DEBATE, TRAIN, EVOLVE: Self-Evolution of Language Model Reasoning

Anonymous EMNLP submission

Abstract

Large language models (LLMs) have improved significantly in their reasoning through extensive training on massive datasets. However, relying solely on additional data for improvement is becoming increasingly impractical, highlighting the need for models to autonomously enhance their reasoning without external supervision. In this paper, we propose DEBATE, TRAIN, EVOLVE (DTE), a novel ground truth-free training framework that uses multi-agent debate traces to evolve a single language model. We also introduce a new prompting strategy REFLECT-CRITIQUE-REFINE, to improve debate quality by explicitly instructing agents to critique and refine their reasoning. Extensive evaluations on **five** reasoning benchmarks with **six** open-weight models show that our DTE framework achieve substantial improvements, with an average accuracy gain of **8.92%** on the challenging GSM-PLUS dataset. Furthermore, we observe strong cross-domain generalization, with an average accuracy gain of **5.8%** on all other benchmarks, suggesting that our method captures general reasoning capabilities.

1 Introduction

Over the past few years, the advancements in large language models (LLMs) have largely depended on training over massive datasets (Abdin et al., 2024, 2025). However, eventually, we will approach a saturation point where feeding more data into these models may not further improve their reasoning capabilities (Costello et al., 2025). This motivates a new research question: *How can language models continue to improve without relying on additional external supervision?*

Recent approaches attempt to overcome the data bottleneck by enabling models to generate and learn from synthetic data, which is generated by automatically expanding a small set of seed tasks into large synthetic instruction datasets (Wang et al., 2022; Zeng et al., 2024). Other methods (Madaan et al., 2023; Jiang et al., 2023; Gou et al., 2023; Peng et al., 2023; Zelikman et al., 2024; Costello et al., 2025) refine model-generated outputs through iterative self-feedback or preference optimization. Despite their effectiveness, these self-evolution strategies predominantly rely on judgments from a single model or a teacher-student configuration, often leading to confirmation bias and insufficient reasoning diversity.

To address these limitations, one promising direction emerged is multi-agent debate (MAD) (Du et al., 2023). It involves multiple models independently generating and critically analyzing each other's answers, helping to reveal subtle reasoning errors often overlooked by individual models (Liang et al., 2023; Wang et al., 2024). Although MAD shows improved reasoning accuracy, current works predominantly use MAD as an inference-time technique (Smit et al., 2023), requiring multiple models to be run simultaneously for each query. This substantially increases computational overhead and latency (Subramaniam et al., 2025), making MAD impractical for large-scale deployments. This motivates our research question: *Can we **evolve** a single model reasoning by fine-tuning on these debate traces?*

Building upon this intuition, we propose DEBATE, TRAIN, EVOLVE (DTE), a novel framework that combines the strengths of MAD with efficient single-model inference. Specifically, we introduce a ground-truth-free training approach in which a model learns from its own debate traces generated during MAD, thereby evolving autonomously over iterative training cycles. Our framework addresses key challenges of existing methods by extracting high-quality reasoning insights from diverse multi-agent interactions, thus avoiding single-model biases and computational inefficiencies.

First, we conduct a large-scale empirical analysis of MAD using open-source models, where we identify limitations of the original MAD (Du

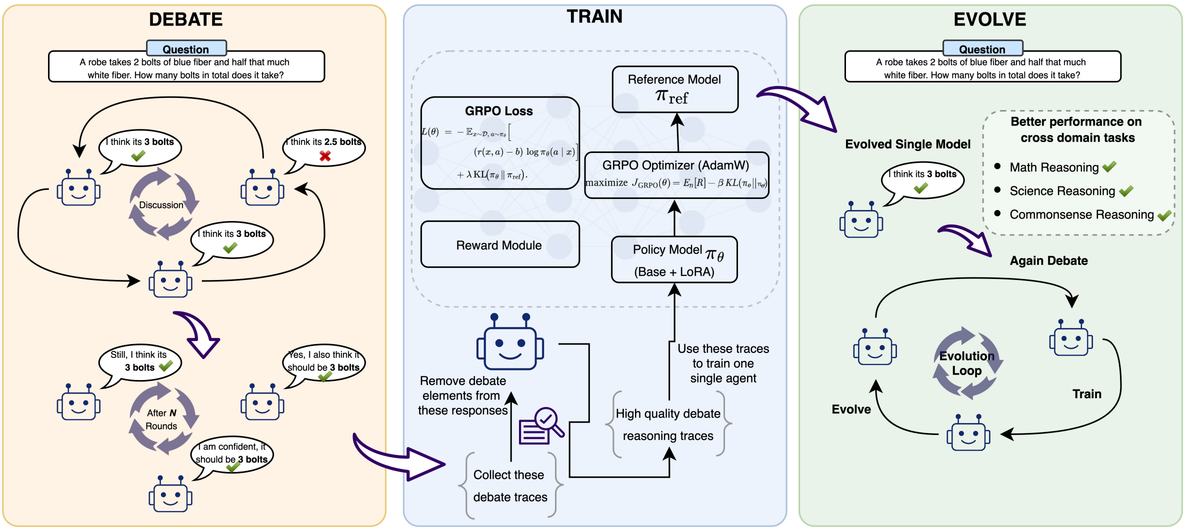

Figure 1: Overview of the proposed DEBATE–TRAIN–EVOLVE framework. *Left*—**Debate**: Several agents debate until they converge on a consensus (green ✓) or expose a wrong path (red ✗). *Centre*—**Train**: we remove pure debate elements, keep the high-quality reasoning traces and consensus answer, and use them to fine-tune a single policy with GRPO. *Right*—**Evolve**: the evolved agent replaces its earlier self, so future inference require just one forward pass yet they outperform the committee on maths, science, and commonsense benchmarks.

et al., 2023) prompting approach, particularly in smaller models. To address this, we propose a REFLECT-CRITIQUE-REFINE (RCR) prompting strategy, which explicitly forces agents to identify, critique, and correct reasoning errors in both their own and peers' answers. **Second,** using this prompting strategy, we build our DTE framework (Figure 1). **Finally**, we find that models with $< 3B$ parameters suffer accuracy loss after second evolution round; our controlled study shows that the problem correlates with large temperature-induced variance and high KL divergence from the base policy. Lowering the sampling temperature from 0.7 to 0.3 cuts the KL drift by $1/3rd$ and recovers up to 76% of the lost performance, preventing catastrophic forgetting without extra supervision.

Our experiments show significant gains in reasoning performance across multiple datasets. Specifically, our evolved models show an average accuracy improvement of **8.92%** on the challenging GSM-PLUS dataset compared to their original versions. Moreover, our framework achieves notable cross-domain generalization, enhancing model performance across datasets not seen during training. These results confirm that our DEBATE, TRAIN, EVOLVE method successfully distills multi-agent debate's insights into efficient single-model inference, bridging the gap between computational efficiency and advanced reasoning capabilities.

2 Related Work

Multi-Agent Debate Approaches Du et al. (2023) first showed that letting several large models debate improves accuracy on maths, strategy, and factual QA without any new parameters. Later, Liang et al. (2023) highlighted the risk of *degeneration-of-thought*: a single agent quickly converges on one path, whereas a two-debater plus judge setup maintains diversity and outperforms GPT-4 on tricky arithmetic. RECONCILE (Chen et al., 2023) mixes agents from different model families, reaches consensus through confidence-weighted votes, and adds up to eleven points on seven reasoning benchmarks. Smit et al. (2023) shows that MAD beats sampling ensembles only after careful tuning. Finally, works like PREDICT (Park et al., 2024) apply multi-agent debate to tasks beyond QA, such as hate-speech classification, where agents reason under different guidelines. Recent advances further incorporate explicit reinforcement learning into the debate process. For example, the ACC-Collab framework (Estornell et al., 2024) utilized an actor-critic approach to explicitly optimize agent collaboration, yielding superior performance on reasoning tasks.

Self-Evolution in Language Models SELF-INSTRUCT (Wang et al., 2022) prompts GPT-3 to write 52000 novel instructions plus answers and then fine-tunes on its own output, reducing the gap to InstructGPT by thirty-three points on Super-Natural-Instructions without extra human labels. STAR (Zelikman et al., 2024) augments a few chain-of-thought exemplars by letting the model explain wrong answers in reverse, doubling CommonsenseQA accuracy for a 350M model. SELF-REFINE (Madaan et al., 2023) and the broader SELF framework (Lu et al., 2023) turn one model into writer, critic and re-writer, looping feedback at inference or during fine-tuning to improve on GSM8K by around seven points. Instruction-tuning variants refine the idea: SELF-REFINE INSTRUCTION-TUNING (Ranaldi and Freitas, 2024) pairs Llama-2 and Mistral students with large teacher rationales and then lets each student prefer its own better reasoning, closing the size gap on commonsense and math tasks. More recently, THINK, PRUNE, TRAIN, IMPROVE (Costello et al., 2025) shows that careful filtering of self-generated traces can raise Gemma-2B to 58% on GSM8K and push Llama-3-70B beyond GPT-4o. These studies confirm that single-agent loops, with or without ground truth, can expand a model's ability.

Despite these works, two things remain unexplored: ***1)*** Fully autonomous, ground-truth-free self-evolution; ***2)*** Integration of MAD into model evolution. Our work addresses this by the DEBATE, TRAIN, EVOLVE framework, which combines MAD with self-supervised reinforcement learning (GRPO) to enable models to autonomously evolve their reasoning capabilities.

3 Method

Preliminary Study: Multi-Agent Debate (MAD) First, we conduct a large-scale empirical study on MAD. Our initial objective was to understand how MAD affects performance across different reasoning tasks (eg., math, science), model families, and across scales, and to identify its limitations. Our ultimate goal was to generate high-quality reasoning traces using MAD. Following MAD (Du et al., 2023) prompting strategy, however, our initial experiments show that the gains from MAD were minimal across reasoning benchmarks.

Our manual inspection uncovered two primary issues: **(i) sycophancy:** agents often discarded their correct answer and copied a other agent's wrong but confident solution; **(ii) verbosity bias:** when two answers disagreed, agents preferred the longer rationale even if it was logically weak (Saito et al., 2023). For example, on average, a three-agent Qwen-1.5B committee, sycophancy reached 0.28 (28% of the debates have sycophancy rounds). These effects lead to a substantial fraction of [correct $\rightarrow$ incorrect] transitions during debate.

Algorithm 1: MAD using RCR prompting

Input: query q, agents $\mathcal{A}$, max rounds $T = 5$
Round 0: each $a_i \in \mathcal{A}$ outputs $(y_i^{(0)}, r_i^{(0)})$
if *all* $y_i^{(0)}$ *identical* **then**
  return this answer
end
for $t = 1$ **to** T **do**
  each a_i receives $\{(y_j^{(t-1)}, r_j^{(t-1)})\}_{j \neq i}$
  reflect: list potential error in $y_i^{(t-1)}$
  critique: point out one flaw in two distinct peers' rationales
  refine: output $(y_i^{(t)}, r_i^{(t)})$, adding at least one novel step if $y_i^{(t)} \neq y_i^{(t-1)}$
  if *all* $y_i^{(t)}$ *identical* **then**
    return consensus answer
  end
end
return majority_vote$(\{y_i^{(T)}\})$

3.1 REFLECT–CRITIQUE–REFINE Prompting

Inspired by previous works that had iterative self-feedback procedures (Madaan et al., 2023; Gou et al., 2023; Peng et al., 2023), we propose REFLECT-CRITIQUE-REFINE (RCR) prompting, which mitigates this problem. Unlike the original MAD prompt that simply asks agents to revise their answer by looking at others, RCR prompting encourages agents to present confident reasoning. 1) Every agent must *reflect* by stating why its current answer could be wrong; 2) Each agent then *critiques* exactly two peer rationales, naming a specific flaw or a step it cannot verify; 3) Finally the agent may *refine* its answer—but only if it supplies at least one new reasoning step.

Phrases like *"identify any errors"* reliably trigger negative tokens **("error", "mistake", "step X is wrong")** which LLMs have learned during supervised finetuning. By specifying valid next moves (defend/correct/adopt), we implicitly shape the log-probability mass toward useful trajectories, shrinking the space of rambling answers. The single-step explanation requirement forces agents to think before copying, while the fixed quota of peer critiques limits needless verbosity. This prompting helped reduce sycophancy and improved performance.

Debate protocol Let $\mathcal{A} = \{a_1, \ldots, a_N\}$ be a set of agents and q a query. Round 0: each a_i produces $(y_i^{(0)}, r_i^{(0)})$, an answer and a rationale. If *consensus* (all answers identical) occurs, the process stops. Otherwise we enter a debate round $t \geq 1$ following Algorithm 1: agents receive all previous rationales, execute REFLECT–CRITIQUE–REFINE, and output $(y_i^{(t)}, r_i^{(t)})$. The debate ends when either (i) consensus is reached or (ii) a maximum of N rounds is exhausted, after which the final answer is decided by simple majority vote.

3.2 DEBATE, TRAIN, EVOLVE Framework

So far we have seen that RCR prompting yields high-quality debate traces and measurable accuracy gains. We now move from *learning* from these debate traces to train a single language model. This results in DEBATE, TRAIN, EVOLVE (DTE), a closed loop that distills the collective insight of multiple agents into a single, faster model.

Framework overview Let π_{θ_0} be a frozen base policy with parameters θ_0. At iteration k we run Algorithm 1 on a batch of queries $\mathcal{Q}_k$, producing a set of debate traces $\mathcal{D}_k = \{(x, y^\star, R)\}$ where x is the original query, $y^\star$ the consensus answer and R a compressed rationale extracted from the highest-scoring agent. **Extraction step** keeps any explanation steps that appear in at least two agents or that introduce a symbolic manipulation absent from the previous round.

