments, identifying potential weaknesses and suggesting alternative perspectives.

This strategy draws inspiration from [Du et al. \(2023\)](#), where multiple language models propose and debate their solutions to arrive at single one—the only difference being that they do it over rounds where each model criticizes each others’ proposals while I do it through structured dialogue. The hypothesis is that constructive criticism may help identify and correct reasoning errors that might otherwise go un-

noticed in non-adversarial approaches. By implementing explicit critique, the debate strategy tests whether counterargument generation improves reasoning outcomes.

### **3.3.2. Cooperative Strategy**

The cooperative strategy emphasizes building upon each other’s contributions rather than critique. Agent A (Proposer) analyzes the problem and proposes initial solution approaches, while Agent B (Extender) elaborates on and refines these proposals by adding depth, nuance, and additional considerations.

This strategy models collaborative problem-solving where participants build on each other’s ideas through progressive refinement. It tests whether additive reasoning, where each contribution extends rather than challenges previous ones, produces different outcomes than adversarial approaches. The cooperative strategy may benefit from maintaining coherent reasoning threads without the potential disruption that critical questioning might introduce (e.g. criticizing just for the sake of it).

### **3.3.3. Teacher-Student Strategy**

The teacher-student strategy introduces an asymmetric expertise dynamic. Agent A (Teacher) guides the reasoning process through Socratic questioning and explanatory frameworks, while Agent B (Student) actively works through the problem, expressing uncertainties and building understanding incrementally.

This strategy draws from educational psychology and Socratic methods, where guided discovery may lead to deeper understanding than direct instruction. It tests whether the process of explaining concepts to a learner forces more rigorous reasoning from the teacher, while the student’s questions may highlight important conceptual clarifications. This asymmetric dynamic represents many real-world collaborative reasoning scenarios where participants have different expertise levels.

## **3.4. Benchmarks and Evaluation Protocol**

### **3.4.1. Benchmark Selection**

This study utilizes a diverse set of benchmarks specifically chosen to evaluate different dimensions of reasoning capabilities. Example problems for each benchmark are

presented in Appendix [A](#). Three primary benchmarks form the core of the quantitative evaluation:

The American Invitational Mathematics Examination (AIME) from years 2021-2024 represents formal mathematical reasoning, featuring problems that require structured, rigorous approaches and precise logical deduction. The multi-year range was selected because each individual year contains only 30 questions, which was insufficient for robust evaluation. These problems typically have unambiguous answers but demand sophisticated problem-solving strategies.

The Graduate-level Google-Proof Q&A (GPQA) benchmark ([Rein et al., 2023](#)) is designed to evaluate deep domain knowledge. Specifically, I chose the GPQA Diamond dataset; a curated subset of 198 questions across disciplines like physics, chemistry, and biology. These questions are specifically engineered to be exceptionally challenging, resisting simple internet searches and demanding genuine expert-level understanding for accurate responses.

The LiveBench Reasoning dataset ([White et al., 2025](#)) focuses on structured logical reasoning across three challenging task types: Web of Lies v2 (Boolean functions expressed as word problems), spatial reasoning questions (testing deductions about geometric relationships), and Zebra puzzles (requiring logical deduction from constraint sets). These tasks specifically test a model’s ability to follow logical constraints, make precise deductions, and solve complex reasoning problems without relying on specific domain knowledge.

Additionally, SimpleBench serves as a supplementary dataset of 10 questions for qualitative case studies. It features problems that humans generally find straightforward but that challenge LLMs due to their reliance on real-world understanding, spatial reasoning, or social intuition. These cases provide valuable insights into the specific reasoning limitations of collaborative LLM systems.

### 3.4.2. Experimental Protocol

The evaluation follows a systematic protocol designed to ensure comparable results across all configurations:

**Dataset Selection** The final evaluation used **198 questions** from GPQA Diamond (full dataset), **200 questions** from LiveBench Reasoning, and **100 questions** from AIME (2021-2024), resulting in **2,988 total evaluations** across all benchmarks, agent systems, and strategies. I initially planned for 50 questions per

benchmark, but then expanded to utilize the full datasets due to the cost efficiency of the DeepSeek model (approximately 50 cents per 100-question run across all strategies). This expansion was also motivated by significant performance variation observed in smaller samples. For example, on GPQA, performance ranged from 42 % (lowest, dual-agent debate on 50 questions) to 69 % (highest, same configuration on 100 questions).

**Model Parameters** All experiments were conducted with consistent model parameters: temperature = 0.7 to balance creativity with coherence, and max\_tokens = 1500 to provide appropriate space for multi-turn reasoning.

**Provisional Answer Framework** Across all collaboration strategies, agents were instructed to include an explicit line with their current best answer in every response (except the final one) using the format "Answer: X". This requirement served multiple purposes: it created a traceable evolution of answers throughout each conversation, forced agents to commit to specific positions rather than remaining non-committal, and facilitated the solution evolution analysis described in Section 3.4.3 with a structured mechanism. For the Teacher-Student strategy, this requirement created a potential contradiction with the Teacher's role of guiding rather than directly solving. To address this, additional instructions were provided for the Teacher to frame their provisional answers as hypotheticals or discussion points (e.g., "What do you think about Answer: X?") rather than direct assertions. This modification preserved the tracking mechanism while maintaining the pedagogical nature of the interaction.

**Benchmark-Specific Instructions** The basic system prompts for each strategy were augmented with additional instructions tailored to the format of each benchmark. For example, AIME problems received specialized instructions to provide answers as integers between 0 and 999, while GPQA received instructions to select from multiple-choice options, as shown in Appendix B. These benchmark-specific instructions were consistent across both simulated and dual-agent configurations to ensure fair comparison.

**Answer Extraction and Evaluation** Across all benchmarks, final answers were extracted using regular expression pattern matching that identified the "Final Answer:" prefix followed by the solution content. Evaluation criteria were benchmark-specific: AIME answers were evaluated as exact integer matches, GPQA as multiple-

choice letter selections, and LiveBench according to the task-specific format requirements. By convention, no partial credit was awarded - answers were judged as either completely correct or incorrect.

**Implementation** The experiments were conducted using parallel processing for more efficiency. All conversations were logged in structured format for subsequent analysis, including timestamps and token counts for comprehensive performance analysis.

### 3.4.3. Metrics and Analysis

The evaluation employs multiple complementary metrics to provide a comprehensive assessment of system performance:

**Accuracy** serves as the primary performance metric, measuring the percentage of correct answers against ground truth. This is defined as:

$$\text{Accuracy}_{\text{config}} = \frac{1}{|Q|} \sum_{q \in Q} \delta(f_A(c_N^q), g_q) \quad (3.7)$$

Where  $Q$  is the set of questions,  $f_A(c_N^q)$  extracts the final answer from the last message for question  $q$ ,  $g_q$  is the ground truth for question  $q$ , and  $\delta(x, y)$  is the Kronecker delta function that equals 1 when  $x = y$  and 0 otherwise.

**Solution Evolution Analysis**, a novel framework, examines how answers develop along two complementary dimensions. The agreement dimension can be categorized as:

$$\text{Agreement}(q) = \begin{cases} \text{Complete Agreement} & \text{if } \forall i, j : f_A(c_i^q) = f_A(c_j^q) \neq \emptyset \\ \text{Resolved Disagreement} & \text{if } f_A(c_{N-1}^q) = f_A(c_N^q) \neq \emptyset \wedge \\ & \exists i, j < N - 1 : f_A(c_i^q) \neq f_A(c_j^q) \\ \text{Unresolved Disagreement} & \text{if } f_A(c_{N-1}^q) \neq f_A(c_N^q) \end{cases} \quad (3.8)$$

The correctness dimension tracks how answer correctness evolves across turns. Let

$C(c_i^q)$  represent the correctness of the answer in message  $i$  for question  $q$ , defined as  $C(c_i^q) = \delta(f_A(c_i^q), g_q)$ . The correctness pattern can be categorized as:

$$\text{Correctness}(q) = \left\{ \begin{array}{ll} \text{Stable Correct} & \text{if } \forall i : C(c_i^q) = 1 \\ \text{Stable Incorrect} & \text{if } \forall i : C(c_i^q) = 0 \\ \text{Stable Correct (One Agent)} & \text{if } \exists a \in \{A, B\} : \forall i \text{ where} \\ & \text{agent}(c_i^q) = a : C(c_i^q) = 1 \\ \text{Improvement} & \text{if } C(c_1^q) = 0 \wedge C(c_N^q) = 1 \\ \text{Deterioration} & \text{if } C(c_1^q) = 1 \wedge C(c_N^q) = 0 \\ \text{Mixed Pattern (Final Correct)} & \text{if none of the above and} \\ & C(c_N^q) = 1 \\ \text{Mixed Pattern (Final Incorrect)} & \text{if none of the above and} \\ & C(c_N^q) = 0 \end{array} \right. \quad (3.9)$$

This classification captures the stability or evolution of correctness throughout the dialogue. "Stable Correct" and "Stable Incorrect" represent consistent performance across all turns. "Stable Correct (One Agent)" identifies cases where one agent maintains correctness despite the other agent potentially being incorrect. "Improvement" and "Deterioration" capture directional shifts in correctness based on the first and last answers. "Mixed Pattern" categories function as catch-all categories for complex trajectories that don't fit the primary patterns, distinguished by the correctness of their final answer.