We then fine-tune a student policy π_{θ_k} on $\mathcal{D}_k$ with Group Relative Policy Optimization (Shao et al., 2024) and then place π_{θ_k} back into the agent pool, replacing its earlier version. The loop repeats until the mean validation reward stagnates for two iterations or a maximum of five cycles is reached. Algorithm 2 gives the exact procedure.

Reward shaping and GRPO For a query x and a student output y we define the shaped reward

$$r(x, y) = w_{\text{vote}}\,[y = y^\star] + w_{\text{fmt}}\,\text{isFmt}(y) + w_{\text{brev}}\,\exp(-|y|/\tau),$$

where $[\,\cdot\,]$ is the indicator function, isFmt checks the <reasoning> <answer> XML template, and $|y|$ is the token length. Weights are set to $(w_{\text{vote}}, w_{\text{fmt}}, w_{\text{brev}}) = (2, 0.5, 0.5)$ and $\tau = 120$.

GRPO maximises the clipped surrogate

Algorithm 2: DEBATE, TRAIN, EVOLVE

Input: base policy π_{θ_0}, agent set $\mathcal{A}_0 = \{\pi_{\theta_0}\} \cup \mathcal{B}$, queries $\mathcal{Q}$, max iterations K

for $k = 1$ **to** K **do**
    sample batch $\mathcal{Q}_k \subset \mathcal{Q}$
    $\mathcal{D}_k \leftarrow \{\}$
    foreach $x \in \mathcal{Q}_k$ **do**
        run Algorithm 1 with agents $\mathcal{A}_{k-1}$ on x
        store $(x,\, y^\star,\, R)$ in $\mathcal{D}_k$
    end
    update $\theta_k \leftarrow \theta_{k-1}$ by maximising $\mathcal{L}_{\text{GRPO}}(\theta; \mathcal{D}_k)$
    $\mathcal{A}_k \leftarrow \big(\mathcal{A}_{k-1} \setminus \{\pi_{\theta_{k-1}}\}\big) \cup \{\pi_{\theta_k}\}$
    if *validation reward* $<1\%$ *better than previous* **then**
        break
    end
end
return final student π_{θ_k}

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_k,\, y\sim\pi_\theta}\big[\min\big(\rho\, r, \operatorname{clip}(\rho, 1-\epsilon, 1+\epsilon)\, r\big) - \beta\, \text{KL}\big(\pi_\theta \,\|\, \pi_{\theta_0}\big)\big],$$

where $\rho = \pi_\theta(y\,|\,x)/\pi_{\theta_{k-1}}(y\,|\,x)$, $\epsilon = 0.2$, and $\beta = 0.02$. The first term is identical to PPO but uses the debate-derived reward; the second anchors the student to the base policy, a key factor for avoiding catastrophic forgetting (Luo et al., 2025; Kotha et al., 2024). Gradients are estimated with the standard REINFORCE algorithm and a value network baseline.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on **five** public reasoning benchmarks: ***1)* GSM8K** (Cobbe et al., 2021), ***2)* GSM-Plus** (Li et al., 2024) (harder numeric reasoning), ***3) ARC-Easy***, ***4) ARC-Challenge*** (Clark et al., 2018), and ***5) CommonsenseQA*** (Talmor et al., 2019).

Baselines and models. We conduct of RCR prompting study on **ten** open-weight models—Qwen (0.5-32B), Llama-3/8B, Mistral-7B, Phi-mini—and **two** proprietary models, GPT-4o and GPT-4o-mini. We study our DTE framework with 6 models (Qwen 1.5B-14B, Llama-3B and Llama-8B). **Baselines are:** (i) the single *original* model; (ii) *vanilla MAD* with the original MAD prompt (Du et al., 2023).

Parameter settings. During debate we sample each agent once per query at temperature $T = 1.0$ (exploratory) or 0.0 (deterministic); mixed-teams

Model	GSM8K			GSM-Plus			ARC-Challenge		
	Original Model	3 Agent MAD	Evolved Single Model (DTE)	Original Model	3 Agent MAD	Evolved Single Model (DTE)	Original Model	3 Agent MAD	Evolved Single Model (DTE)
Qwen-2.5-1.5B	62.77	72.33	73.09 (+10.32 ↑)	42.00	53.33	55.92 (+13.92 ↑)	69.21	68.52	68.36 (-0.85 ↓)
Qwen-2.5-3B	84.08	85.14	86.05 (+1.97 ↑)	61.75	68.00	69.50 (+7.75 ↑)	83.53	84.64	83.95 (-0.42 ↓)
Qwen-2.5-7B	90.67	91.21	88.32 (-2.35 ↓)	68.62	74.17	74.71 (+6.09 ↑)	87.22	91.64	90.89 (+3.67 ↑)
Qwen-2.5-14B	92.80	93.33	93.74 (+0.94 ↑)	71.79	77.25	78.88 (+7.09 ↑)	90.27	93.77	93.13 (+2.86 ↑)
Llama-3.2-3B	72.55	73.84	75.06 (+2.51 ↑)	45.67	51.12	53.79 (+8.12 ↑)	73.12	76.19	77.23 (+4.11 ↑)
Llama-3.1-8B	81.73	82.18	86.81 (+5.08 ↑)	55.62	60.79	66.17 (+10.55 ↑)	77.65	85.07	86.53 (+8.88 ↑)

Table 1: **Performance of one DEBATE–TRAIN–EVOLVE round.** For six open-weight models we report test accuracy on three reasoning benchmarks in three settings: the single *base* model ("Original"), a **3-agent** debate using our RCR prompt ("MAD"), and the *evolved single* student obtained after one DTE round. **Green** numbers denote the absolute gain of the evolved model over its Original Model, **red** numbers a decrease in performance.

use one exploratory and two deterministic agents. For evolution we adopt LoRA fine-tuning (rank 128, dropout 0.05) on attention and MLP projections, freezing embeddings and layer norms. GRPO is optimized with AdamW (learning rate 2×10^{-5}, weight decay 0.01, 50-step linear warm-up). Each evolution epoch processes 8k debate traces ($\sim$2 M tokens) and runs on A100-80 GB GPUs for a 7B model; larger models scale near-linearly.

Evaluation metrics. Task performance is *exact match* for GSM-style datasets and *accuracy* for MC-QA. We also track **Sycophancy-Rate**: the fraction of agents switching to an incorrect peer answer without adding new reasoning; [correct $\rightarrow$ incorrect] and [incorrect $\rightarrow$ correct] instances during MAD.

4.2 Main Results

Our main results are organized into three main parts: ***(1)*** First, we evaluate the effectiveness of DEBATE–TRAIN–EVOLVE (DTE) framework, ***(2)*** Next, we test its generalization across different reasoning tasks by transferring evolved models to new datasets, and ***(3)*** Finally, we analyze the extent of model self-evolution through iterative rounds.

1) OVERALL DTE PERFORMANCE. Evolved model using DTE shows an average gain of 8.92% ACCURACY on GSM-PLUS compared to its vanilla performance. Table 1 contrasts three settings: the single base model ("Original"), a three-agent debate with our RCR prompt ("MAD"), and the *evolved single model* produced by one DEBATE–TRAIN–EVOLVE pass. On **GSM-Plus**—the hard math dataset—DTE improves every model, with an average gain of **+2.38 points** over three-agent MAD. Qwen-1.5B shows the largest jump (+13.92 pts), confirming that ***evolution is most helpful when the base model has head-room and the debate provides diverse traces.*** On **GSM8K** the average gain is smaller (+0.84 pts) because several models were already near their ceiling after debate. **ARC-Challenge** sees a mixed results: large models benefit (+3.67 pts for Qwen-7B, +8.88 pts for Llama-8B) while small models drift by <1 pt. Overall, DTE shows a mean improvement of **3.06 pts** over single model and **+1.09 pts** over MAD while restoring single-pass inference.

2) CROSS-DOMAIN GENERALIZATION. Our results suggests that DTE improves reasoning that travels beyond the source data, with larger models showing the most stable improvements. Table 2 reports how well the evolved models generalize on other datasets. We test two scenarios: evolve using ***(i)*** GSM8K; ***(ii)*** GSM-Plus and test on four unseen datasets. When trained on GSM8K, every model gains on GSM-Plus (average **+5.8** pts) and on ARC-Challenge (+2.5 pts on average). ARC-Easy also sees small but consistent gains except for the 1.5B model, which drops 1.6 pts. CommonsenseQA improves for 5/6 models, indicating that the reward shaped from mathematical traces still helps improve on commonsense reasoning. Negative deltas are confined to the smallest model (Qwen-1.5B) and to a lesser degree Qwen-3B, suggesting that small models struggles to reconcile new skills with prior knowledge. In contrast, models $\geq$ 7B never lose more than 0.2 pts on any transfer task. Training on GSM-Plus and testing on GSM8K yields similar behaviour: large gains on the GSM8K (+3.7 pts on average) and moderate gains on others. ***The symmetry suggests that DTE learns general reasoning heuristics (e.g. numeric decomposition, unit tracking) rather than memorising dataset-specific patterns.***

Model	Fine-tuned on GSM8K				Fine-tuned on GSM-Plus			
	GSM-Plus (Δ)	ARC-Easy (Δ)	ARC-Challenge (Δ)	CommonsenseQA (Δ)	GSM8K (Δ)	ARC-Easy (Δ)	ARC-Challenge (Δ)	CommonsenseQA (Δ)
Qwen-2.5-1.5B	+9.21 ↑	-1.60 ↓	+0.67 ↑	-2.23 ↓	+10.32 ↑	-1.52 ↓	+0.24 ↑	-2.31 ↓
Qwen-2.5-3B	+3.79 ↑	+1.27 ↑	+0.83 ↑	+3.26 ↑	+1.36 ↑	+1.09 ↑	+0.60 ↑	+3.26 ↑
Qwen-2.5-7B	+1.01 ↑	+1.73 ↑	+4.50 ↑	+3.40 ↑	+1.14 ↑	+1.69 ↑	+3.65 ↑	+3.32 ↑
Qwen-2.5-14B	+1.67 ↑	+2.53 ↑	+3.42 ↑	+1.33 ↑	+0.53 ↑	+2.32 ↑	+4.01 ↑	-0.14 ↓
Llama-3.2-3B	+6.71 ↑	+2.48 ↑	-1.11 ↓	+3.10 ↑	+3.80 ↑	+1.93 ↑	-3.92 ↓	+3.51 ↑
Llama-3.1-8B	+8.13 ↑	+3.91 ↑	+6.74 ↑	+1.10 ↑	+5.15 ↑	+4.88 ↑	+7.84 ↑	+0.85 ↑

Table 2: **Cross-domain generalisation of evolved students.** Each cell shows the change in test accuracy (Δ, in points) after one DTE pass, relative to the same model before evolution. The table is split by the dataset used for fine-tuning—GSM8K (left block) or GSM-Plus (right block)—and reports transfer to four unseen targets. **Green** numbers signal gains, **red** numbers losses.

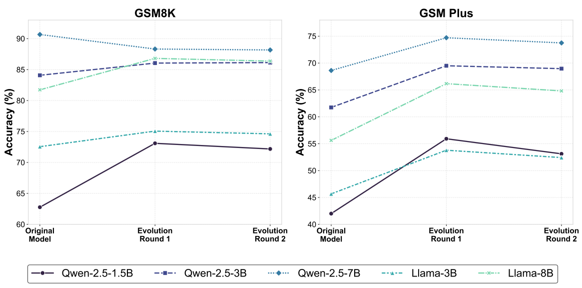

Figure 2: **Accuracy vs. evolution round.**

3) How far can a model evolve? Results show that one evolution round captures nearly all of the available gains. Figure 2 reports accuracy over two evolution rounds for five models on GSM8K and GSM-Plus. Round 1 almost always helps: the smallest model (Qwen-1.5B) jumps from 42.0 → 55.9 on GSM-Plus and 62.8 → 73.1 on GSM8K, while Llama-8B gains 10.6 and 5.1 points on the same datasets. The only counter-example is Qwen-7B, which drops 2.4 points on GSM8K despite improving 6.1 on GSM-Plus; upon manual inspection we see that its Round-1 traces over-emphasise shortcut heuristics that hurt easier questions. **In Round 2, we observe little improvement and sometimes the performance even drops.** Large models (≥ 7 B) add at most +0.8 points, for Qwen-3B on GSM8K, and more often lose 0.4–1.4 points. The 1.5B model gives back 0.9 points on GSM8K and 2.8 on GSM-Plus, but still ends well above its starting point. Across all runs the mean forgetting $\mathrm{Fgt}_2 = \max_{t<2}(\mathrm{Acc}_t - \mathrm{Acc}_2)$ is 0.92 pts for models ≥ 7 B and 1.6 pts for smaller ones, confirming that smaller models suffers from catastrophic forgetting.

4.3 Ablation Studies

1) Effectiveness of the RCR prompt in MAD. RCR prompting substantially boost performance over original MAD prompting (Du et al., 2023). Figure 3 compares single-model inference, the original debate prompt (MAD@3), and our Reflect–Critique–Refine (RCR-MAD@3) prompt. Across eight diverse models the RCR prompting raises three-agent accuracy by an average of **+1.9 pts** on GSM8K, **+3.7 pts** on GSM-Plus, and **+0.7 pts** on ARC-Challenge. The gain scales with task difficulty: GSM-Plus, which contains harder adversarial questions, benefits the most (up to +7.9 pts for Qwen-1.5B and +6.1 pts for Qwen-7B). On ARC-Challenge improvements are smaller but still positive for 6/8 models. **RCR prompting also significantly reduces sycophancy.** It *halves* the mean sycophancy rate (from 0.28 to 0.13 on GSM-Plus) and narrows the verbosity gap by 43 %, indicating that agents now switch answers only when they can articulate a new reasoning step. ***These observations confirm that RCR is a necessary pre-step for producing high-quality traces later utilized by the DTE training loop.***

2) How many agents are enough? Results shows that *three* agents MAD captures 85-95 % of the maximum gains. Figure 4 sweeps the agents size from 1 − 7 and reports trends on four benchmark. We observe three clear patterns here: ***1)* Beyond 3-agent the curve plateaus and even oscillates**, suggesting the marginal information added by the 4th or 5th agent. ***2)* Small models benefit most from extra agents.** Already strong single-agent (Qwen-14B) adds minimal improvement upon scaling up after three. ***3)* Harder tasks need (slightly) more agents.** On GSM-Plus the optimum often shifts to four or five agents: Qwen-7B reaches its peak accuracy (76.0%) at 7 agents, 1.04 pts above the three-agent setting. ARC-Easy, a much easier dataset, saturates at 2 agents for every model; extra debaters add noise rather than insight.

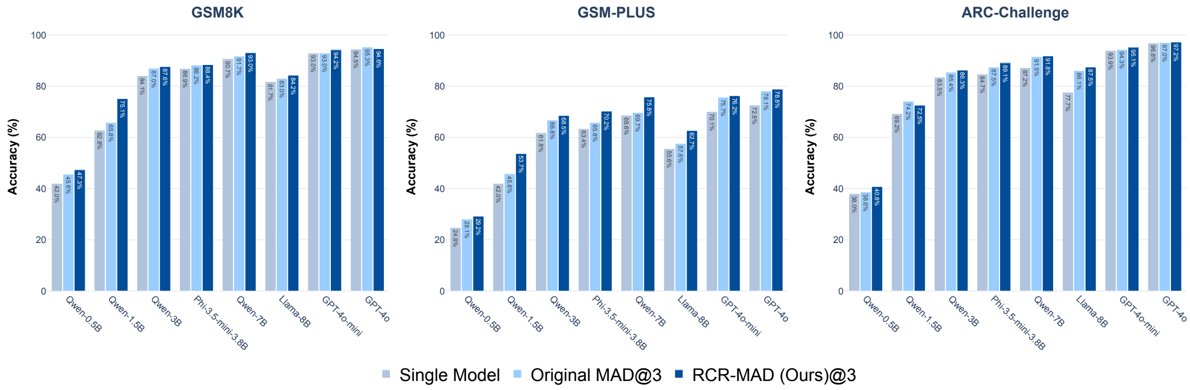

Figure 3: Results (%) on: GSM8K, GSM-PLUS, and ARC-Challenge datasets. Performance is compared across three evaluation settings: single model inference, the Original Multi-Agent Debate (MAD@3) prompt, and our proposed RCR (RCR-MAD (Ours)@3) prompting.

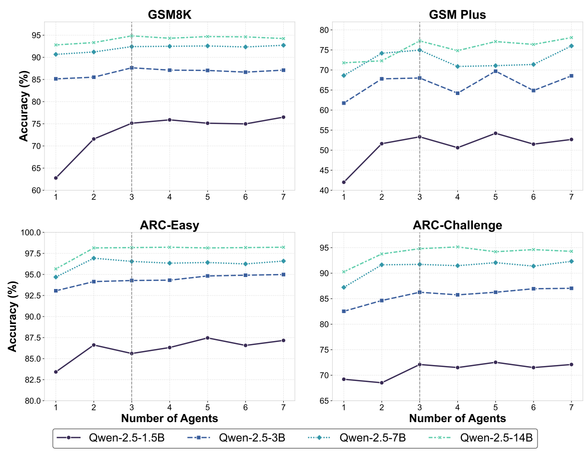

Figure 4: **Scaling up agents** Accuracy of four Qwen model sizes as the number of agents grows from 1-7.

3) Does agent diversity matter? We observe two consistent trends here: **First,** when the individual agents have comparable standalone accuracy, cross-family mixtures beat homogeneous agents team, supporting the idea that architectural diversity yields complementary reasoning paths. **Second,** when the pool mixes a strong and a weaker model, the debate result gravitates toward the stronger member—adding the weaker agent neither helps nor seriously harms, suggesting that diversity only helps when all agents can contribute novel insights. Complete results for every dataset and roster is available in Appendix B.

4) Why GRPO over other fine-tuning methods? **GRPO consistently outperforms the alternatives,** indicating that its relative-advantage reward balances exploration and policy stability better than plain maximum-likelihood (SFT) or preference-only (DPO/PPO) updates. Table 3 compare three update rules under a fixed compute budget: (1) classical supervised fine-tuning on debate answers (SFT); (2) Direct Preference Optimisation using the majority vote as the preferred sample; (3) Group Relative Policy Optimisation (GRPO). GRPO delivers the largest accuracy jump on GSM-Plus for every model size. Both SFT and DPO give smaller gains and even slight regressions on the 3 B model, highlighting the risk of overfitting when the reward ignores policy shift. We also observe that GRPO keeps KL < 0.24 across sizes, whereas DPO averages 0.43. The relative-advantage term in GRPO therefore not only boosts reward but also constrains drift, reducing catastrophic forgetting.

Model	Original (GSM-Plus)	SFT	DPO	GRPO
Qwen-2.5-1.5B	42.00	47.31	51.34	**55.92**
Qwen-2.5-3B	61.75	58.33	64.32	**69.50**
Qwen-2.5-7B	68.62	67.89	69.88	**74.71**

Table 3: Accuracy on GSM-Plus after **10K** training steps using three optimization objectives.

5) Data selection strategy. We test three data sampling schemes on GSM-Plus: *Random-2K* selects 2000 examples uniformly from the full pool (10552); *Debate-Only* keeps only data points where agents entered at least one critique round ($t \geq 1$); *All-Traces* trains on the entire cleaned set. Table 4 shows that accuracy rises monotonically with coverage: the full corpus beats Debate-Only

Model	Random-2K	Debate-Only	All-Traces
Qwen-1.5B	44.82	51.61	**55.92**
Qwen-3B	58.10	62.70	**69.50**
Qwen-7B	69.71	72.53	**74.71**

Table 4: **Effect of training-set size and composition.** GSM-Plus accuracy after one evolution round using three trace-selection schemes.

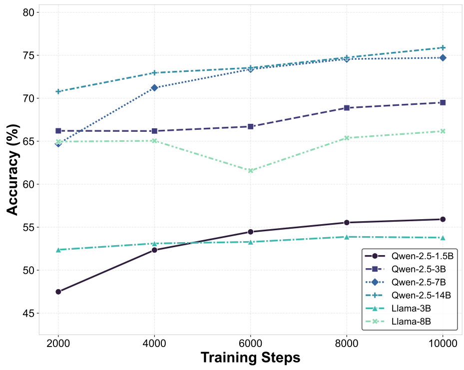

Figure 5: **Diminishing returns in GRPO updates after 8K steps.** GSM-Plus accuracy for five models as a function of the number of training steps during GRPO.

by **4.43 pts** (avg) and Random-2K by **9.17 pts** (avg). The gap is largest for Qwen-1.5B, suggesting that smaller models benefit from easier "round-0" examples that Random-2K may miss and Debate-Only discards. We therefore use the full trace set in all other experiments.

6) How long do we train? Figure 5 plots GSM-Plus accuracy as we grow the number of GRPO training steps from 2K to 10K. All models share the similiar trend: rapid gains up to about 8K steps followed by saturation. Small and mid-size models profit the most from the early updates—Qwen-1.5B climbs 8.0 pts between 2K and 6K samples—whereas larger models such as Qwen-14B rise more slowly but steady. Beyond 8K the curve flattens: the average improvement from 8K ß 10 k is only +0.32 pts while wall-clock time grows by 25%.

7) Does iterative fine-tuning hurt? Figure 6 plots GSM8K and GSM-Plus accuracy for Qwen-1.5B after the first and second evolution rounds under four sampling temperatures. When we keep the original exploratory setting ($T = 1.0$) the model loses 2.0 pts on GSM8K and gains only 13.5 pts on GSM-Plus—well below the +33.5 pts it achieved in Round 1—confirming a clear case of catastrophic forgetting. Lowering the temperature stabilises training: at $T = 0.4$ Round-2 accuracy is within 0.9 pts of Round 1 on GSM-Plus and almost fully recovers on GSM8K; a deterministic schedule ($T = 0.0$) even adds +3.3 pts on GSM8K but plateaus on GSM-Plus.

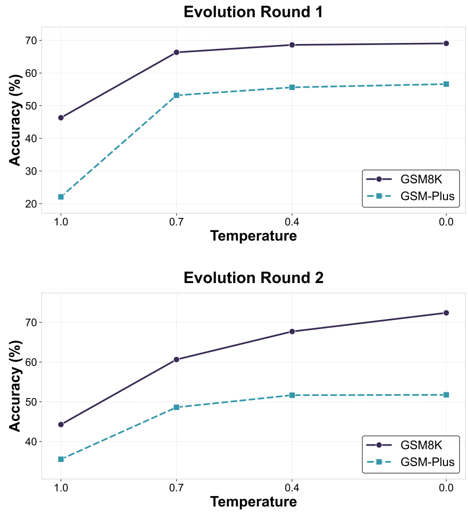

Figure 6: **Iterative fine-tuning and forgetting.** Accuracy of Qwen-1.5 B after the first and second evolution rounds at four sampling temperatures.

The mechanism is visible in the KL divergence between successive students. At $T = 1.0$ we measure $\mathrm{KL}_{\mathrm{evo}}$=0.37 for Qwen-1.5B, whereas $T = 0.4$ cuts this to 0.19 and $T = 0.0$ to 0.11, matching the reduction in forgetting. We therefore adopt a linear decay from 0.7 in Round 1 to 0.3 in later rounds for all models up to 3B parameters; larger models did not require temperature adjustment.

5 Conclusion

In this paper, we introduced the Debate, Train, Evolve (DTE) framework, a novel approach enabling language models to autonomously enhance their reasoning capabilities by leveraging multi-agent debate traces. Our Reflect-Critique-Refine prompting strategy significantly improved debate quality, reducing sycophancy and reasoning errors. Experiments demonstrated substantial accuracy gains, notably an average improvement of **8.92%** accuracy on the challenging GSM-PLUS dataset. Additionally, we showed strong cross-domain generalization, confirming that our approach captures general reasoning skills rather than dataset-specific patterns. Importantly, DTE effectively combines the benefits of multi-agent debate with the computational efficiency of single-model inference.

Limitations

Despite its effectiveness, our approach has certain limitations. Firstly, iterative fine-tuning within the DTE framework can cause catastrophic forgetting, particularly evident in smaller language models (<3B parameters), leading to potential model collapse. Although we explored several mitigation strategies, completely eliminating this issue remains challenging. Secondly, our framework assumes the availability of high-quality initial debate traces; thus, its efficacy may degrade if debates are of poor quality or if initial agent performance is weak.

Moreover, our study primarily focused on structured reasoning tasks like mathematical and commonsense reasoning. The applicability and effectiveness of DTE on less structured or more open-ended tasks, such as natural language generation or dialogue systems, require further investigation. Lastly, although computationally efficient compared to traditional MAD setups, DTE still incurs higher training costs than standard single-model fine-tuning. Future work should aim to optimize the framework further, enhancing its practicality and accessibility.

Ethics Statement

This study explore the self-evolution of language models using publicly available benchmarks and datasets such as GSM8K, ARC, and CommonsenseQA. All data used in our experiments are non-sensitive and freely accessible, ensuring compliance with ethical research standards and reproducibility. Our method involves fine-tuning on model-generated content, without introducing or relying on any human-annotated private data.

AI Assistance: We used ChatGPT assistance for parts of the Appendix, such as generating LaTeX code for tables and refining text written by the authors. All AI-generated content was carefully reviewed and revised by the authors to ensure accuracy and clarity.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

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

Caia Costello, Simon Guo, Anna Goldie, and Azalia Mirhoseini. 2025. Think, prune, train, improve: Scaling reasoning without scaling models. *arXiv preprint arXiv:2504.18116*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. 2024. Acc-debate: An actor-critic approach to multi-agent debate. *arXiv preprint arXiv:2411.00053*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *ArXiv*, abs/2305.11738.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. Self-evolve: A code evolution framework via large language models. *ArXiv*, abs/2306.02907.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding catastrophic forgetting in language models via implicit inference. *Preprint*, arXiv:2309.10105.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.

Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984, Bangkok, Thailand. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Weichao Wang, Xingshan Zeng, Lifeng Shang, and 1 others. 2023. Self: Self-evolution with language feedback. *arXiv preprint arXiv:2310.00533*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *Preprint*, arXiv:2308.08747.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. Predict: Multi-agent-based debate simulation for generalized hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20963–20987.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Lidén, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *ArXiv*, abs/2302.12813.

Leonardo Ranaldi and Andrè Freitas. 2024. Self-refine instruction-tuning for aligning reasoning in language models. *arXiv preprint arXiv:2405.00402*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D Barrett, and Arnu Pretorius. 2023. Should we be going mad? a look at multi-agent debate strategies for llms. *arXiv preprint arXiv:2311.17371*.

Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? In *Annual Meeting of the Association for Computational Linguistics*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. 2024. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In *Proc. the 36th International Conference on Neural Information Processing Systems*, volume 1126.

Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. 2024. Automatic instruction evolving for large language models. *Preprint*, arXiv:2406.00770.

Contents of the Appendix

A Datasets Details

We evaluate our Multi-Agent Debate (MAD) approach on five diverse reasoning benchmarks. In the following, we briefly describe each dataset along with their splits. In this paper, we use the test split to evaluate all Small Language Models (SLMs). Table 5 summarizes the splits for each dataset.

GSM8K (Cobbe et al., 2021) is a collection of high-quality grade school math word problems that require multi-step reasoning. In the `main` configuration, the dataset contains a total of 8,790 examples, with 7,473 examples in the training split and 1,319 examples in the test split. These problems typically require between 2 and 8 steps to solve, making it an excellent benchmark for evaluating mathematical reasoning capabilities.

Dataset	Train	Validation	Test
GSM8K	7,473	–	1,319
GSM+	–	10,552	2,400
ARC-Easy	2,251	570	2,376
ARC-Challenge	1,119	299	1,172
CommonsenseQA	9,741	1,221	1,140

Table 5: Dataset splits and example counts. Note that GSM8K is provided with only training and test splits.