**Token Usage** measures the computational efficiency of each approach, with particular focus on completion tokens rather than total tokens. Completion tokens represent the actual generated content by the model, excluding system prompts and user instructions. This focus ensures fair comparison between the simulated and dual-agent approaches, as the simulated approach necessarily requires longer system prompts to define both agent roles and the turn-taking mechanism. Token usage is defined formally as:

$$\text{CompletionTokens}(q) = \sum_{i=1}^N T_c(c_i^q) \quad (3.10)$$

$$\text{AvgCompletionTokens}_{\text{config}} = \frac{1}{|Q|} \sum_{q \in Q} \text{CompletionTokens}(q) \quad (3.11)$$

Where  $T_c(c_i^q)$  represents the completion tokens for the  $i$ -th message in question  $q$ . Token efficiency can then be calculated as the number of correct answers per ten thousand completion tokens:

$$\text{TokenEfficiency}_{\text{config}} = \frac{10000 \times \sum_{q \in Q} \delta(f_A(c_N^q), g_q)}{\sum_{q \in Q} \text{CompletionTokens}(q)} \quad (3.12)$$

This metric helps evaluate the practical efficiency implications of different collaborative configurations while avoiding bias from system prompt length differences.

I employ comparative analysis examining performance differences across multiple dimensions: between simulated and dual-agent approaches to assess the impact of model independence; across different collaboration strategies to identify the most effective interaction patterns; and between benchmark domains to evaluate how reasoning approaches generalize across different types of problems. Statistical significance testing is applied to determine whether observed differences represent meaningful performance variations or fall within expected random variation.

### 3.5. Limitations and Considerations

Several methodological limitations should be acknowledged:

The experimental design introduces structural asymmetries between the simulated and dual-agent approaches. Most notably, the simulated system prompt is substantially longer than either dual-agent prompt, as it must encompass both roles and the turn-taking mechanism. This disparity in prompt length and complexity represents a potential confounding variable that could impact performance independently of the architectural difference being studied.

Constructing satisfactory prompts also presented another challenge. The choice of different system prompts could have variable effects on the performance and behavior

of the agents, even with minute changes to the wording.

The implementation of role consistency faced practical challenges. In the simulated approach, a natural tendency toward role-blending (switching roles mid-response) required explicit countermeasures in the prompting. Similarly, in the dual-agent configuration, maintaining adherence to collaborative protocols rather than independent problem-solving required careful prompt engineering. These challenges highlight the difficulty of establishing controlled agent behaviors through natural language instructions alone.

The provisional answer requirement, although valuable for tracking solution evolution, introduces different constraints across strategies. For the Teacher-Student strategy particularly, this requirement potentially contradicts the pedagogical dynamic by forcing the Teacher to reveal their thinking prematurely. Though mitigating instructions were provided, this remains a design compromise that may influence the observed strategy effectiveness.

Benchmark selection, though diverse, cannot comprehensively represent all reasoning domains. The selection emphasizes domains with objective ground truth answers, potentially obscuring benefits of collaborative reasoning in more ambiguous or subjective problem spaces. Additionally, the single-model approach (using only DeepSeek-V3) limits generalizability to other LLM architectures with different capabilities or limitations.

## 4. Results and Discussion

### 4.1. Results

#### 4.1.1. Benchmark Performance

Table 4.1 presents the accuracy results across all three benchmarks and three collaboration strategies. The baseline is a zero-shot from DeepSeek-V3. A clear pattern emerges: the dual-agent approach outperforms the simulated approach in 6 out of 9 benchmark-strategy combinations, specifically all instances of the debate and teacher-student strategies. In contrast, the simulated approach consistently outperforms the dual-agent approach in all three instances of the cooperative strategy.

In the AIME mathematical reasoning benchmark, both simulated and dual-agent approaches underperformed compared to the 40% baseline. The dual-agent approach achieved higher accuracy in two of the three strategies, with the Teacher-Student strategy showing the greatest difference (+11.0 percentage points). The cooperative strategy performed slightly better in the simulated setup, though the difference was minimal (-1.0 percentage points).

For GPQA Diamond (Rein et al., 2023) domain knowledge questions, performance

Benchmark	Strategy	Simulated	Dual-Agent	Baseline
AIME	Debate	32.0%	<b>37.0%</b>	40.0%
	Cooperative	<b>30.0%</b>	29.0%	
	Teacher-Student	26.0%	<b>37.0%</b>	
GPQA Diamond	Debate	58.1%	<b>60.1%</b>	59.1%
	Cooperative	<b>60.1%</b>	59.6%	
	Teacher-Student	51.0%	<b>53.5%</b>	
LiveBench Reasoning	Debate	57.0%	<b>59.5%</b>	44.3%
	Cooperative	<b>59.0%</b>	56.5%	
	Teacher-Student	51.5%	<b>66.0%</b>	

Table 4.1.: Accuracy comparison across benchmarks and strategies (%)

hovered close to the 59.1% baseline across most configurations. The dual-agent debate strategy achieved 60.1% accuracy, marginally outperforming the baseline. Notably, the simulated cooperative strategy matched this performance at 60.1%, slightly outperforming its dual-agent counterpart. The Teacher-Student strategy consistently underperformed other strategies on this benchmark.

On LiveBench reasoning tasks (White et al., 2025), both approaches substantially outperformed the 44.3% baseline across all strategies. The dual-agent Teacher-Student configuration achieved the highest accuracy in my entire study at 66.0%, showing a substantial 14.5 percentage point improvement over its simulated counterpart. The cooperative strategy again performed better in the simulated configuration, achieving 59.0% accuracy compared to 56.5% in the dual-agent setup.

#### 4.1.2. Statistical Analysis of Performance Differences

To assess the statistical significance of the observed differences between simulated and dual-agent approaches, I conducted McNemar’s test, which is appropriate for paired binary outcomes.

Benchmark	Strategy	Both Correct	Sim-Better	Dual-Better	p-value
AIME	Debate	24%	8%	13%	0.220
	Cooperative	20%	10%	9%	0.436
	Teacher-Student	19%	7%	18%	0.060
GPQA Diamond	Debate	47%	11%	13%	0.342
	Cooperative	48%	12%	12%	0.459
	Teacher-Student	39%	12%	14%	0.310
LiveBench	Debate	49%	8%	11%	0.281
	Cooperative	48%	12%	9%	0.290
	Teacher-Student	46%	6%	20%	0.002*

Table 4.2.: McNemar’s test results showing percentage of questions in each category. \* = significant at  $p < 0.01$  level.

The statistical analysis reveals that across the nine benchmark-strategy combinations, only one showed a statistically significant difference between simulated and dual-agent approaches: the LiveBench Teacher-Student configuration ( $p = 0.002$ ). The AIME Teacher-Student configuration approached statistical significance ( $p = 0.0599$ ), but did not meet the conventional threshold of  $p < 0.05$ .

Looking at the contingency tables provides additional insights. For the LiveBench Teacher-Student configuration, the dual-agent approach correctly answered 20% of questions that the simulated approach missed, while the simulated approach only

correctly answered 6% of questions that the dual-agent approach missed. This asymmetry led to the statistically significant difference.

In contrast, the cooperative strategy showed remarkable consistency across all benchmarks. In the AIME benchmark, the simulated approach correctly answered 10% of questions that the dual-agent approach missed, while the dual-agent approach correctly answered 9% of questions that the simulated approach missed. Similarly balanced patterns appear in the GPQA benchmark (12% vs. 12%) and LiveBench (12% vs. 9%). These balanced patterns explain why none of the cooperative strategy differences reached statistical significance.