GSM-Plus (Li et al., 2024) extends the GSM8K benchmark with more challenging and diverse mathematical word problems. GSM+ problems generally require more sophisticated multi-step reasoning and often involve more complex mathematical concepts than those in the original GSM8K dataset.

ARC (Clark et al., 2018) comprises two subsets of multiple-choice science questions:

- **ARC-Easy**: Contains 2,251 train, 570 validation, and 2,376 test examples. These questions are answerable by most middle school students and test basic science knowledge.
- **ARC-Challenge**: Contains 1,119 train, 299 validation, and 1,172 test examples. These questions are more challenging and typically answered incorrectly by both retrieval-based algorithms and word co-occurrence algorithms.

CommonsenseQA (Talmor et al., 2019) requires commonsense reasoning to answer multiple-choice questions. It has 9,741 training examples, 1,221 validation examples, and 1,140 test examples. The questions are specifically designed to test commonsense knowledge and reasoning capabilities that go beyond simple factual recall.

B Implementation Details

Training For model fine-tuning, we used GRPO to enhance Language Models on the targeted reasoning tasks. Our training pipeline utilized the Unsloth [1] and TRL [2] libraries for efficient parameter-efficient fine-tuning with QLoRA (Dettmers et al.,

[1] https://github.com/unslothai/unsloth
[2] https://github.com/huggingface/trl

2023). Models were trained with a LoRA (Hu et al., 2021) rank of 128 and target modules including query, key, value, output, gate, up, and down projections. We used 8-bit Adam optimizer with beta parameters of (0.9, 0.99) and a weight decay of 0.1. The learning rate was set to 5e-6 with a cosine decay schedule and 10% warmup ratio. Training proceeded for 10,000 steps with a per-device batch size of 8. To improve output formatting, we implemented a multi-component reward function consisting of: (1) an answer correctness reward, (2) format adherence rewards for XML tags structure, (3) a numeric response reward, and (4) a tag-counting reward to incentivize proper tag usage. Each model was instructed to output responses in a structured XML format with separate <reasoning> and <answer> tags to facilitate consistent answer extraction and evaluation. To manage memory constraints on high-end GPUs, we set the maximum sequence length to 2048 tokens with a 512-token maximum for prompts and 1536 tokens for model completions.

Inference We conducted all model inferences using NVIDIA H100-80GB, A100-80GB, L40-48GB, and A40-48GB GPUs. For efficient inference, we used the vLLM library (Kwon et al., 2023) [3], dynamically allocating the required number of GPUs to load each model. Multi-GPU utilization was enabled using Hugging Face Accelerate [4] for model sharding and speed optimization.

[3] https://docs.vllm.ai/en/latest/

[4] https://github.com/huggingface/accelerate

C REFLECT–CRITIQUE–REFINE Prompt Design

Prompt 1: RCR Prompting for Math Reasoning Datasets (GSM8K, GSM-Plus)

Prompt Template

You are Agent {self.agent_id} in a multi-agent debate to solve the following math problem:
Problem: {question}
{own_previous}
Here are the solutions from other agents: {context}
This is debate round {round_num}. Please carefully analyze all solutions—including your own—identify any errors in reasoning, and provide your revised solution.

- **If you believe your previous answer is correct, explain why and defend it.**
- **If you believe you made an error, explain the error and provide a corrected solution.**
- **If you believe another agent's answer is correct, explain why you agree with it.**

Your final answer must be in the format $\boxed{\{answer\}}$ at the end.

Prompt 2: RCR Prompting for Science Reasoning Datasets (ARC-E, ARC-C)

Prompt Template You are Agent {self.agent_id} in a multi-agent debate to solve the following scientific problem:
Problem: {question}
{own_previous}
Here are the solutions from other agents:
{context}
This is debate round {round_num}. Please carefully analyze all solutions—including your own—identify any misconceptions or flawed scientific reasoning, and provide your revised solution.

- **If you believe your previous answer is correct, explain the scientific principles supporting your answer.**
- **If you believe you made an error, explain the scientific misconception and provide a corrected solution.**
- **If you believe another agent's answer is correct, explain why their scientific reasoning is sound.**

Your final answer must be in the format $\boxed{answer}$ at the end.

Prompt 3: RCR Prompting for Commonsense Reasoning Datasets (CSQA)

Prompt Template You are Agent {self.agent_id} in a multi-agent debate to solve the following commonsense reasoning problem:
Problem: {question}
{own_previous}
Here are the solutions from other agents:
{context}
This is debate round {round_num}. Please carefully analyze all solutions—including your own—identify any flawed assumptions or logical inconsistencies, and provide your revised solution.

- **If you believe your previous answer is correct, explain the logical reasoning and real-world knowledge supporting it.**
- **If you believe you made an error, explain the flawed assumption or inconsistency and provide a corrected solution.**
- **If you believe another agent's answer is correct, explain why their reasoning aligns with commonsense knowledge.**

Your final answer must be in the format $\boxed{answer}$ at the end.

D Additional Self-Evolution Results

In this section, we present a comprehensive analysis of our DEBATE, TRAIN, EVOLVE framework across multiple experimental settings. We first examine the impact of various Guided Reinforcement Policy Optimization (GRPO) configurations, followed by analyses of multi-round training effects, and finally cross-domain generalization results. Our experiments utilize a diverse set of models ranging from 1.5B to 14B parameters and evaluate performance on challenging reasoning benchmarks including GSM8K, GSM-Plus, ARC-Challenge, ARC-Easy, and CommonsenseQA.

D.1 Complete GRPO results (all steps, temperature)

We begin by investigating how different GRPO hyperparameters affect model performance. Tables 6, 7, and 8 present results across three datasets (GSM8K, GSM-Plus, and ARC-Challenge) for six different model configurations, varying training steps (2000, 5000, and 10000) and sampling temperatures (0.8 and 0.2).

Several key patterns emerge from these results. First, we observe that larger models (7B+) generally maintain or improve their performance through GRPO fine-tuning, while smaller models (particularly Llama-3B) occasionally exhibit catastrophic forgetting at higher step counts. Second, lower temperature (0.2) typically yields more stable optimization trajectories for most model configurations, especially at higher step counts. This supports our hypothesis that constraining policy drift during fine-tuning is crucial for successful reasoning evolution.

Notably, the Qwen-2.5-3B model demonstrates remarkable stability across configurations, with consistent performance gains on GSM-Plus (from 61.75% to 69.50%) and robust maintenance of GSM8K performance. In contrast, the Llama-3B model shows significant performance degradation at higher step counts with 0.8 temperature, dropping to near-random performance (2.73%) after 10000 steps on GSM8K, while maintaining better stability at 0.2 temperature.

For ARC-Challenge, we observe that all models benefit from MAD evolution, with particularly strong gains for Qwen-2.5-7B (from 87.22% to 91.64%) and Llama-8B (from 77.65% to 85.07%). These results suggest that our framework effectively generalizes across both mathematical reasoning and scientific question-answering domains.

D.2 Complete Round 2 MAD Results

After the first round of GRPO fine-tuning, we evaluated the performance of models in a multi-agent debate setting to assess how evolution affects collaborative reasoning. Table 10 presents these results across different debate configurations: exponential temperature scaling (Exp), default settings (Default), temperature-4 settings (temp4), and deterministic setting (Det).

The MAD Round 2 results demonstrate that evolved models generally maintain their collaborative reasoning capabilities after GRPO fine-tuning. For most models, MAD performance after evolution either improves or remains comparable to the original MAD results. The Qwen-2.5-7B model, for instance, achieves 77.75% accuracy on GSM-Plus under the temp4 configuration, which represents a 3.58% improvement over its original MAD performance.

Interestingly, we observe that different debate configurations yield varying results across model sizes. Smaller models like Qwen-2.5-1.5B show significant performance variation across configurations, with deterministic settings yielding the best results (69.07% on GSM8K and 56.62% on GSM-Plus). In contrast, larger models like Qwen-2.5-7B demonstrate more consistent performance across configurations.

The exponential temperature scaling configuration generally underperforms other settings, particularly for smaller models. This suggests that controlled diversity in debate is beneficial, but excessive exploration may hinder collaborative reasoning effectiveness.

D.3 GRPO round 2 results

To investigate the effects of iterative evolution, we conducted a second round of GRPO fine-tuning on models that had already undergone one round of evolution. Table 9 presents these results for four model configurations across two datasets (GSM8K and GSM-Plus).

The second round of GRPO training reveals interesting dynamics in model evolution. For the Qwen family of models, we observe continued performance improvements or stability across most configurations. The Qwen-2.5-7B model, for instance, achieves further gains on GSM-Plus, reaching 73.75% accuracy (a 5.13% improvement over its first round GRPO performance).

However, the Llama-3B model exhibits significant performance degradation in certain configurations, particularly at higher step counts with 0.8 temperature (dropping to 35.63% on GSM8K and 23.02% on GSM-Plus). This reinforces our finding that smaller models are more sensitive to optimization instability during iterative fine-tuning. Importantly, using a lower temperature of 0.2 substantially mitigates this issue, allowing the Llama-3B model to maintain competitive performance (73.62% on GSM8K) even after two rounds of evolution.

These results highlight the importance of careful hyperparameter selection during iterative self-evolution, particularly for smaller models that may be more susceptible to catastrophic forgetting or excessive policy drift.

D.4 Complete Round 3 MAD Results

To investigate the long-term stability of collaborative reasoning capabilities through multiple evolution iterations, we conducted a third round of multi-agent debate after the second round of GRPO fine-tuning. Table 11 presents these results for three Qwen models across the same four debate configurations.

The Round 3 MAD results reveal interesting trends in iterative evolution. For the Qwen-2.5-3B and Qwen-2.5-7B models, performance remains relatively stable across debate configurations, indicating robust retention of reasoning capabilities through multiple fine-tuning iterations. However, the Qwen-2.5-1.5B model shows more variable performance, particularly under the exponential temperature scaling configuration where it drops to 44.28% on GSM8K.

Notably, the deterministic debate setting (Det) consistently produces the best or near-best performance across all models and datasets, suggesting that reduced randomness in collaborative reasoning becomes increasingly important after multiple evolution rounds. This aligns with our hypothesis that controlling policy drift is crucial for successful iterative evolution.

The stability of larger models (3B+) across multiple evolution rounds indicates that our DEBATE, TRAIN, EVOLVE framework can support continuous improvement without substantial performance degradation when applied to sufficiently capable base models.

D.5 Complete Cross Domain Task Results

A key question for self-evolution frameworks is whether improvements generalize beyond the training domain. Table 12 presents results for models fine-tuned on either GSM8K or GSM-Plus and evaluated on multiple out-of-domain tasks including ARC-Easy, ARC-Challenge, and CommonsenseQA.

The cross-domain results reveal impressive generalization capabilities. Models fine-tuned on mathematical reasoning tasks (GSM8K and GSM-Plus) show substantial performance improvements not only on the alternative math dataset but also on science and commonsense reasoning benchmarks. For instance, the Qwen-2.5-14B model fine-tuned on GSM8K achieves 98.19% accuracy on ARC-Easy, 93.69% on ARC-Challenge, and 83.70% on CommonsenseQA.

Interestingly, models fine-tuned on GSM-Plus generally perform better on GSM8K than vice versa. For example, the Qwen-2.5-1.5B model achieves 73.09% on GSM8K when fine-tuned on GSM-Plus, but only 51.21% on GSM-Plus when fine-tuned on GSM8K. This asymmetry suggests that GSM-Plus may require more diverse reasoning strategies that transfer well to simpler tasks.

The strong cross-domain performance demonstrates that our DEBATE, TRAIN, EVOLVE framework does not simply optimize for specific datasets but instead enhances fundamental reasoning capabilities that generalize across tasks. This is a critical advantage over traditional supervised fine-tuning approaches that often exhibit limited transferability.

Model	Base Performance		MAD	GRPO (Temperature 0.8)			GRPO (Temperature 0.2)		
	Train	Test		2k steps	5k steps	10k steps	2k steps	5k steps	10k steps
Qwen-2.5-1.5B	81.55	62.77	72.33	67.78	71.42	71.04	73.09	66.49	53.98
Qwen-2.5-3B	91.28	84.08	85.14	85.06	85.14	86.13	84.00	86.05	84.38
Qwen-2.5-7B	94.29	90.67	91.21	88.32	86.73	84.00	86.96	86.35	88.02
Llama-3B	83.90	72.55	73.84	69.22	21.53	2.73	72.40	75.06	3.26
Llama-8B	89.08	81.73	82.18	84.61	85.29	85.22	86.81	84.91	0.15
Qwen-2.5-14B	94.89	92.80	93.33	87.72	89.84	91.81	86.58	89.34	93.74

Table 6: **Complete GRPO Results on GSM8K Dataset.** Results show accuracy (%) for different models under various GRPO configurations. Training hyperparameters include learning rate of 5e-6 and context length of 256 tokens. MAD refers to Multi-Agent Debate baseline performance.

Model	Base Performance		MAD	GRPO (Temperature 0.8)			GRPO (Temperature 0.2)		
	Train	Test		2k steps	5k steps	10k steps	2k steps	5k steps	10k steps
Qwen-2.5-1.5B	42.40	42.00	51.62	47.49	54.46	19.00	52.33	53.04	55.92
Qwen-2.5-3B	61.14	61.75	67.79	66.21	66.71	69.13	64.04	67.25	68.25
Qwen-2.5-7B	68.27	68.62	74.17	64.71	73.38	74.71	67.75	72.54	74.50
Llama-3B	47.68	45.67	51.12	52.38	53.29	52.33	51.79	49.54	53.79
Llama-8B	58.56	55.62	60.79	64.96	61.58	66.17	65.08	63.46	60.46
Qwen-2.5-14B	71.11	71.79	77.25	70.79	73.54	75.88	73.00	73.42	75.62

Table 7: **Complete GRPO Results on GSM-Plus Dataset.** Results show accuracy (%) for different models under various GRPO configurations on the more challenging GSM-Plus dataset. Training hyperparameters include learning rate of 5e-6.

Model	Base Performance		MAD	GRPO (Temperature 0.8)			GRPO (Temperature 0.2)		
	Train	Test		2k steps	5k steps	10k steps	2k steps	5k steps	10k steps
Qwen-2.5-1.5B	—	69.21	68.52	30.03	62.63	68.36	47.27	51.88	67.51
Qwen-2.5-3B	—	83.53	84.64	81.66	80.29	83.63	81.91	79.78	83.95
Qwen-2.5-7B	—	87.22	91.64	88.57	88.48	90.63	88.43	88.57	90.89
Llama-3B	—	73.12	76.19	75.51	74.32	76.87	76.79	74.57	77.23
Llama-8B	—	77.65	85.07	83.70	84.45	86.03	84.98	85.53	86.53
Qwen-2.5-14B	—	90.27	93.77	91.81	92.49	93.13	91.47	91.47	92.67

Table 8: **Complete GRPO Results on ARC-Challenge Dataset.** Results show accuracy (%) for different models under various GRPO configurations on the ARC-Challenge dataset. Training hyperparameters include learning rate of 5e-6 and context length of 128 tokens. Base train performance was not evaluated for this dataset.

Model	Dataset	GRPO Round 2 (Temp 0.8)		GRPO Round 2 (Temp 0.2)	
		2k steps	5k steps	2k steps	5k steps
Qwen-2.5-1.5B	GSM8K	65.73	68.54	69.98	72.18
	GSM-Plus	47.38	50.12	46.37	48.04
Qwen-2.5-3B	GSM8K	84.84	86.05	84.46	84.08
	GSM-Plus	65.71	67.96	65.67	67.00
Qwen-2.5-7B	GSM8K	86.28	87.19	88.17	87.34
	GSM-Plus	69.42	73.75	70.54	73.12
Llama-3B	GSM8K	55.88	35.63	73.62	64.29
	GSM-Plus	48.75	23.02	52.42	25.08

Table 9: **Complete GRPO Round 2 Results.** Results show accuracy (%) after second round of GRPO training across different step counts and temperature settings. All models were trained with learning rate of 5e-6 and context length of 128 tokens.

Model	Dataset	MAD Configuration			
		Exp	Default	temp4	Det
Qwen-2.5-1.5B	GSM8K	46.32	66.34	68.61	69.07
	GSM-Plus	22.09	53.18	55.62	56.62
Qwen-2.5-3B	GSM8K	84.08	86.66	86.35	86.50
	GSM-Plus	69.62	70.25	69.67	70.29
Qwen-2.5-7B	GSM8K	91.36	90.75	91.05	89.99
	GSM-Plus	76.42	77.00	77.75	77.62
Llama-3B	GSM8K	66.26	75.97	75.51	75.36
	GSM-Plus	53.62	54.58	55.96	56.04
Llama-8B	GSM8K	84.69	85.90	86.96	85.60
	GSM-Plus	65.00	65.92	66.46	66.50

Table 10: **Complete MAD Round 2 Results.** Results show accuracy (%) for different models in multi-agent debate after first round of GRPO fine-tuning. Exp = exponential temperature scaling, Default = standard configuration, temp4 = temperature-4 settings, Det = deterministic configuration.

Model	Dataset	MAD Configuration			
		Exp	Default	temp4	Det
Qwen-2.5-1.5B	GSM8K	44.28	60.65	67.70	72.40
	GSM-Plus	35.54	48.62	51.67	51.75
Qwen-2.5-3B	GSM8K	83.78	85.60	85.75	86.13
	GSM-Plus	63.67	63.42	64.16	64.47
Qwen-2.5-7B	GSM8K	89.76	91.05	90.90	91.13
	GSM-Plus	69.67	69.85	70.50	69.88

Table 11: **Complete MAD Round 3 Results.** Results show accuracy (%) for different models in multi-agent debate after second round of GRPO fine-tuning. Exp = exponential temperature scaling, Default = standard configuration, temp4 = temperature-4 settings, Det = deterministic configuration.

Model	Fine-tuned on	Evaluation Dataset				
		GSM8K	GSM-Plus	ARC-Easy	ARC-Challenge	CommonsenseQA
Qwen-2.5-1.5B	GSM8K	—	51.21	85.02	69.88	64.29
	GSM-Plus	73.09	—	85.10	69.45	64.21
Qwen-2.5-3B	GSM8K	—	65.54	93.94	84.30	75.92
	GSM-Plus	86.50	—	94.15	84.13	75.92
Qwen-2.5-7B	GSM8K	—	69.63	96.42	91.72	82.96
	GSM-Plus	91.81	—	96.38	90.87	82.88
Llama-3B	GSM8K	—	52.38	87.12	72.01	68.14
	GSM-Plus	76.35	—	86.57	69.20	68.55
Llama-8B	GSM8K	—	63.75	93.01	84.39	74.12
	GSM-Plus	86.88	—	93.98	85.49	73.87
Qwen-2.5-14B	GSM8K	—	73.46	98.19	93.69	83.70
	GSM-Plus	93.33	—	97.98	94.28	82.23

Table 12: **Complete Cross Domain Task Results.** Results show accuracy (%) on various datasets after fine-tuning on either GSM8K or GSM-Plus. Dashes (—) indicate that evaluation was not performed on the same dataset used for fine-tuning.

E Complete Results of Large-scale Empirical Study on MAD using RCR Prompting

This section presents a comprehensive analysis of our large-scale empirical investigation into Multi-Agent Debate (MAD) using Recursive Critical Reflection (RCR) prompting across five diverse benchmarks: GSM8K, GSM-Plus, ARC-Easy, ARC-Challenge, and CommonsenseQA. Through extensive experimentation involving various model combinations and parameter settings, we evaluate how collaborative reasoning among multiple language model agents affects problem-solving performance.

E.1 Evaluation Metrics and Methodology

To facilitate systematic comparison and analysis of debate outcomes, we track the following key metrics across all debate configurations:

- **Accuracy**: The primary performance measure, representing the percentage of problems correctly solved after the debate process concludes.
- **Δ (Performance Delta)**: Measures the performance change relative to appropriate baselines. We report several variants including:
  - Δ (vs Base): Change compared to the single agent's performance
  - Δ (vs Lower Agent): Change compared to the lower-performing agent in cross-agent debates
  - Δ (vs Upper Agent): Change compared to the better-performing agent in cross-agent debates
  - Δ (vs Lowest): Change compared to the lowest-performing agent in three-agent settings
- **Debate Rounds**: The average number of interaction rounds required to reach consensus or the maximum allowed limit, indicating debate efficiency.
- **Sycophancy**: A normalized measure (per data points) quantifying the tendency of agents to abandon their answers in favor of matching another agent's previous response, providing insights into social influence dynamics.
- **State Transitions**: Tracked as C$\rightarrow$I (correct to incorrect) and I$\rightarrow$C (incorrect to correct) counts, these reveal the qualitative nature of answer changes during debate.
- **Debate Helped**: The overall count of instances where the debate process improved the final outcome compared to initial responses.

Our evaluation spans multiple dimensions of agent configuration:

- **Agent Settings**: We systematically vary temperature parameter across four settings:
  - Default: Balanced temperature
  - Deterministic (Det.): Lower temperature for more consistent outputs
  - Exploratory (Exp.): Higher temperature for more diverse responses
  - Mixed: Combinations of the above settings across different agents
- **Debate Structures**: We investigate four primary debate configurations:
  - Single-Model Debate: Multiple instances of the same model with varied parameter settings
  - Cross-Agent Debate: Two different models debating with various parameter settings
  - Three Identical Agents: Three instances of the same model with potentially different settings
  - Three Varied Agents: Three different models engaging in debate

E.2 Overview of Results Organization

Our extensive experimental results are organized in Tables 13-32, systematically covering all five datasets with the four debate configurations described above. For each dataset, we present:

- Table set 1 (Tables 13-16): Performance on GSM8K
- Table set 2 (Tables 17-20): Performance on GSM-Plus
- Table set 3 (Tables 21-24): Performance on ARC-Easy
- Table set 4 (Tables 25-28): Performance on ARC-Challenge
- Table set 5 (Tables 29-32): Performance on CommonsenseQA

E.3 Key Findings and Patterns

E.3.1 Impact of Agent Settings

Our analysis reveals that agent parameter settings significantly influence debate outcomes across all datasets. We observe that while the Default setting provides reliable performance, Exploratory settings often lead to higher variance in outcomes, sometimes yielding exceptional improvements but also risking performance degradation. The Deterministic setting generally produces more consistent but potentially conservative results.

The sycophancy metric proves particularly informative, showing higher values in debates between models with substantial performance gaps. This suggests that lower-performing models tend to defer to higher-performing ones, which can be either beneficial or detrimental depending on the initial state distribution.

E.3.2 Cross-Model Debate Dynamics

In cross-agent debates (Tables 10-14), we find that pairing models with complementary strengths often produces synergistic effects. The Δ metrics relative to both upper and lower agents reveal important patterns: when a high-performing model debates with a weaker one, the debate outcome typically falls between their individual performances but closer to the stronger model's baseline.

State transitions (C→I and I→C) provide valuable insights into debate quality. A high I→C rate coupled with a low C→I rate indicates constructive debate where correct reasoning prevails, while the opposite pattern signals problematic dynamics where convincing but incorrect reasoning dominates.

E.3.3 Three-Agent Debate Effectiveness

The introduction of a third agent creates more complex interaction patterns. Three-agent debates consistently show lower sycophancy rates compared to two-agent settings, suggesting that the presence of multiple perspectives reduces blind conformity. When all three agents are identical, we observe that diversity in parameter settings typically outperforms homogeneous settings.

In three varied agent debates, we find particularly interesting results when combining models of different sizes and architectures. As shown in Table 16, certain combinations like "Qwen-2.5-3B + Phi-mini-3.8B + Llama-3.1-3B" achieve accuracy improvements even compared to the highest-performing individual agent, suggesting effective complementarity between these models' reasoning approaches.

E.3.4 Dataset-Specific Patterns

Our results indicate substantial variation in debate effectiveness across different datasets:

- **GSM8K and GSM+**: Harder Mathematical reasoning tasks (GSM-Plus) show the most consistent benefits from debate, with average debate rounds typically higher than other datasets, suggesting that step-by-step verification is particularly valuable for these problems.
- **ARC-Easy and ARC-Challenge**: Multiple-choice science questions reveal interesting patterns where sycophancy is generally lower, but debate can still improve performance when appropriately configured.
- **CommonsenseQA**: This dataset exhibits unique characteristics where debates tend to conclude more quickly, suggesting that commonsense reasoning may be less amenable to explicit verification through debate.

E.4 Conclusion

Tables 13-32 collectively present a comprehensive empirical foundation for understanding the effects of Multi-Agent Debate using RCR prompting across diverse reasoning tasks. The metrics reveal nuanced patterns in how debate influences performance, with clear evidence that appropriate configuration of debate participants and settings can yield substantial improvements over single-agent performance.

The consistent tracking of accuracy, deltas, debate rounds, sycophancy, and state transitions provides a multi-dimensional view of debate quality beyond simple performance measures. These results demonstrate that MAD is not universally beneficial but rather depends critically on the specific combination of models, parameter settings, and problem domains. Our findings establish an important baseline for future research on collaborative reasoning between language models, highlighting both the potential and the challenges of multi-agent approaches to complex problem-solving.