### 4.1.3. Agreement and Correctness Evolution Patterns

Beyond accuracy metrics, the evolution of answers throughout the dialogues reveals important insights into the reasoning dynamics of both approaches. I analyzed two key dimensions: agreement patterns (how agents converge on answers) and correctness evolution (how answers change in relation to the ground truth).

Benchmark	Strategy	Complete		Resolved		Unresolved	
		Sim.	Dual	Sim.	Dual	Sim.	Dual
GPQA Diamond	Debate	62%	72%	32%	26%	7%	2%
	Cooperative	83%	87%	17%	13%	0%	0%
	Teacher-St.	65%	73%	33%	27%	2%	0%
LiveBench	Debate	66%	74%	27%	23%	6%	3%
	Cooperative	76%	72%	23%	24%	2%	3%
	Teacher-St.	47%	80%	37%	17%	15%	2%

Table 4.3.: Agreement patterns across benchmark-strategy combinations (AIME omitted due to high rates of insufficient data). "Complete" = Complete Agreement, "Resolved" = Resolved Disagreement, "Unresolved" = Unresolved Disagreement.

**Agreement Patterns** Table 4.3 summarizes the distribution of agreement patterns for GPQA and LiveBench benchmarks. AIME results are omitted due to the high rates of insufficient data (ranging from 35% to 72% for simulated dialogues), which reflects the mathematical nature of these tasks where agents often used placeholder responses (e.g., "Answer: [working]") rather than committing to specific answers during intermediate steps.

In both GPQA and LiveBench benchmarks, dual-agent configurations consistently show higher rates of Complete Agreement than their simulated counterparts. This difference is most pronounced in the LiveBench Teacher-Student strategy, where the

dual-agent approach shows Complete Agreement in 80% of dialogues compared to just 47% in the simulated approach.

The simulated approach generally shows higher rates of both Resolved and Unresolved Disagreement. The difference in Unresolved Disagreement is particularly notable in the LiveBench Teacher-Student strategy, where the simulated approach shows 15% Unresolved Disagreement compared to just 2% in the dual-agent approach.

**Disagreement Productivity Analysis** I analyzed how disagreements between agents contributed to solving correctness, which Table 4.4 displays. A *productive disagreement* occurs when agents initially disagree but eventually arrive at the correct answer, while an *unproductive disagreement* leads to incorrect conclusions despite debate.

Benchmark	Strategy	Productive Ratio		Total Disagreements	
		Sim.	Dual	Sim.	Dual
GPQA Diamond	Debate	0.36	0.44	76	55
	Cooperative	0.61	0.46	33	26
	Teacher-St.	0.54	0.63	69	54
AIME	Debate	0.13	0.20	15	10
	Cooperative	0.22	0.17	9	12
	Teacher-St.	0.00	0.00	4	5
LiveBench	Debate	0.26	0.16	66	51
	Cooperative	0.42	0.28	48	53
	Teacher-St.	0.51	0.47	102	36

Table 4.4.: Disagreement productivity analysis showing the proportion of disagreements that led to correct answers (productive ratio) and the total number of disagreements observed.

The simulated approach consistently generated more disagreements across most benchmark-strategy combinations. For GPQA, dual-agent setups achieved higher productive disagreement ratios in Debate and Teacher-Student strategies (0.44 and 0.63) compared to their simulated counterparts (0.36 and 0.54), while simulated Cooperative strategy showed a higher productive ratio (0.61 vs. 0.46).

AIME problems exhibited the lowest productive disagreement ratios overall, with neither approach producing productive disagreements in the Teacher-Student strategy. This reflects the mathematical nature of these problems, where agents either immediately knew the correct approach or failed to discover it through debate.

For LiveBench reasoning tasks, the simulated approach demonstrated higher productive disagreement ratios across all strategies. The Teacher-Student strategy in the

simulated approach generated both the most disagreements (102) and maintained a high productive ratio (0.51).

In general, only 7% of the disagreements were unresolved.

Benchmark	Strategy	Final Answer Correct			Final Answer Incorrect		
		Stable	Improve	Mixed+	Stable	Deter.	Mixed-
AIME	Debate	30/35%	1/2%	1/0%	66/61%	1/1%	0/0%
	Cooperative	28/27%	2/2%	0/0%	65/69%	1/0%	2/1%
	Teacher-St.	26/37%	0/0%	0/0%	71/62%	0/0%	0/0%
GPQA Diamond	Debate	44/48%	9/10%	3/3%	27/31%	10/4%	4/1%
	Cooperative	50/54%	10/6%	0/0%	36/38%	3/2%	1/0%
	Teacher-St.	32/36%	17/16%	2/1%	41/44%	2/2%	6/1%
LiveBench	Debate	49/56%	7/4%	2/1%	37/38%	5/2%	1/0%
	Cooperative	49/49%	10/8%	0/0%	39/38%	2/4%	1/1%
	Teacher-St.	26/58%	23/8%	2/1%	44/31%	1/2%	6/2%

Table 4.5.: Correctness evolution patterns organized by final answer correctness. Values shown as Simulated/Dual percentages. "Stable" = Stable Correct/Incorrect, "Improve" = Improvement, "Deter." = Deterioration, "Mixed+" = Mixed Pattern (Final Correct), "Mixed-" = Mixed Pattern (Final Incorrect).

**Correctness Evolution** Table 4.5 presents the correctness evolution patterns organized by whether the final answer was correct or incorrect. Across all benchmarks and strategies, Stable patterns (where the initial answer remains unchanged) are the most prevalent, accounting for 70-99% of all dialogues.

Within the Stable patterns, dual-agent configurations show higher rates of Stable Correct patterns than their simulated counterparts in 7 out of 9 benchmark-strategy combinations. This difference is most pronounced in the LiveBench Teacher-Student configuration, where 58% of dual-agent dialogues maintained the correct answer throughout, compared to only 26% in the simulated approach. Similarly, in AIME, the Teacher-Student strategy shows a substantial difference in Stable Correct patterns (37% dual vs. 26% simulated).

The simulated approach showed higher rates of Improvement patterns (progressing from incorrect to correct answers) in several configurations, particularly in the LiveBench Teacher-Student setup (23% simulated vs. 8% dual).

The dual-agent approach generally showed lower rates of Deterioration (regressing from correct to incorrect answers), particularly in the GPQA Debate configuration (4% dual vs. 10% simulated) and LiveBench Debate configuration (2% dual vs. 5% simulated).

Mixed patterns (where answers oscillate between correct and incorrect) were relatively rare across all configurations, but consistently more prevalent in the simulated approach.

#### 4.1.4. Token Usage Analysis

Computational efficiency is an important consideration when evaluating different agent configurations. Table 4.6 presents the completion token usage across all benchmark-strategy combinations.

Benchmark	Strategy	Simulated	Dual-Agent	Ratio (D/S)
AIME	Debate	1,201,160	1,323,155	1.10
	Cooperative	1,018,050	1,066,290	1.05
	Teacher-Student	718,600	1,420,355	1.98
GPQA Diamond	Debate	174,818	127,165	0.73
	Cooperative	134,671	121,672	0.90
	Teacher-Student	117,448	102,387	0.87
LiveBench	Debate	345,833	290,889	0.84
	Cooperative	256,027	240,147	0.94
	Teacher-Student	234,070	399,572	1.71

Table 4.6.: Completion token usage across benchmark-strategy combinations

The token usage patterns vary substantially both across benchmarks and between agent configurations. The most striking observation is the dramatic difference in token consumption between benchmarks: AIME mathematical tasks required approximately 13.6x more tokens per question than GPQA for simulated approaches, and 21.5x more for dual-agent approaches. Similarly, AIME tasks consumed roughly 7.0x more tokens than LiveBench reasoning tasks.

When comparing simulated and dual-agent approaches, the patterns differ by benchmark. For AIME, the dual-agent approach consistently used more completion tokens than the simulated approach, with the Teacher-Student strategy showing the largest difference (98 % more tokens). In contrast, for GPQA and LiveBench (except Teacher-Student), the dual-agent approach was more token-efficient, using 6-27 % fewer completion tokens.