Agent 1	Agent 2	Agent Settings	MAD Accuracy (RCR Prompting)	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 1319)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Default	47.38	5.38 ↑	1.60	1.17	156.00	251	220
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Deterministic	47.31	5.31 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Exploratory	39.20	2.8 ↓	2.19	1.25	185.00	274	234
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Det. & Exp.	43.14	1.14 ↑	1.89	1.09	185.00	262	226
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Default	70.89	8.12 ↑	0.86	0.70	101.00	352	317
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Deterministic	63.46	0.69 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Exploratory	71.57	8.8 ↑	1.05	0.84	94.00	449	399
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Det. & Exp.	72.33	9.56 ↑	0.98	0.71	99.00	423	377
Qwen-2.5-3B	Qwen-2.5-3B	Both: Default	86.05	0.91 ↑	0.31	0.21	55.00	115	104
Qwen-2.5-3B	Qwen-2.5-3B	Both: Deterministic	84.99	0.15 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Both: Exploratory	85.52	0.38 ↑	0.35	0.26	62.00	116	103
Qwen-2.5-3B	Qwen-2.5-3B	Both: Det. & Exp.	86.28	1.14 ↑	0.34	0.19	50.00	106	101
Qwen-2.5-7B	Qwen-2.5-7B	Both: Default	91.74	1.07 ↑	0.16	0.13	28.00	53	49
Qwen-2.5-7B	Qwen-2.5-7B	Both: Deterministic	90.60	0.07 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Both: Exploratory	91.21	0.54 ↑	0.18	0.15	27.00	59	57
Qwen-2.5-7B	Qwen-2.5-7B	Both: Det. & Exp.	91.51	0.84 ↑	0.18	0.15	33.00	57	55
Qwen-2.5-14B	Qwen-2.5-14B	Both: Default	93.48	0.68 ↑	0.11	0.13	22.00	46	43
Qwen-2.5-14B	Qwen-2.5-14B	Both: Deterministic	93.18	0.38 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-14B	Qwen-2.5-14B	Both: Exploratory	93.33	0.53 ↑	0.11	0.12	20.00	48	48
Qwen-2.5-14B	Qwen-2.5-14B	Both: Det. & Exp.	93.63	0.83 ↑	0.13	0.15	24.00	44	39
Qwen-2.5-32B	Qwen-2.5-32B	Both: Default	95.00	0.08 ↑	0.05	0.06	11.00	21	20
Qwen-2.5-32B	Qwen-2.5-32B	Both: Deterministic	94.77	0.15 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Both: Exploratory	95.38	0.46 ↑	0.07	0.08	9.00	32	31
Qwen-2.5-32B	Qwen-2.5-32B	Both: Det. & Exp.	95.30	0.38	0.04	0.05	12.00	23	21
Llama-3.1-3B	Llama-3.1-3B	Both: Default	74.91	2.36 ↑	0.73	0.49	106.00	208	183
Llama-3.1-3B	Llama-3.1-3B	Both: Deterministic	74.37	1.82 ↑	0.00	0.00	0.00	0	0
Llama-3.1-3B	Llama-3.1-3B	Both: Exploratory	72.40	0.15 ↓	0.94	0.57	138.00	225	202
Llama-3.1-3B	Llama-3.1-3B	Both: Det. & Exp.	73.84	1.29 ↑	0.80	0.48	133.00	193	175
Llama-3.1-8B	Llama-3.1-8B	Both: Default	82.56	0.83 ↑	0.48	0.38	86.00	116	105
Llama-3.1-8B	Llama-3.1-8B	Both: Deterministic	81.50	0.23 ↓	0.00	0.00	0.00	0	0
Llama-3.1-8B	Llama-3.1-8B	Both: Exploratory	80.67	1.06 ↓	0.60	0.40	98.00	162	149
Llama-3.1-8B	Llama-3.1-8B	Both: Det. & Exp.	82.18	0.45 ↑	0.56	0.39	97.00	142	126
Phi-mini-3.8B	Phi-mini-3.8B	Both: Default	87.72	0.84 ↑	0.29	0.27	51.00	101	95
Phi-mini-3.8B	Phi-mini-3.8B	Both: Deterministic	86.73	0.15 ↓	0.02	0.00	0.00	2	1
Phi-mini-3.8B	Phi-mini-3.8B	Both: Exploratory	87.95	1.07 ↑	0.30	0.26	48.00	112	99
Phi-mini-3.8B	Phi-mini-3.8B	Both: Det. & Exp.	87.34	0.46 ↑	0.33	0.26	62.00	103	95
Mistral-7B	Mistral-7B	Both: Default	33.74	12.36 ↑	1.65	0.73	101.00	454	340
Mistral-7B	Mistral-7B	Both: Deterministic	20.02	1.36	0.04	0.00	0.00	0	0
Mistral-7B	Mistral-7B	Both: Exploratory	35.71	14.33 ↑	1.85	0.80	110.00	509	381
Mistral-7B	Mistral-7B	Both: Det. & Exp.	33.51	12.13	1.53	0.68	97.00	433	334

Table 13: Performance in Multi-Agent Debate Settings on the **GSM8K** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters like Default, Deterministic, Exploratory, and a combination) on the **MAD Accuracy (RCR Prompting)** of various language models. The Δ column quantifies the **improvement (or decline) over the single base model performance**. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 1319 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent Settings	Accuracy	Δ (Lower Agent)	Δ (Upper Agent)	Debate Rounds (Avg)	Sycophancy (Avg / 1319)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	1: Default & 2: Default	62.40	20.4 ↑	0.37 ↓	1.52	0.96	168.00	434	387
Qwen-2.5-0.5B	Qwen-2.5-1.5B	1: Det. & 2: Det.	62.32	20.32 ↑	0.45 ↓	1.27	0.72	155.00	357	323
Qwen-2.5-0.5B	Qwen-2.5-1.5B	1: Exp. & 2: Exp.	58.91	16.91 ↑	3.86 ↓	1.95	1.03	175.00	531	448
Qwen-2.5-0.5B	Qwen-2.5-1.5B	1: Det. & 2: Exp.	60.88	18.88 ↑	1.89 ↓	1.54	0.83	147.00	416	344
Qwen-2.5-0.5B	Qwen-2.5-1.5B	1: Exp. & 2: Det.	61.18	19.18 ↑	1.59 ↓	1.67	0.87	164.00	474	425
Qwen-2.5-1.5B	Llama-3.1-3B	1: Default & 2: Default	76.42	13.65 ↑	3.87 ↑	1.09	0.56	107.00	388	342
Qwen-2.5-1.5B	Llama-3.1-3B	1: Det. & 2: Det.	75.59	12.82 ↑	3.04 ↑	1.14	0.36	93.00	285	258
Qwen-2.5-1.5B	Llama-3.1-3B	1: Exp. & 2: Exp.	76.57	13.8 ↑	4.02 ↑	1.17	0.65	96.00	416	355
Qwen-2.5-1.5B	Llama-3.1-3B	1: Det. & 2: Exp.	75.06	12.29 ↑	2.51 ↑	1.22	0.48	111.00	362	326
Qwen-2.5-1.5B	Llama-3.1-3B	1: Exp. & 2: Det.	76.04	13.27 ↑	3.49 ↑	1.12	0.59	129.00	383	331
Qwen-2.5-3B	Phi-mini-3.8B	1: Default & 2: Default	87.41	2.27 ↑	0.53 ↑	0.39	0.22	53.00	128	114
Qwen-2.5-3B	Phi-mini-3.8B	1: Det. & 2: Det.	85.97	0.83 ↑	0.91 ↓	0.43	0.17	74.00	82	72
Qwen-2.5-3B	Phi-mini-3.8B	1: Exp. & 2: Exp.	88.63	3.49 ↑	1.75 ↑	0.44	0.27	46.00	155	142
Qwen-2.5-3B	Phi-mini-3.8B	1: Det. & 2: Exp.	86.73	1.59 ↑	0.15 ↓	0.40	0.20	63.00	105	99
Qwen-2.5-3B	Phi-mini-3.8B	1: Exp. & 2: Det.	88.10	2.96 ↑	1.22 ↑	0.41	0.23	57.00	135	126
Qwen-2.5-1.5B	Qwen-2.5-3B	1: Default & 2: Default	82.71	19.94 ↑	2.43 ↓	0.71	0.51	67.00	370	359
Qwen-2.5-1.5B	Qwen-2.5-3B	1: Det. & 2: Det.	81.27	18.5 ↑	3.87 ↓	0.62	0.48	94.00	284	275
Qwen-2.5-1.5B	Qwen-2.5-3B	1: Exp. & 2: Exp.	83.17	20.4 ↑	1.97 ↓	0.80	0.56	68.00	414	392
Qwen-2.5-1.5B	Qwen-2.5-3B	1: Det. & 2: Exp.	82.87	20.1 ↑	2.27 ↓	0.76	0.48	74.00	328	310
Qwen-2.5-1.5B	Qwen-2.5-3B	1: Exp. & 2: Det.	82.26	19.49 ↑	2.88 ↓	0.75	0.52	82.00	384	372
Llama-3.1-3B	Llama-3.1-8B	1: Default & 2: Default	78.54	5.99 ↑	3.19 ↓	0.77	0.51	122.00	213	195
Llama-3.1-3B	Llama-3.1-8B	1: Det. & 2: Det.	79.23	6.68 ↑	2.5 ↓	0.68	0.48	130.00	159	143
Llama-3.1-3B	Llama-3.1-8B	1: Exp. & 2: Exp.	77.10	4.55 ↑	4.63 ↓	0.93	0.58	127.00	238	224
Llama-3.1-3B	Llama-3.1-8B	1: Det. & 2: Exp.	79.83	7.28 ↑	1.9 ↓	0.81	0.45	123.00	211	183
Llama-3.1-3B	Llama-3.1-8B	1: Exp. & 2: Det.	77.18	4.63 ↑	4.55 ↓	0.87	0.56	141.00	183	173
Qwen-2.5-7B	Qwen-2.5-14B	1: Default & 2: Default	92.19	1.52 ↑	0.61 ↓	0.16	0.13	39.00	63	61
Qwen-2.5-7B	Qwen-2.5-14B	1: Det. & 2: Det.	92.04	1.37 ↑	0.76 ↓	0.17	0.13	47.00	53	50
Qwen-2.5-7B	Qwen-2.5-14B	1: Exp. & 2: Exp.	93.10	2.43 ↑	0.3 ↑	0.16	0.15	33.00	72	68
Qwen-2.5-7B	Qwen-2.5-14B	1: Det. & 2: Exp.	92.19	1.52 ↑	0.61 ↓	0.15	0.11	37.00	58	58
Qwen-2.5-7B	Qwen-2.5-14B	1: Exp. & 2: Det.	92.80	2.13 ↑	0.00	0.17	0.16	39.00	64	60

Table 14: Performance Analysis of Cross-Agent Debates on the **GSM8K** Dataset. This table details the outcomes of debates between different language models (Agent 1 and Agent 2). **Agent Settings** specify the configuration (e.g., Default, Deterministic (Det.), Exploratory (Exp.)) applied to Agent 1 and Agent 2 respectively, influencing temperature and top_p parameters. The table presents overall **Accuracy**, along with **Δ (Lower Agent)** and **Δ (Upper Agent)** indicating the performance change for each agent relative to a baseline. Additional metrics include average **Debate Rounds**, normalized **Sycophancy** (per 1319 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C) to show debate impact.

Agent 1	Agent 2	Agent 3	Agent Settings	Accuracy	Δ (Improvement)	Debate Rounds (Avg)	Sycophancy (Avg / 1319)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	All: Default	41.70	0.3 ↓	2.77	3.17	414.00	393.00	236.00
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	All: Deterministic	47.31	5.31 ↑	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	All: Exploratory	36.09	5.91 ↓	3.47	3.33	438.00	450.00	282.00
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	1 Det, 2 Exp	38.36	3.64 ↓	3.13	2.90	412.00	370.00	246.00
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	2 Det, 1 Exp	43.06	1.06 ↑	1.97	1.42	306.00	300.00	211.00
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	All: Default	72.48	9.71 ↑	1.35	1.64	193.00	652.00	469.00
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	All: Deterministic	63.99	1.22 ↑	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	All: Exploratory	75.13	12.36 ↑	1.57	1.82	181.00	796.00	547.00
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	1 Det, 2 Exp	74.83	12.06 ↑	1.51	1.71	170.00	741.00	534.00
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	2 Det, 1 Exp	72.25	9.48 ↑	0.97	1.03	131.00	510.00	329.00
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	All: Default	86.96	1.82 ↑	0.49	0.52	85.00	191.00	147.00
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	All: Deterministic	84.99	0.15 ↓	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	All: Exploratory	87.64	2.5 ↑	0.60	0.65	85.00	256.00	200.00
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	1 Det, 2 Exp	86.73	1.59 ↑	0.63	0.56	110.00	236.00	179.00
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	2 Det, 1 Exp	86.05	0.91 ↑	0.40	0.32	75.00	130.00	99.00
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	All: Default	93.03	2.36 ↑	0.22	0.22	33.00	110.00	88.00
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	All: Deterministic	90.60	0.07 ↓	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	All: Exploratory	92.42	1.75 ↑	0.24	0.24	52.00	110.00	87.00
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	1 Det, 2 Exp	92.12	1.45 ↑	0.24	0.24	44.00	106.00	86.00
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	2 Det, 1 Exp	91.96	1.29 ↑	0.17	0.17	28.00	76.00	52.00
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	All: Default	94.09	1.29	0.11	0.13	18.00	67.00	59.00
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	All: Deterministic	92.95	0.15	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	All: Exploratory	94.24	1.44	0.14	0.16	26.00	88.00	78.00
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	1 Det, 2 Exp	94.31	1.51	0.13	0.16	17.00	81.00	68.00
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	2 Det, 1 Exp	92.87	0.07	0.09	0.08	30.00	33.00	29.00
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	All: Default	95.30	0.38	0.07	0.07	18.00	44.00	39.00
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	All: Deterministic	94.77	0.15	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	All: Exploratory	94.84	0.08	0.08	0.09	21.00	51.00	47.00
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	1 Det, 2 Exp	95.30	0.38	0.07	0.07	16.00	49.00	41.00
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	2 Det, 1 Exp	95.22	0.30	0.05	0.05	11.00	34.00	24.00
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	All: Default	88.40	1.52 ↑	0.42	0.55	86.00	168.00	129.00
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	All: Deterministic	86.66	0.22 ↓	0.01	0.01	0.00	0.00	0.00
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	All: Exploratory	88.10	1.22 ↑	0.48	0.59	99.00	197.00	145.00
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	1 Det, 2 Exp	87.87	0.99 ↑	0.46	0.53	95.00	178.00	132.00
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	2 Det, 1 Exp	87.72	0.84 ↑	0.32	0.41	64.00	121.00	80.00
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	All: Default	72.63	0.08 ↑	1.29	1.29	265.00	317.00	238.00
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	All: Deterministic	73.16	0.61 ↑	0.00	0.00	0.00	0.00	0.00
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	All: Exploratory	72.78	0.23 ↑	1.49	1.39	246.00	414.00	312.00
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	1 Det, 2 Exp	73.69	1.14 ↑	1.39	1.28	251.00	407.00	283.00
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	2 Det, 1 Exp	72.93	0.38 ↑	1.08	0.87	203.00	229.00	147.00
Mistral-7B	Mistral-7B	Mistral-7B	All: Default	37.83	16.45 ↑	2.37	1.97	203.00	894.00	454.00
Mistral-7B	Mistral-7B	Mistral-7B	All: Deterministic	20.02	1.36 ↓	0.04	0.00	0.00	0.00	0.00
Mistral-7B	Mistral-7B	Mistral-7B	All: Exploratory	39.27	17.89 ↑	2.81	2.30	189.00	904.00	480.00
Mistral-7B	Mistral-7B	Mistral-7B	1 Det, 2 Exp	38.89	17.51 ↑	2.61	2.13	222.00	940.00	476.00
Mistral-7B	Mistral-7B	Mistral-7B	2 Det, 1 Exp	35.33	13.95 ↑	1.82	1.39	135.00	694.00	360.00
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	All: Default	84.23	2.5 ↑	0.72	0.82	135.00	429.00	192.00
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	All: Deterministic	81.50	0.23 ↓	0.00	0.00	0.00	0.00	0.00
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	All: Exploratory	83.70	1.97 ↑	0.88	0.89	162.00	310.00	230.00
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	1 Det, 2 Exp	83.32	1.59 ↑	0.86	0.86	160.00	284.00	211.00
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	2 Det, 1 Exp	82.26	0.53 ↑	0.67	0.63	129.00	199.00	132.00

Table 15: Performance Analysis of Three Identical Agents Debating on GSM8K. This table shows results when three instances of the same model (**Agent 1**, **Agent 2**, **Agent 3** being identical) engage in a debate. **Agent Settings** describe the configuration mix across these three agents (e.g., All Default, or a mix like 1 Deterministic (Det), 2 Exploratory (Exp)). **Accuracy** is the debate outcome, and Δ **(Improvement)** is the change from the single agent's baseline. Standard metrics like **Debate Rounds**, normalized **Sycophancy** (per 1319 data points), and error transition rates (C→I, I→C) are also included.