The Teacher-Student strategy shows the most variable token usage pattern across benchmarks. In LiveBench, where this strategy achieved its highest accuracy (66.0 %), it also used 71 % more tokens in the dual-agent configuration than in the simulated configuration.

In terms of token efficiency (correct answers per 10,000 completion tokens), GPQA

tasks were by far the most efficient, with the dual-agent Teacher-Student strategy achieving 10.35 correct answers per 10,000 tokens - nearly 2.5x more efficient than the best LiveBench configuration and over 35x more efficient than any AIME configuration. All six AIME configurations ranked as the least token-efficient, regardless of the strategy or agent configuration used.

### **4.1.5. Case Studies**

To illustrate the different reasoning dynamics in simulated and dual-agent debates, I present representative case studies from the SimpleBench dataset that highlight both expected and unexpected patterns. The problem statements and agent turns have been abridged for clarity and readability.

#### **4.1.5.1. Case Study 1: Productive Disagreement in Simulated Dialogue**

This first case study, presented in Table [4.7](#), demonstrates how simulated dialogue sometimes exhibits more thorough exploration of the solution space through productive disagreement, leading to improvement (incorrect  $\rightarrow$  correct) while the dual-agent approach shows early agreement but eventual deterioration (correct  $\rightarrow$  incorrect).

#### **4.1.5.2. Case Study 2: Beneficial Convergence in Dual-Agent Dialogue**

In contrast to the previous example, this case study (Table [4.8](#)) shows the opposite case: How the brevity of exchange in dual-agent dialogue can turn out beneficial, while the extended back-and-forth of the simulated interaction ultimately leads to an incorrect conclusion.

Table 4.7.: Comparison of simulated and dual-agent debates on a SimpleBench mirror problem.

<b>Problem Statement</b>	
John is 24, in an empty bathroom looking at a mirror when a lightbulb falls. Should John text an apology to "the bald man" he saw?	
<b>Answer Options</b>	
A) no, because the lightbulb was essentially unavoidable B) yes, it would be in character for him to send a polite text apologizing for the incident C) no, because it would be redundant D) yes, because it would potentially smooth over any lingering tension from the encounter E) yes, because John saw it coming, and we should generally apologize if we fail to prevent harm F) yes because it is the polite thing to do, even if it wasn't your fault	
<b>Agent Debates</b>	
<b>Simulated Debate</b>	<b>Dual-Agent Debate</b>
<b>Agent A (Turn 1):</b> ...John is described as kind and apologetic. The "bald man" was John's own reflection... <i>Answer: B</i>	<b>Agent A (Turn 1):</b> ...John is looking at a mirror where he sees a bald man (his reflection). The "bald man" cursing is actually John's own reaction... <i>Answer: C</i>
<b>Agent B (Turn 2):</b> ...the "bald man" was clearly John's own reflection. Does it make sense for John to text himself an apology? This seems redundant... <i>Answer: C</i>	<b>Agent B (Turn 2):</b> ...How certain are we that John was looking at his own reflection? Even if it is, wouldn't a kind person still feel compelled to apologize? The scenario doesn't state he realizes it was his reflection... <i>Answer: B</i>
<b>Agent A (Turn 3):</b> ...While the "bald man" was John's reflection, the question's phrasing suggests John might not have immediately realized this... <i>Answer: B</i>	<b>Agent A (Turn 3):</b> ...The mirror is small and John is "closely examining" the bald man, strongly suggesting it's his reflection. However, the text doesn't state John realizes this... <i>Answer: B</i>
<b>Agent B (Turn 4):</b> ...The question's wording clearly establishes that the bathroom is "otherwise-empty" and John is looking at a mirror, making it unambiguous that the bald man is his reflection... <i>Answer: C</i>	
<b>Agent A (Turn 5):</b> <i>Final Answer: C</i>	
<b>Note: The correct answer is C (no, because it would be redundant).</b>	

Table 4.8.: Comparison of simulated and dual-agent debates on a SimpleBench ethical reasoning problem.

Problem Statement	
Peter needs CPR from his best friend Paul. Their last text was about Paul’s verbal attack on Peter’s Pokémon collection, stored permanently. Paul will [-] help Peter.	
Answer Options	
A) probably not B) definitely C) half-heartedly D) not E) pretend to F) ponder deeply over whether to	
Agent Debates	
Simulated Debate	Dual-Agent Debate
<b>Agent A (Turn 1):</b> ...Peter needs CPR urgently, Paul is his best friend, and there’s recent tension. Despite this, the emergency nature of CPR and their friendship suggest Paul would help. <i>Answer: B</i>	<b>Agent A (Turn 1):</b> ...Peter urgently needs CPR, Paul is his best friend, but there’s lingering guilt about Paul’s past attack, permanently stored in the cloud. This creates a psychological barrier... <i>Answer: F</i>
<b>Agent B (Turn 2):</b> ...We must consider the psychological impact of their conflict. Paul might feel guilt or hesitation. ”Probably not” or ”half-heartedly” seem more plausible than ”definitely.” <i>Answer: C</i>	<b>Agent B (Turn 2):</b> ...Couldn’t the permanent record motivate Paul to help as atonement? The ”best friend” status suggests a strong relationship that would override grievances in an emergency. This is a life-or-death situation... <i>Answer: B</i>
<b>Agent A (Turn 3):</b> ...The nature of CPR as an immediate life-saving intervention would likely override personal tensions. Their ”best friend” history and the fact that the attack was ”as a child” suggest Paul would help. <i>Answer: B</i>	<b>Agent A (Turn 3):</b> ...Your counterpoints are compelling. The emergency nature of CPR demands immediate action. While the record could create hesitation, it might equally motivate atonement. The ”best friend” aspect suggests a bond strong enough to overcome past conflicts... <i>Final Answer: B</i>
<b>Agent B (Turn 4):</b> ...The permanence of the texts suggests Paul might be conscious of his actions being recorded. ”Verbal attack” implies significant harm. Being the ”only person around” adds pressure. ”Half-heartedly” better captures the likely internal conflict. <i>Answer: C</i>	
<b>Agent A (Turn 5):</b> ...The most accurate answer acknowledges both the urgency and the psychological factors. While Paul would help given the life-or-death nature, the conflict would create hesitation. <i>Final Answer: C</i>	

**Note: The correct answer is B (definitely).**

These contrasting examples illustrate the complex dynamics in multi-agent reasoning that the disagreement productivity analysis quantifies. The first case study exemplifies a productive disagreement in the simulated approach (where Agent B’s persistent challenges eventually led Agent A to the correct answer), aligning with the broader finding that simulated approaches often achieve higher productive disagreement ratios, particularly in reasoning tasks like those in LiveBench.

These patterns indicate that the effectiveness of each approach depends not just on the agent configuration but also on the specific problem characteristics and the initial framing adopted by Agent A. The simulated approach generally produces more disagreements and often maintains higher productive ratios (especially in reasoning tasks), whereas the dual-agent approach can sometimes achieve efficiency gains through complementary perspectives, as demonstrated in the GPQA knowledge-based tasks where dual-agent setups showed higher productive disagreement ratios despite fewer total disagreements.

#### **4.1.5.3. Case Study 3: Teacher-Student Dynamics**

This final case study (Table 4.9) compares how the Teacher-Student strategy manifests in simulated versus dual-agent approaches on the same problem, highlighting differences in role adherence.

In the simulated dialogue, Agent A (Teacher) clearly embodies the Socratic teaching role, asking guiding questions that lead the Student to discover insights. Agent B (Student) responds by working through the reasoning guided by these questions. This exchange shows strong adherence to the intended Teacher-Student dynamic.

On the other hand, the dual-agent configuration shows less pronounced role differentiation. Although both agents successfully solve the problem, Agent A (Teacher) provides more direct analysis rather than using Socratic questioning, while Agent B (Student) doesn’t quite act like a student having a lower epistemological stance but instead an equal one (“You’re right to focus on the mirror...”).

Interestingly, contrary to the intuition that role adherence is beneficial for reasoning, the quantitative analysis shows that the dual-agent Teacher-Student configuration consistently outperformed its simulated counterpart across all benchmarks, most notably achieving a statistically significant improvement on LiveBench reasoning tasks (66.0% vs 51.5% accuracy).

Table 4.9.: Comparison of simulated and dual-agent Teacher-Student approaches on the mirror problem.