Agent 1	Agent 2	Agent 3	Agent Settings	Accuracy	Δ (vs Lowest)	Debate Rounds (Avg)	Sycophancy (Avg / 1319)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-3B	All: Default	80.82	4.32 ↓	1.81	1.58	154.00	859.00	639.00
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Llama-3.1-3B	All: Default	69.52	3.03 ↓	2.43	1.76	271.00	718.00	508.00
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Phi-mini-3.8B	All: Default	76.04	10.84 ↓	2.20	1.47	267.00	727.00	532.00
Qwen-2.5-0.5B	Qwen-2.5-3B	Llama-3.1-3B	All: Default	79.15	5.99 ↓	2.10	1.36	184.00	696.00	536.00
Qwen-2.5-0.5B	Qwen-2.5-3B	Phi-mini-3.8B	All: Default	83.62	3.24 ↓	1.82	1.08	150.00	618.00	534.00
Qwen-2.5-0.5B	Llama-3.1-3B	Phi-mini-3.8B	All: Default	76.57	10.31 ↓	2.39	1.16	255.00	515.00	402.00
Qwen-2.5-1.5B	Qwen-2.5-3B	Llama-3.1-3B	All: Default	82.71	2.43 ↓	1.24	1.06	156.00	544.00	436.00
Qwen-2.5-1.5B	Qwen-2.5-3B	Phi-mini-3.8B	All: Default	85.22	1.66 ↓	1.08	0.85	139.00	460.00	388.00
Qwen-2.5-1.5B	Llama-3.1-3B	Phi-mini-3.8B	All: Default	81.20	5.68 ↓	1.33	1.05	196.00	560.00	446.00
Qwen-2.5-3B	Phi-mini-3.8B	Llama-3.1-3B	All: Default	86.96	0.08 ↑	0.89	0.71	127.00	372.00	297.00
Qwen-2.5-3B	Qwen-2.5-3B	Phi-mini-3.8B	All: Default	87.64	0.76 ↑	0.60	0.55	97.00	227.00	175.00
Qwen-2.5-3B	Phi-mini-3.8B	Phi-mini-3.8B	All: Default	87.79	0.91 ↑	0.58	0.53	111.00	209.00	167.00
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	All: Default	68.46	5.69 ↑	2.10	2.09	221.00	795.00	570.00
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-1.5B	All: Default	55.12	7.65 ↓	2.60	2.52	364.00	628.00	407.00

Table 16: Performance Analysis of Three-Agent Debates (Varied Models) on **GSM8K**. This table presents outcomes from debates involving three potentially different language models (**Agent 1**, **Agent 2**, **Agent 3**). All debates use default agent settings. The Δ **(vs Lowest)** column indicates the performance change of the debate outcome (Accuracy) compared to the baseline performance of the lowest-performing agent among the three in that specific debate. Standard metrics like **Debate Rounds**, normalized **Sycophancy** (per 1319 data points), and error transition rates (C→I, I→C) are also included.

Agent 1	Agent 2	Agent Settings	MAD Accuracy (RCR Prompting)	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 2400)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Default	27.33	2.54 ↑	2.00	1.51	248.00	348	295
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Deterministic	29.25	4.46 ↑	0.02	0.00	0.00	2	1
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Exploratory	23.12	1.67 ↓	2.56	1.43	284.00	351	289
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Det. & Exp.	27.33	2.54 ↑	2.26	1.33	267.00	396	336
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Default	53.12	11.12 ↑	1.14	0.91	210.00	555	502
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Deterministic	47.29	5.29 ↑	0.03	0.00	0.00	0	0
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Exploratory	51.62	9.62 ↑	1.40	1.08	218.00	647	551
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Det. & Exp.	52.29	10.29 ↑	1.17	0.85	181.00	528	477
Qwen-2.5-3B	Qwen-2.5-3B	Both: Default	67.42	5.67 ↑	0.62	0.39	133.00	225	213
Qwen-2.5-3B	Qwen-2.5-3B	Both: Deterministic	67.38	5.63 ↑	0.05	0.00	0.00	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Both: Exploratory	67.79	6.04 ↑	0.69	0.46	132.00	296	265
Qwen-2.5-3B	Qwen-2.5-3B	Both: Det. & Exp.	66.46	4.71 ↑	0.67	0.36	163.00	223	208
Qwen-2.5-7B	Qwen-2.5-7B	Both: Default	74.17	5.55 ↑	0.35	0.26	62.00	135	127
Qwen-2.5-7B	Qwen-2.5-7B	Both: Deterministic	73.62	5.00 ↑	0.04	0.00	0.00	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Both: Exploratory	74.17	5.55 ↑	0.39	0.30	88.00	158	150
Qwen-2.5-7B	Qwen-2.5-7B	Both: Det. & Exp.	74.46	5.84 ↑	0.33	0.25	78.00	126	118
Qwen-2.5-14B	Qwen-2.5-14B	Both: Default	77.21	5.42 ↑	0.32	0.32	47.00	102	100
Qwen-2.5-14B	Qwen-2.5-14B	Both: Deterministic	76.25	4.46 ↑	0.06	0.00	0.00	0	0
Qwen-2.5-14B	Qwen-2.5-14B	Both: Exploratory	77.25	5.46 ↑	0.33	0.32	45.00	128	123
Qwen-2.5-14B	Qwen-2.5-14B	Both: Det. & Exp.	76.96	5.17 ↑	0.31	0.29	48.00	99	93
Qwen-2.5-32B	Qwen-2.5-32B	Both: Default	73.33	0.87 ↑	0.24	0.19	29.00	62	59
Qwen-2.5-32B	Qwen-2.5-32B	Both: Deterministic	72.79	0.33 ↑	0.08	0.00	0.00	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Both: Exploratory	73.42	0.96 ↑	0.27	0.23	32.00	91	88
Qwen-2.5-32B	Qwen-2.5-32B	Both: Det. & Exp.	73.46	1.00 ↑	0.26	0.19	26.00	70	68
Phi-mini-3.8B	Phi-mini-3.8B	Both: Default	69.62	6.20 ↑	0.60	0.47	113.00	204	191
Phi-mini-3.8B	Phi-mini-3.8B	Both: Deterministic	69.21	5.79 ↑	0.13	0.02	0.00	6	3
Phi-mini-3.8B	Phi-mini-3.8B	Both: Exploratory	70.38	6.96 ↑	0.67	0.50	117.00	267	242
Phi-mini-3.8B	Phi-mini-3.8B	Both: Det. & Exp.	69.42	6.00 ↑	0.62	0.45	114.00	203	188
Mistral-7B	Mistral-7B	Both: Default	23.42	8.38 ↑	1.91	0.77	159.00	576	434
Mistral-7B	Mistral-7B	Both: Deterministic	14.33	0.71 ↓	0.15	0.01	0.00	4	2
Mistral-7B	Mistral-7B	Both: Exploratory	23.29	8.25 ↑	2.13	0.85	149.00	586	437
Mistral-7B	Mistral-7B	Both: Det. & Exp.	22.75	7.71 ↑	1.93	0.77	147.00	556	414
Llama-3.1-3B	Llama-3.1-3B	Both: Default	51.58	5.91 ↑	1.20	0.82	232.00	439	378
Llama-3.1-3B	Llama-3.1-3B	Both: Deterministic	50.50	4.83 ↑	0.01	0.00	0.00	0	0
Llama-3.1-3B	Llama-3.1-3B	Both: Exploratory	51.12	5.45 ↑	1.47	0.87	233.00	482	406
Llama-3.1-3B	Llama-3.1-3B	Both: Det. & Exp.	50.75	5.08 ↑	1.28	0.74	218.00	381	333
Llama-3.1-8B	Llama-3.1-8B	Both: Default	62.04	6.42 ↑	0.95	0.72	202.00	313	274
Llama-3.1-8B	Llama-3.1-8B	Both: Deterministic	61.04	5.42 ↑	0.00	0.00	0.00	0	0
Llama-3.1-8B	Llama-3.1-8B	Both: Exploratory	60.79	5.17 ↑	1.12	0.77	197.00	340	303
Llama-3.1-8B	Llama-3.1-8B	Both: Det. & Exp.	60.96	5.34 ↑	1.01	0.72	214.00	304	273

Table 17: Comparative Analysis of Language Model Performance in Multi-Agent Debate Settings on the **GSM-Plus** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters like Default, Deterministic, Exploratory, and a combination) on the **MAD Accuracy (RCR Prompting)** of various language models. The Δ column quantifies the **improvement (or decline) over the single base model performance**. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 2400 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent Settings	MAD Accuracy	Δ Lower	Δ Upper	Debate Rounds (Avg)	Sycophancy (Avg / 2400)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Default	41.38	16.59 ↑	0.62 ↓	1.85	1.12	314	628	548
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Deterministic	42.67	17.88 ↑	0.67 ↑	1.58	0.89	292	565	505
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Exploratory	39.54	14.75 ↑	2.46 ↓	2.30	1.20	320	722	604
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Det. & Exp.	40.04	15.25 ↑	1.96 ↓	1.97	1.04	301	588	492
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Exp. & Det.	44.25	19.46 ↑	2.25 ↑	2.00	1.04	278	750	664
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Default	54.42	12.42 ↑	8.75 ↑	1.56	0.75	232	612	532
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Deterministic	54.37	12.37 ↑	8.70 ↑	1.56	0.50	224	489	435
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Exploratory	54.21	12.21 ↑	8.54 ↑	1.77	0.89	255	696	602
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Det. & Exp.	53.29	11.29 ↑	7.62 ↑	1.65	0.62	249	555	488
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Exp. & Det.	54.58	12.58 ↑	8.91 ↑	1.51	0.77	249	603	533
Qwen-2.5-3B	Phi-mini-3.8B	Both: Default	70.21	8.46 ↑	6.79 ↑	0.79	0.41	132	304	275
Qwen-2.5-3B	Phi-mini-3.8B	Both: Deterministic	69.83	8.08 ↑	6.41 ↑	0.78	0.29	128	224	200
Qwen-2.5-3B	Phi-mini-3.8B	Both: Exploratory	69.71	7.96 ↑	6.29 ↑	0.83	0.47	136	339	303
Qwen-2.5-3B	Phi-mini-3.8B	Both: Det. & Exp.	69.88	8.13 ↑	6.46 ↑	0.79	0.31	133	241	216
Qwen-2.5-3B	Phi-mini-3.8B	Both: Exp. & Det.	70.58	8.83 ↑	7.16 ↑	0.81	0.38	134	307	276
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Default	63.79	21.79 ↑	2.04 ↑	1.05	0.67	154	573	537
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Deterministic	63.92	21.92 ↑	2.17 ↑	0.85	0.60	180	500	471
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Exploratory	63.79	21.79 ↑	2.04 ↑	1.12	0.76	165	680	639
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Det. & Exp.	62.58	20.58 ↑	0.83 ↑	1.09	0.61	174	525	483
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Exp. & Det.	64.25	22.25 ↑	2.50 ↑	1.08	0.68	189	640	608
Llama-3.1-3B	Llama-3.1-8B	Both: Default	56.75	11.08 ↑	1.13 ↑	1.29	0.88	264	422	381
Llama-3.1-3B	Llama-3.1-8B	Both: Deterministic	57.08	11.41 ↑	1.46 ↑	1.13	0.74	278	348	316
Llama-3.1-3B	Llama-3.1-8B	Both: Exploratory	57.17	11.50 ↑	1.55 ↑	1.43	0.89	241	490	424
Llama-3.1-3B	Llama-3.1-8B	Both: Det. & Exp.	57.21	11.54 ↑	1.59 ↑	1.27	0.72	259	420	362
Llama-3.1-3B	Llama-3.1-8B	Both: Exp. & Det.	56.67	11.00 ↑	1.05 ↑	1.27	0.80	298	411	364
Qwen-2.5-7B	Qwen-2.5-14B	Both: Default	75.88	7.26 ↑	4.09 ↑	0.38	0.28	88	165	159
Qwen-2.5-7B	Qwen-2.5-14B	Both: Deterministic	75.54	6.92 ↑	3.75 ↑	0.32	0.24	83	119	112
Qwen-2.5-7B	Qwen-2.5-14B	Both: Exploratory	75.08	6.46 ↑	3.29 ↑	0.39	0.30	111	168	153
Qwen-2.5-7B	Qwen-2.5-14B	Both: Det. & Exp.	76.12	7.50 ↑	4.33 ↑	0.36	0.25	92	155	148
Qwen-2.5-7B	Qwen-2.5-14B	Both: Exp. & Det.	76.33	7.71 ↑	4.54 ↑	0.35	0.31	78	143	133

Table 18: Comparative Analysis of Mixed-Model Performance in Multi-Agent Debate Settings on the **GSM-Plus** Dataset. This table showcases the impact of different **Agent Settings** on the **MAD Accuracy** when pairing different language models together. The Δ **Lower** and Δ **Upper** columns quantify the improvement (or decline) over each individual model's base performance. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 2400 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the dynamics when models of different capabilities debate together.