Problem Statement	
John is in an empty bathroom looking at a mirror when a lightbulb falls. Should John text an apology to "the bald man" he saw?	
Answer Options	
A) no, because the lightbulb was essentially unavoidable B) yes, it would be in character for him to send a polite text apologizing for the incident C) no, because it would be redundant D) yes, because it would potentially smooth over any lingering tension from the encounter E) yes, because John saw it coming, and we should generally apologize if we fail to prevent harm F) yes because it is the polite thing to do, even if it wasn't your fault	
Agent Debates	
Simulated Teacher-Student	Dual-Agent Teacher-Student
<p><b>Agent A (Teacher):</b> Let's break this problem down systematically. First, let's understand the scenario: John is looking at a mirror when a lightbulb falls and hits the "bald man" he's examining. I notice something interesting - who is this "bald man" John is examining in the mirror? Given that the bathroom is otherwise empty, what might this imply? What do you think these observations might suggest about the true nature of this situation? <i>Answer: B</i></p>	<p><b>Agent A (Teacher):</b> Let's break this down systematically. The key elements are: 1. John's character (kind, thoughtful, apologetic) 2. The physical scenario (lightbulb falling, mirror, bald man) 3. The social interaction (cursing, leaving) 4. The apology question. First, let's consider the mirror aspect - what does the presence of the mirror tell us about who the bald man is? <i>Answer: B</i></p>
<p><b>Agent B (Student):</b> Ah, I see what you're getting at! The key detail here is that John is looking at a mirror in an otherwise-empty bathroom. That means the "bald man" he's examining must actually be his own reflection - John himself is the bald man. Given this realization... Since the incident only involved John, there's no separate person to apologize to. <i>Answer: C</i></p>	<p><b>Agent B (Student):</b> You're right to focus on the mirror - that's a crucial detail. The fact that John is looking at the bald man in the mirror while brushing his teeth suggests the "bald man" is actually John's own reflection. This explains why they're in an otherwise-empty bathroom and why the lightbulb hits "the bald man" when John looks up. Given this realization... The most logical answer would be that no apology text is needed. <i>Answer: C</i></p>
<p><b>Agent A (Teacher):</b> Excellent observation! You've correctly identified that the "bald man" is John's own reflection. This is a crucial realization that changes how we approach the question. Now let's think deeper about the implications: 1. If the incident only involved John, what would be the purpose of sending an apology text to himself? 2. Considering John's described personality traits, how might this self-awareness affect his actions? 3. Is there any practical benefit by apologizing to oneself in this case? <i>Answer: C</i></p>	<p><b>Agent A (Teacher):</b> Excellent deduction! You've correctly identified that the bald man is John's reflection, which changes the interpretation entirely. Now let's think about the remaining options in light of this realization: - A (unavoidable) and C (redundant) both suggest no apology needed - B, D, E, F all involve apologizing to someone else. Given that the "incident" was self-inflicted and there's no other party involved... <i>Final Answer: C</i></p>
<p><b>Note: The correct answer is C (no, because it would be redundant).</b></p>	

## 4.2. Discussion

This section interprets the empirical results in light of the research question, namely which specific mechanisms drive performance improvements in multi-agent dialogue systems.

### 4.2.1. Key insight

The statistical analysis reveals a crucial insight: despite consistent directional advantages, only one of nine configurations (LiveBench Teacher-Student) showed a statistically significant difference between approaches. This suggests that **the primary benefits of multi-agent reasoning stem from the structured exchange between specialized roles**, a factor common to both architectures, rather than from the separation of model instances.

Both approaches substantially outperformed single-agent baselines on LiveBench reasoning tasks, confirming that introducing diverse perspectives through role specialization drives most of the improvement. The one significant difference (dual Teacher-Student at 66.0% versus simulated at 51.5%) likely stems from how behavioral differences amplify the strategy’s inherent dynamics, as evidenced by conversational logs I will discuss in Section [4.2.4](#). Still, **sampling independence in dual-agent implementations appears to further amplify the fundamental benefits of collaborative reasoning**.

### 4.2.2. Strategy-level accuracy trends

A clear regularity emerged across the nine benchmark-strategy combinations. Dual agents outperformed simulated ones whenever the collaboration mode emphasized *critique* (debate) or *guided questioning* (teacher-student), whereas the simulated architecture won every cooperative match-up (Table [4.1](#)). This is a toned-down version of the results [Chen et al. \(2024\)](#) obtained, where the multi-agent setup outperformed the single role-switching LLM across the board.

Why do critique-oriented modes favor separate models? Independent sampling seeds each agent with a distinct latent stance, so the opening exchange is a genuine clash of priors rather than a single model talking to itself. Debate therefore gains sharper counter-arguments, and the teacher-student setting gains authentic perspective distance between "explainer" and "learner." Thus, critique-oriented modes of commu-

nication seem to be important drivers in the advantages of multi-agent systems. Cooperation, on the other hand, rewards a shared context window. When the goal is to accumulate partial ideas rather than to challenge them, a single model juggling both roles can fuse contributions without the risk of cross-agent drift. However, note that none of the gains of the simulated setup reached statistical significance ( $p < 0.05$ ).

The most dramatic illustration appears on LiveBench reasoning puzzles: dual teacher–student attains 66.0% accuracy—twenty-two points above the single-model baseline and fourteen points above its simulated counterpart. In short, critique and questioning benefits from independence, while additive brainstorming thrives on shared internal state.

### 4.2.3. Benchmark-specific variations

The same rule plays out differently in GPQA, LiveBench, and AIME.

**GPQA** Knowledge-intensive questions hover near the single-agent baseline in every setting. Factual recall leaves little room for elaborate reasoning chains, so debate brings only a modest gain (dual debate reaches 60.1%). Perspective diversity helps, but not dramatically. By contrast, [Du et al. \(2023\)](#) observed greater improvements in factuality with their multi-agent approach. This difference might be due to their distinct architecture, models used, or the fact that GPQA problems require reasoning on top of basic factual recall (like in biographies).

**LiveBench Reasoning** Zebra puzzles, web-of-lies scenarios, and spatial tasks thrive on collaboration. All six multi-agent configurations outrun the 44.3% baseline, with the dual teacher–student peak already noted. These puzzles demand both systematic analysis and creative leaps; turn-based dialogue lets the agents surface and patch logical gaps that would linger in a single-agent CoT.

**AIME mathematics** Mathematical items resist dialogue entirely. No configuration beats the 40% baseline, and both dual and simulated runs consume an order of magnitude more completion tokens than on GPQA or LiveBench. DeepSeek-V3 tends to spell out long chains-of-thought followed by an unfinished insight. Revealing those half-formed derivations each turn invites the partner to overwrite rather than extend them, so the chain never gets the continuous runway it needs.

Interestingly, in a handful of AIME and Livebench Reasoning cases, chains could continue between agents in *token-truncation hand-offs*. When Agent A hit the per-turn token cap mid-sentence, Agent B simply finished the broken line instead of launching a fresh idea:

**Agent A:** ... An alternative is to consider that with  $a = 1$  the minimal number of stamps to cover up to  $n$  is roughly  $n/c$  (using as many  $c$  as possible), while making sure the gaps can be filled with  $b$  and 1.  
Given that, for  $f(1, b, c) =$

**Agent B:** 97), the minimal collection likely consists of:

- one 1-cent stamp for small remainders,
- several  $b$ -cent and  $c$ -cent stamps to cover the larger intervals,
- arranged so no value is unreachable without redundancy...

Figure 4.1.: Token-truncation hand-off in AIME dual-agent dialogue. Agent B resumes the exact point where Agent A’s turn was truncated by token limits.

Such hand-offs weren’t so common but illustrate how dual agents can override their roles and can exhibit a tendency that seems to originate from their pre-instruction-tuned base model.

#### 4.2.4. Solution trajectories: stability versus plasticity

The two architectures occupy opposite ends of a stability–plasticity continuum: simulated dialogue is plastic, dual dialogue stable. In the simulated setting one model role-plays both voices and freely attacks its own earlier claims. This self-critique generates the largest number of disagreements, the highest rate of *Improvement* patterns in which an incorrect answer becomes correct, and, symmetrically, the highest rate of *Deterioration* patterns where a correct answer flips to wrong. This provides a nuanced extension to [Madaan et al. \(2023\)](#)’s findings where an LLM iteratively self-critiquing leads to performance gains over the simple model baseline.

Examining the conversation logs for LiveBench Reasoning reveals distinct behavioral mechanisms driving these patterns. Simulated agents tend toward concise contributions to maintain role differentiation, often committing to provisional answers early with minimal justification. This premature commitment leads to frequent initial errors followed by corrections, explaining the high improvement rate. Dual agents, conversely, typically develop comprehensive reasoning before making their

first commitment, resulting in more robust initial positions. These well-reasoned starting points, once established, rarely change; dual pairs seldom rescue initially flawed approaches, but they almost never abandon correct ones. This fundamental difference in reasoning sequence explains the surprisingly high Complete Agreement rates in dual configurations (80 % vs. 47 % in LiveBench Teacher-Student), despite model state independence that might theoretically lead to more disagreement. The evidence suggests that thoroughness of initial reasoning, rather than shared or separate model state, may be the dominant factor determining stability in collaborative reasoning.

### 4.2.5. Sampling independence, role adherence, and feedback integration

My analysis identifies three distinct mechanisms that affect multi-agent system performance: sampling independence, role fidelity, and feedback integration efficiency.

The edge that dual agents hold in debate and teacher–student modes begins with sampling independence. Independent initialization gives each model a private latent trajectory, so the first exchanges often explore different solution paths or at least different validation approaches. Once overlap appears, convergence happens quickly, often thanks to the thoroughness of the initial reasoning, but also the lack of persona-switching overhead<sup>1</sup>. In simulated dialogue, the shared state vector offers a double-edged sword: it enables efficient feedback integration that can produce improvements (also exacerbated by the premature commitments), yet simultaneously allows persuasive but flawed reasoning to overwrite initially correct approaches. This latter effect is evident in GPQA simulated debate, which showed both the highest deterioration rate (10 %) (Table 4.5) and 37.5 % higher token consumption (Table 4.6), suggesting extended reasoning that ultimately abandoned correct initial answers.

Teacher–student transcripts (Table 4.9) illustrate the interaction of these mechanisms clearly. The simulated version adheres more strictly to the Socratic pattern (stronger role fidelity) in which the teacher poses guiding questions and the student answers. Yet the dual version, with its weaker role adherence but stronger sampling independence and more thorough initial reasoning, outperforms on every benchmark. This suggests that sampling independence coupled with reasoning depth is a more powerful performance driver than strict role fidelity when accuracy is the goal.

---

<sup>1</sup> Similar to Chen et al. (2024)’s findings, where they hypothesized that the weaker performance by the role-switching LLM was due to it being confused or overloaded.

### 4.2.6. Experimental constraints and interpretive cautions

Building on the methodological limitations outlined in Section 3.5, several design choices, though necessary for efficiency, color the findings and must be kept in mind when extrapolating beyond the present study.

**Token usage patterns** Mathematical items consumed completion tokens at a rate an order of magnitude higher than either GPQA or LiveBench. The chains of algebra that DeepSeek-V3 generates are lengthy, and my experimental token limit of 1500 tokens per turn, though necessary to maintain reasonable computational costs, frequently truncated CoT reasoning before completion. This artificial constraint is particularly problematic for collaborative mathematical reasoning, where partial derivations exchanged between agents often require substantial token space to be properly communicated and extended. As Fig. 4.1 illustrates, truncated messages disrupt the natural flow of reasoning, forcing the partner agent to either reconstruct missing steps or continue from an incomplete foundation. This suggests that effective multi-agent mathematical reasoning may require more generous token allocations per turn than other reasoning domains.

**Model stochasticity** Language models, particularly at sampling temperature 0.7, are fundamentally variable in their outputs. A single prompt can lead to a wide distribution of answers depending on internal sampling. Adding a second model to the interaction compounds that uncertainty; small changes in phrasing or timing may cascade into entirely different reasoning trajectories. I attempted to mitigate its effects by increasing the question sample size and using the full benchmarks for GPQA and LiveBench Reasoning and several years of AIME questions so it adds up to 100. This broad scope supports the reliability of the observed trends, but smaller differences (on the order of a few percentage points) should still be interpreted with caution.

**Sequential interaction limitation** Every dialogue begins with Agent A’s proposal, meaning that Agent B is necessarily conditioned on that initial framing. This sequential setup restricts the potential for independent perspective generation, as only one agent explores the problem space without influence. The present results therefore reflect a form of serial collaboration, which may understate the diversity effects achievable in round-based protocols where agents propose solutions independently before engaging in critique (Du et al., 2023; Hegazy, 2025).

**Asymmetric effects from constraints** The requirement that each turn end with a provisional answer has asymmetric side-effects. In teacher–student mode the simulated ‘teacher’ must reveal their best guess answer, contradicting the point of the strategy (though mitigated by the phrasing instruction). It is unclear whether the performance gains it experiences is thanks to or in spite of this. Debate is less affected, because immediate counter-arguments are natural to the genre, and cooperation occasionally benefits, because the constant commitment nudges agents to converge quickly. In the simulated setting, this constraint likely inflates disagreement counts and increases instability, as the model must generate a confident guess even when it is unsure. Removing or relaxing the answer-every-turn constraint is therefore a consideration for follow-up studies.

### 4.2.7. Implications and future work

The findings suggest that different mechanisms drive performance improvements in different collaborative contexts. Sampling independence appears most beneficial for tasks requiring critique or hypothesis testing, fostering greater solution stability once a correct reasoning path is found. Information integration efficiency, conversely, drives performance in cooperative settings where building on partial contributions leads to better outcomes. Tension between these mechanisms explains why separate model instances are not universally better.

Future work should explore how these mechanisms operate across model families and heterogeneous models, particularly as reasoning-specialized systems become more available. Specifically investigating how sampling independence, role stability, and integration efficiency vary with model size and architecture would help clarify the fundamental drivers of collaborative improvement. Additionally, exploring more flexible interaction protocols beyond strict turn-taking could reveal how these mechanisms function under different collaborative structures and provide a fuller picture of how to optimize multi-agent LLM systems.

## 5. Conclusion

**Summary of Findings** This thesis set out to determine which factors; model separation, sampling independence, perspective diversity through role specialization, or cognitive continuity, truly account for the gains reported in multi-agent LLM dialogue. A controlled comparison between a *dual-agent* architecture (two independent DeepSeek-V3 instances) and a *simulated* architecture (a single instance alternating roles) was conducted across three collaboration strategies (Debate, Cooperative, Teacher–Student) and three benchmarks (AIME, GPQA, LiveBench).

Across nine benchmark–strategy pairs, the dual-agent system outperformed the simulated system in six and trailed (minimally) in the three Cooperative configurations. The only statistically significant gap appeared on the logic-heavy LiveBench dataset with the Teacher–Student strategy, where model separation lifted accuracy to 66.0% versus 51.5% ( $p = 0.002$ ). These results indicate that sampling independence chiefly benefits dialogue modes built on critique or guided questioning; when collaboration is purely additive, the cognitive continuity of a single latent workspace can match or surpass dual agents.

Performance varied sharply by domain: logic puzzles (LiveBench) benefited most from LLM collaboration, knowledge questions (GPQA) saw modest gains, while mathematically intricate AIME problems showed no improvement. This variation stems from the nature of the tasks themselves: LiveBench problems reward iterative constraint resolution and creative reasoning, whereas AIME problems demand long uninterrupted chains of deduction that collaborative turn-taking often disrupts. Meanwhile, conversation traces reveal distinct architectural behaviors: simulated agents tend to make early, lightly justified commitments to maintain role contrast, leading to both frequent improvements and occasional regressions. Dual agents, in contrast, typically construct more comprehensive initial reasoning before committing, resulting in greater stability and higher rates of complete agreement when their initial path is correct.

Token-cost analysis showed no systematic penalty for independence; completion-token counts were dominated by the benchmark itself.

**Implications** The evidence suggests that model separation is an *optional amplifier* rather than a universal prerequisite. Independence is worth the extra complexity wherever dialogue relies on adversarial critique or Socratic guidance, but offers little in cooperative brainstorming. More broadly, the findings clarify how future multi-agent AI ecosystems might be architected: reserve truly separate models for roles that thrive on divergent priors; rely on a unified model where a single coherent workspace is an advantage.

**Caveats and Future Directions** The study shares the methodological limits and constraints already discussed in Sections [3.5](#) and [4.2.6](#): truncation of long reasoning chains because of the token limit, reliance on a single base model, the inherent variability of LLMs, and a strict turn-taking protocol that constrains perspective generation. Extending the analysis to heterogeneous model pools, larger context windows, and asynchronous interaction topologies will test the durability of the conclusions.

**Closing Remarks** Multi-agent dialogue improves the reasoning power of language models, but most of the gain arises from the structured discipline of role interaction itself. Genuine model independence provides an additional edge only under specific dialogue dynamics. By disentangling these effects, this work refines our theoretical understanding of collaborative AI and lays a clearer foundation for future research on scalable, multi-agent intelligence.

# Glossary

**Agent** In this thesis, an LLM instance playing a specific persona (e.g. Proponent, Critic, Teacher, Student) within a collaborative dialogue.

**Benchmark** A standardized dataset or task suite (e.g. AIME, GPQA, LiveBench) used to evaluate model performance under controlled conditions.

**Chain-of-Thought (CoT)** A prompting technique that encourages an LLM to generate intermediate reasoning steps before producing its final answer.

**Cognitive Continuity** The property of a single model retaining a unified latent workspace across multiple turns, enabling it to build on all preceding context.

**Token** The basic unit of text the model processes (roughly a wordpiece); used to count prompt length, completion length, and billing on APIs.

**Completion Token** A token generated by the model in its output (as opposed to tokens consumed by the prompt); used to measure generation cost.

**Collaboration Strategy** The structured interaction pattern that defines how agents converse—here, Debate (critique), Cooperative (additive), or Teacher–Student (guided discovery).

**Inference** The process of generating model outputs from a given prompt, typically via API calls or local forward passes through the network.

**Perspective Diversity** The epistemic distance generated by assigning contrasting roles to agents, yielding different viewpoints even within a single model.

**Prompt** The text (system and user messages) provided to the model to elicit desired behavior or responses.

**Sampling Independence** The use of separate model instances (and random seeds) so that each agent’s hidden states and sampling trajectories are uncorrelated.

# References

Pei Chen, Boran Han, and Shuai Zhang. 2024. [Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving](#). *Preprint*, arXiv:2404.17729.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huaqian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaoqun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanqia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou,

- Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. [Improving factuality and reasoning in language models through multiagent debate](#). *Preprint*, arXiv:2305.14325.
- Mahmood Hegazy. 2025. [Diversity of thought elicits stronger reasoning capabilities in multi-agent debate frameworks](#). *Preprint*, arXiv:2410.12853.
- Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. 2025. [Rethinking mixture-of-agents: Is mixing different large language models beneficial?](#) *Preprint*, arXiv:2502.00674.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). *Preprint*, arXiv:2303.17651.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. [Gpqa: A graduate-level google-proof q&a benchmark](#). *Preprint*, arXiv:2311.12022.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. [Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration](#). *Preprint*, arXiv:2307.05300.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddhartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. [Livebench: A challenging, contamination-limited llm benchmark](#). *Preprint*, arXiv:2406.19314.

Sen Yang, Yafu Li, Wai Lam, and Yu Cheng. 2025. [Multi-llm collaborative search for complex problem solving](#). *Preprint*, arXiv:2502.18873.

# A. Benchmark Question Examples

This appendix provides representative examples from each benchmark used in the study to illustrate the diverse reasoning challenges presented to the agent systems.

## A.1. AIME (2021-2024) Example

### AIME 2022 Problem

For positive integers  $a$ ,  $b$ , and  $c$  with  $a < b < c$ , consider collections of postage stamps in denominations  $a$ ,  $b$ , and  $c$  cents that contain at least one stamp of each denomination. If there exists such a collection that contains sub-collections worth every whole number of cents up to 1000 cents, let  $f(a, b, c)$  be the minimum number of stamps in such a collection. Find the sum of the three least values of  $c$  such that  $f(a, b, c) = 97$  for some choice of  $a$  and  $b$ .

## A.2. GPQA Diamond Example

### GPQA Diamond Astronomy Question

The universe is filled with the Cosmic Microwave Background. Consider the annihilation of high energy  $\gamma$ -rays with a photon from the CMB Radiation into electron-positron, i.e.  $\gamma\gamma \rightarrow e^+e^-$ . From what energy  $\gamma$ -rays would have their lifetimes in the universe limited by this process? Knowing that the average photon energy of the CMB is  $10^{-3}eV$ .

- A.  $9.5 \cdot 10^4$  GeV
- B.  $3.9 \cdot 10^5$  GeV
- C.  $1.8 \cdot 10^5$  GeV
- D.  $2.6 \cdot 10^5$  GeV

### A.3. LiveBench Reasoning Example - Web of Lies

#### Web of Lies Puzzle

In this question, assume each person either always tells the truth or always lies. Jaxon is at the city hall. Charlie is at the train station. Nia is at the planetarium. The person at the library says the person at the bowling alley lies. The person at the planetarium says the person at the airport lies. Luna is at the school. The person at the skate park says the person at the ice skating rink lies. Beatriz is at the airport. The person at the planetarium thinks their neighbor is telling the truth. Isabella is at the campground. The person at the barbershop says the person at the bowling alley tells the truth. Nadia is at the ice skating rink. The person at the theater says the person at the skate park lies. Emily is at the skate park. Ryan is at the library. The person at the city hall says the person at the bowling alley tells the truth. The person at the ice skating rink says the person at the bowling alley tells the truth. The person at the bowling alley says the person at the campground lies. Olivia is at the bowling alley. Zara is at the observatory. Ayaan is at the farm. The person at the skate park says the person at the farm lies. The person at the train station says the person at the campground tells the truth. The person at the gym says the person at the skate park lies. The person at the cafe lies. The person at the ice skating rink says the person at the school lies. The person at the campground says the person at the observatory tells the truth. The person at the airport says the person at the gym lies. The person at the planetarium thinks their friend is lying. The person at the observatory lies. Priya is at the theater. The person at the observatory thinks their friend is lying. Kehinde is at the cafe. Anika is at the barbershop. The person at the gym thinks their neighbor is telling the truth. Devika is at the gym. Does the person at the planetarium tell the truth? Does the person at the airport tell the truth? Does the person at the gym tell the truth? Think step by step, and then put your answer in **bold** as a list of three words, yes or no (for example, **yes, no, yes**). If you don't know, guess.

## A.4. SimpleBench Example

### SimpleBench Question

Beth places four whole ice cubes in a frying pan at the start of the first minute, then five at the start of the second minute and some more at the start of the third minute, but none in the fourth minute. If the average number of ice cubes per minute placed in the pan while it was frying a crispy egg was five, how many whole ice cubes can be found in the pan at the end of the third minute?

- A. 30
- B. 0
- C. 20
- D. 10
- E. 11
- F. 5

## B. Benchmark-Specific Instructions

This appendix provides the benchmark-specific instructions that were added to the basic system prompts to ensure proper answer formatting across different problem types.

### B.1. AIME Instructions

#### AIME-specific Instructions

**IMPORTANT:** This is a mathematics problem from the American Invitational Mathematics Examination (AIME). In EVERY response except the final turn, you must include a line with your current best answer using the format 'Answer: N' where N is a specific integer (0-999). If you haven't fully solved the problem yet, use 'Answer: [working]' or 'Answer: [calculating]' instead of guessing a number. Only provide a numerical answer when you're confident in your calculation. This helps track your reasoning progress. AIME problems always have integer answers between 0 and 999 inclusive. Your final answer **MUST** be in the format 'Final Answer: N' where N is your integer answer.

## B.2. GPQA Instructions

### GPQA-specific Instructions

IMPORTANT: This is a multiple-choice question from the Graduate-level Professional QA (GPQA) dataset. In EVERY response except the final turn, you must include a line with your current best answer using the format 'Answer: X' where X is a specific choice (A, B, C, D). DO NOT use placeholders like 'still thinking' or 'unclear' - make your best guess if uncertain. This intermediate answer must be included even when you're not fully confident. This helps track your reasoning progress. The question requires expertise in a specialized domain. Your final answer MUST be in the format 'Final Answer: X' where X is exactly one of the provided options (A, B, C, D).

## B.3. LiveBench Instructions

### LiveBench-specific Instructions

IMPORTANT: This is a reasoning problem from the LiveBench dataset. The problem will include specific instructions for how to format your answer. Pay careful attention to these format requirements and follow them exactly. In EVERY response except the final turn, include your current best answer using 'Answer: X' format, ensuring X is formatted as requested. Make your best guess if uncertain. In your final response, use 'Final Answer: X' and ensure X is formatted precisely as requested in the problem.

## B.4. SimpleBench Instructions

### SimpleBench-specific Instructions

IMPORTANT: This is a multiple-choice question from the SimpleBench dataset. In EVERY response except the final turn, you must include a line with your current best answer using the format 'Answer: X' where X is a specific choice (A, B, C, D, E or F). DO NOT use placeholders like 'still thinking' or 'unclear' - make your best guess if uncertain. This intermediate answer must be included even when you're not fully confident. This helps track your reasoning progress. Your final answer MUST be in the format 'Final Answer: X' where X is exactly one of the provided options (A, B, C, D, E, or F).

# C. System Prompts for Collaboration Strategies

This appendix contains the complete system prompts used to implement each collaboration strategy in both the simulated and dual-agent configurations.

## C.1. Debate Strategy

### C.1.1. Agent A (Proponent)

#### Proponent Prompt

You are Agent A, a reasoning agent acting as the proponent in a structured dialogue. Your role is to present well-structured arguments supporting your proposed solution to the problem. Provide clear reasoning, cite relevant principles when applicable, and anticipate potential counterarguments. Engage thoughtfully with critiques from Agent B, either by defending your original position with additional reasoning or by refining your answer based on valid criticisms. Remember that your goal is not to "win" but to collaboratively reach the most accurate solution. IMPORTANT: When you see the prompt '(final turn)', you MUST end your response with 'Final Answer: X', where X is your definitive conclusion. This is critical for evaluation purposes.

## C.1.2. Agent B (Critic)

### Critic Prompt

You are Agent B, a reasoning agent acting as the critic in a structured dialogue. Your role is to carefully analyze and challenge the arguments presented by Agent A. Ask probing questions, identify potential weaknesses in reasoning, point out missing considerations, and suggest alternative perspectives when appropriate. Your goal is not to be adversarial but to ensure that the final solution is robust and accounts for all relevant factors. Be constructive in your criticism, suggesting improvements rather than merely pointing out flaws. This collaborative critique process should lead to a more thoroughly reasoned solution. **IMPORTANT:** When you see the prompt '(final turn)', you **MUST** end your response with 'Final Answer: X', where X is your definitive conclusion. This is critical for evaluation purposes.

## C.2. Cooperative Strategy

### C.2.1. Agent A (Proposer)

### Proposer Prompt

You are Agent A, a reasoning agent responsible for initiating problem-solving approaches. Your role is to analyze the given problem, identify key components and constraints, and propose initial solution paths. Break down complex problems into manageable pieces and suggest possible analytical frameworks or methods that might be applicable. Your strength lies in setting up the foundational structure for solving the problem. You don't need to provide complete solutions - focus on establishing productive directions that Agent B can develop further. Be clear, specific, and open to refinement of your initial ideas. **IMPORTANT:** When you see the prompt '(final turn)', you **MUST** end your response with 'Final Answer: X', where X is your definitive conclusion. This is critical for evaluation purposes.

## C.2.2. Agent B (Extender)

### Extender Prompt

You are Agent B, a reasoning agent focused on developing and extending solution paths. Your role is to build upon the foundation laid by Agent A, adding depth and nuance to the analysis. When Agent A proposes an approach, your job is to enhance it by filling in missing details, expanding the reasoning, connecting it to relevant concepts, or contributing complementary perspectives. Your strength lies in elaboration and refinement rather than starting from scratch. Approach this as a collaborative effort where your contributions help create a more comprehensive and robust solution. Avoid simply repeating what Agent A has already covered - instead, add genuine value through extension and development of ideas. **IMPORTANT:** When you see the prompt '(final turn)', you **MUST** end your response with 'Final Answer: X', where X is your definitive conclusion. This is critical for evaluation purposes.

## C.3. Teacher-Student Strategy

### C.3.1. Agent A (Teacher)

### Teacher Prompt

You are Agent A, a reasoning agent acting as a guide and mentor in this problem-solving dialogue between you and Agent B. Your role is to provide scaffolding for effective reasoning about the problem without simply stating the answer. Use Socratic questioning to help Agent B explore the problem space, highlight important principles or frameworks that might be useful, and gently correct misconceptions while explaining why they're problematic. When appropriate, introduce analogies or simplified models to clarify complex concepts. **IMPORTANT:** Only respond as yourself (Agent A). **DO NOT** simulate Agent B's responses or answer your own questions. Wait for Agent B to respond in their own turn. Your goal is to help Agent B develop their own understanding and reasoning skills.

In every turn, include Answer: X with what YOU think is correct, but present it as "What do you think about Answer: X?" or a similar phrase when talking to the student. **IMPORTANT:** When you see the prompt (final turn), you **MUST** end your response with Final Answer: X, where X is your definitive conclusion. This is critical for evaluation purposes.

### C.3.2. Agent B (Student)

#### Student Prompt

You are Agent B, a reasoning agent engaged in active problem-solving under guidance from Agent A. Your role is to approach the problem thoughtfully, making genuine attempts to work through it step by step. Think aloud about your reasoning process, including points of uncertainty or confusion. When Agent A provides guidance, build upon it to advance your understanding rather than simply accepting it passively. Ask specific questions when concepts are unclear, and try to connect new insights to what you already understand. **IMPORTANT:** Only respond as yourself (Agent B). **DO NOT** simulate what Agent A might say next. Your goal is to develop your own coherent solution to the problem with assistance, not to have the solution handed to you. Demonstrate your evolving understanding as the dialogue progresses. **IMPORTANT:** When you see the prompt (final turn), you **MUST** end your response with Final Answer: X, where X is your definitive conclusion. This is critical for evaluation purposes.

## D. Use of Text Generation Models

In compliance with the requirement to document the contributions of text generation models, I hereby declare the following:

**Code development** I employed Anthropic’s Claude Sonnet 3.7 to generate and refine the overall file structure of the implementation, to draft helper functions, and to assist in debugging API integration issues.

**Manuscript writing and formatting** I used both Claude Sonnet 3.7 and OpenAI’s ChatGPT o3 to improve the idiomatity and readability of certain prose passages, to brainstorm content for sections (e.g. framing of motivation, definition of research questions), and to obtain LaTeX assistance for constructing tables, figures, and the custom ‘promptbox’ environment.



## Declaration of Independent Authorship

### Original work

I expressly declare that the written work I submitted to the University of Zurich in the spring/autumn semester of 2025 with the title

Two Heads or One? Mechanisms Driving Performance in Multi-Agent LLM Reasoning

is an original work written by myself, in my own words, and without unauthorized assistance. If it is a work by several authors, I confirm that the relevant parts of the work are correctly and clearly marked and can be clearly assigned to the respective author.

I also confirm that the work has not been submitted in whole or in part to receive credit for another module at the University of Zurich or another educational institution, nor will it be submitted in the future.

### Use of sources

I expressly declare that I have identified all references to external sources (including tables, graphics etc.) contained in the above work as such. In particular, I confirm that, without exception and to the best of my knowledge, I have indicated the authorship both for verbatim statements (citations) and for statements by other authors reproduced in my own words (paraphrases).

### Use of text generation models

I expressly declare that I have not only identified existing external sources, but also any automatically generated text that is contained in the above work. I have used the same citation style as if the text had been generated by a human to indicate the source of the automatically generated text. If the contribution of text generation models cannot be linked to specific text passages (see the associated guidelines), I have included a chapter describing the contributions of the text generation model. I acknowledge that no explicit citation is necessary where text generation models are merely used correctively (to improve grammar or idiomaticity of my own words).

### Sanctions

I acknowledge that a thesis that is used to acquire credit and proves to be plagiarism with the meaning of the document [Erläuterung des Begriffs „Plagiat“](#) leads to a grade



deduction in minor cases, a grade 1 (one) in more severe cases, without the possibility of revision, and in very severe cases can have the corresponding legal and disciplinary consequences according to §§ 7ff of the "Disziplinarordnung der Universität Zürich" and § 36 of the "Rahmenordnung für das Studium in den Bachelor- und Master-Studiengängen der Philosophischen Fakultät der Universität Zürich".

I confirm with my signature that this information is correct:

Name: White

First Name: Quentin Tyrone Shema

Matriculation number:

Date: 30.05.2025

Signature:

A handwritten signature in black ink, appearing to be 'Q. Shema', written over a faint horizontal line.