Agent 1	Agent 2	Agent 3	Agent Settings	Accuracy	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 2400)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Default	25.00	0.21 ↑	3.21	3.75	583	473	299
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Deterministic	29.21	4.42 ↑	0.02	0.00	0	0	0
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Exploratory	20.75	4.04 ↓	3.88	3.78	645	578	344
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	1 Det. & 2 Exp.	22.67	2.12 ↓	3.66	3.40	667	467	296
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	2 Det. & 1 Exp.	25.42	0.63 ↑	2.45	1.96	454	394	279
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Default	53.04	11.04 ↑	1.87	2.28	446	995	676
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Deterministic	47.29	5.29 ↑	0.03	0.00	0	0	0
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Exploratory	53.33	11.33 ↑	2.24	2.74	357	1159	774
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	1 Det. & 2 Exp.	53.67	11.67 ↑	2.03	2.35	394	1116	756
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	2 Det. & 1 Exp.	53.17	11.17 ↑	1.31	1.41	265	793	514
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Default	67.38	5.63 ↑	0.97	1.01	273	423	326
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Deterministic	67.38	5.63 ↑	0.05	0.00	0	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Exploratory	68.00	6.25 ↑	1.09	1.12	223	537	404
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	1 Det. & 2 Exp.	68.54	6.79 ↑	1.08	0.94	235	428	343
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	2 Det. & 1 Exp.	67.12	5.37 ↑	0.78	0.61	202	274	208
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Default	75.79	7.17 ↑	0.51	0.52	84	272	209
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Deterministic	73.62	5.00 ↑	0.04	0.00	0	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Exploratory	74.96	6.34 ↑	0.55	0.54	117	270	220
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	1 Det. & 2 Exp.	75.25	6.63 ↑	0.50	0.50	120	267	214
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	2 Det. & 1 Exp.	74.42	5.80 ↑	0.39	0.39	97	181	135
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Default	77.92	6.13 ↑	0.35	0.35	55	166	140
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Deterministic	76.54	4.75 ↑	0.05	0.00	0	3	1
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Exploratory	77.29	5.50 ↑	0.38	0.40	69	188	159
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	1 Det. & 2 Exp.	77.21	5.42 ↑	0.38	0.37	72	172	143
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	2 Det. & 1 Exp.	77.21	5.42 ↑	0.28	0.25	48	105	81
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Default	73.46	1.00 ↑	0.29	0.23	48	112	96
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Deterministic	72.79	0.33 ↑	0.08	0.00	0	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Exploratory	73.46	1.00 ↑	0.33	0.31	46	123	109
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	1 Det. & 2 Exp.	73.88	1.42 ↑	0.29	0.23	42	131	106
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	2 Det. & 1 Exp.	73.12	0.66 ↑	0.24	0.17	40	75	60
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Default	70.21	6.79 ↑	0.90	1.12	226	389	284
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Deterministic	69.17	5.75 ↑	0.12	0.04	0	3	1
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Exploratory	70.25	6.83 ↑	0.95	1.11	219	423	327
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	1 Det. & 2 Exp.	69.83	6.41 ↑	0.93	1.02	232	390	293
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	2 Det. & 1 Exp.	69.54	6.12 ↑	0.73	0.81	191	292	202
Mistral-7B	Mistral-7B	Mistral-7B	Default	24.04	8.99 ↑	2.75	2.12	312	979	525
Mistral-7B	Mistral-7B	Mistral-7B	Deterministic	14.37	0.67 ↓	0.15	0.02	0	8	3
Mistral-7B	Mistral-7B	Mistral-7B	Exploratory	27.04	12.00 ↑	3.03	2.49	325	1234	628
Mistral-7B	Mistral-7B	Mistral-7B	1 Det. & 2 Exp.	23.92	8.88 ↑	2.90	2.25	349	1046	544
Mistral-7B	Mistral-7B	Mistral-7B	2 Det. & 1 Exp.	23.00	7.96 ↑	2.16	1.55	232	855	458
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Default	51.54	5.87 ↑	1.89	1.93	454	733	476
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Deterministic	50.67	5.00 ↑	0.01	0.00	0	0	0
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Exploratory	50.71	5.04 ↑	2.26	2.12	520	857	544
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	1 Det. & 2 Exp.	50.17	4.50 ↑	2.12	1.96	515	744	493
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	2 Det. & 1 Exp.	51.33	5.66 ↑	1.50	1.23	309	493	322
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Default	62.67	7.05 ↑	1.43	1.60	345	572	407
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Deterministic	61.04	5.42 ↑	0.00	0.00	0	0	0
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Exploratory	61.08	5.46 ↑	1.69	1.85	385	624	446
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	1 Det. & 2 Exp.	62.12	6.50 ↑	1.51	1.64	374	588	413
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	2 Det. & 1 Exp.	61.12	5.50 ↑	1.20	1.20	335	414	269

Table 19: Comparative Analysis of Language Model Performance in Multi-Agent Debate Settings on the **GSM-Plus** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters like Default, Deterministic, Exploratory, and combinations) on the **Accuracy** of various language models in three-agent configurations. The Δ column quantifies the **improvement (or decline) over the single base model performance**. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 2400 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent 3	Agent Settings	Accuracy	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 2400)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-3B	Default	60.00	1.75 ↓	2.35	2.05	338	1356	951
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Llama-3.1-3B	Default	47.46	1.79 ↑	3.11	2.23	596	1086	718
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Phi-mini-3.8B	Default	56.62	6.80 ↓	2.83	1.93	503	1168	857
Qwen-2.5-0.5B	Qwen-2.5-3B	Llama-3.1-3B	Default	59.62	2.13 ↓	2.83	1.90	364	1202	895
Qwen-2.5-0.5B	Qwen-2.5-3B	Phi-mini-3.8B	Default	65.25	1.83 ↑	2.42	1.48	353	1190	946
Qwen-2.5-0.5B	Llama-3.1-3B	Phi-mini-3.8B	Default	56.92	6.50 ↓	3.13	1.64	536	980	724
Qwen-2.5-1.5B	Qwen-2.5-3B	Llama-3.1-3B	Default	64.00	2.25 ↑	1.91	1.59	321	1048	773
Qwen-2.5-1.5B	Qwen-2.5-3B	Phi-mini-3.8B	Default	67.25	3.83 ↑	1.61	1.25	299	857	692
Qwen-2.5-1.5B	Llama-3.1-3B	Phi-mini-3.8B	Default	63.50	0.08 ↑	2.02	1.57	405	1079	766
Qwen-2.5-3B	Phi-mini-3.8B	Llama-3.1-3B	Default	69.08	5.66 ↑	1.58	1.20	255	825	653
Qwen-2.5-3B	Qwen-2.5-3B	Phi-mini-3.8B	Default	68.79	7.04 ↑	1.13	0.90	291	461	340
Qwen-2.5-3B	Phi-mini-3.8B	Phi-mini-3.8B	Default	69.21	5.79 ↑	1.10	0.92	279	424	317
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Default	49.88	7.88 ↑	2.44	2.50	456	1197	794
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-1.5B	Default	37.21	4.79 ↓	3.07	3.24	589	969	607

Table 20: Comparative Analysis of Mixed Multi-Agent Debate Settings on the **GSM-Plus** Dataset. This table examines performance when combining different language models in three-agent debate configurations. The first section shows combinations of three different models, while the second section explores configurations with duplicate models. The Δ column indicates performance changes relative to the best single model in each combination, with improvements in green and declines in red. Metrics include **Debate Rounds**, normalized **Sycophancy** (per 2400 data points), and transitions between states (C→I, I→C).

Agent 1	Agent 2	Agent Settings	Accuracy	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 2376)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Default	52.90	1.73 ↓	1.15	0.99	460.00	550	482
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Deterministic	53.24	1.39 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Exploratory	49.07	5.56 ↓	1.46	1.09	558.00	628	530
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Det. & Exp.	52.99	1.64 ↓	1.15	0.97	426.00	572	516
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Default	86.15	0.47 ↓	0.38	0.38	130.00	415	403
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Deterministic	84.60	2.02 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Exploratory	83.42	3.20 ↓	0.55	0.55	160.00	574	547
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Det. & Exp.	86.62	0.00	0.41	0.42	135.00	449	434
Qwen-2.5-3B	Qwen-2.5-3B	Default	94.02	0.96 ↑	0.14	0.13	56.00	117	114
Qwen-2.5-3B	Qwen-2.5-3B	Deterministic	93.35	0.29 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Exploratory	94.15	1.09 ↑	0.16	0.15	49.00	158	157
Qwen-2.5-3B	Qwen-2.5-3B	Det. & Exp.	94.07	1.01 ↑	0.15	0.13	70.00	126	124
Qwen-2.5-7B	Qwen-2.5-7B	Default	96.17	1.48 ↑	0.05	0.05	31.00	39	37
Qwen-2.5-7B	Qwen-2.5-7B	Deterministic	96.55	1.86 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Exploratory	96.93	2.24 ↑	0.05	0.05	21.00	57	53
Qwen-2.5-7B	Qwen-2.5-7B	Det. & Exp.	96.46	1.77 ↑	0.05	0.04	30.00	35	34
Qwen-2.5-14B	Qwen-2.5-14B	Default	98.19	2.53 ↑	0.03	0.02	15.00	21	21
Qwen-2.5-14B	Qwen-2.5-14B	Deterministic	97.77	2.11 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-14B	Qwen-2.5-14B	Exploratory	98.15	2.49 ↑	0.02	0.02	8.00	20	20
Qwen-2.5-14B	Qwen-2.5-14B	Det. & Exp.	97.94	2.28 ↑	0.03	0.02	16.00	24	24
Qwen-2.5-32B	Qwen-2.5-32B	Default	98.53	0.21 ↑	0.02	0.03	10.00	14	13
Qwen-2.5-32B	Qwen-2.5-32B	Deterministic	98.36	0.04 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Exploratory	98.53	0.21 ↑	0.02	0.03	8.00	14	14
Qwen-2.5-32B	Qwen-2.5-32B	Det. & Exp.	98.36	0.04 ↑	0.02	0.02	9.00	10	8
Phi-mini-3.8B	Phi-mini-3.8B	Default	95.88	3.92 ↑	0.11	0.16	40.00	71	60
Phi-mini-3.8B	Phi-mini-3.8B	Deterministic	95.37	3.41 ↑	0.00	0.00	0.00	0	0
Phi-mini-3.8B	Phi-mini-3.8B	Exploratory	94.74	2.78 ↑	0.16	0.21	59.00	126	116
Phi-mini-3.8B	Phi-mini-3.8B	Det. & Exp.	94.95	2.99 ↑	0.14	0.19	56.00	89	78
Mistral-7B	Mistral-7B	Default	81.06	0.04 ↑	0.35	0.28	158.00	227	219
Mistral-7B	Mistral-7B	Deterministic	80.43	0.59 ↓	0.00	0.00	0.00	0	0
Mistral-7B	Mistral-7B	Exploratory	80.18	0.84 ↓	0.43	0.32	203.00	261	251
Mistral-7B	Mistral-7B	Det. & Exp.	82.41	1.39 ↑	0.37	0.27	129.00	240	235
Llama-3.1-3B	Llama-3.1-3B	Default	87.71	3.07 ↑	0.26	0.21	128.00	163	153
Llama-3.1-3B	Llama-3.1-3B	Deterministic	86.66	2.02 ↑	0.00	0.00	0.00	0	0
Llama-3.1-3B	Llama-3.1-3B	Exploratory	88.09	3.45 ↑	0.28	0.26	118.00	216	208
Llama-3.1-3B	Llama-3.1-3B	Det. & Exp.	86.91	2.27 ↑	0.28	0.22	127.00	181	172
Llama-3.1-8B	Llama-3.1-8B	Default	94.44	5.34 ↑	0.11	0.11	54.00	79	75
Llama-3.1-8B	Llama-3.1-8B	Deterministic	93.64	4.54 ↑	0.00	0.00	0.00	0	0
Llama-3.1-8B	Llama-3.1-8B	Exploratory	93.60	4.50 ↑	0.15	0.17	60.00	118	109
Llama-3.1-8B	Llama-3.1-8B	Det. & Exp.	94.53	5.43 ↑	0.12	0.13	54.00	95	93

Table 21: Comparative Analysis of Language Model Performance in Multi-Agent Debate Settings on the **ARC-Easy** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters like Default, Deterministic, Exploratory, and a combination) on the **Accuracy** of various language models. The Δ column quantifies the **improvement (or decline) over the single base model performance**. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 2376 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent Settings	Accuracy	∆ Lower	∆ Upper	Debate Rounds (Avg)	Sycophancy (Avg / 2376)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Default	76.98	22.35 ↑	9.64 ↓	0.95	0.75	262.00	804	760
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Deterministic	79.38	24.75 ↑	7.24 ↓	0.81	0.62	200.00	734	711
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Exploratory	73.19	18.56 ↑	13.43 ↓	1.16	0.85	300.00	899	828
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Det. & Exp.	75.21	20.58 ↑	11.41 ↓	0.95	0.78	260.00	846	790
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Exp. & Det.	77.65	23.02 ↑	8.97 ↓	1.07	0.75	275.00	829	794
Qwen-2.5-1.5B	Llama-3.1-3B	Default	88.55	1.93 ↑	3.91 ↑	0.40	0.39	146.00	376	357
Qwen-2.5-1.5B	Llama-3.1-3B	Deterministic	88.13	1.51 ↑	3.49 ↑	0.29	0.24	150.00	242	239
Qwen-2.5-1.5B	Llama-3.1-3B	Exploratory	88.05	1.43 ↑	3.41 ↑	0.49	0.48	161.00	483	457
Qwen-2.5-1.5B	Llama-3.1-3B	Det. & Exp.	86.99	0.37 ↑	1.35 ↑	0.37	0.39	172.00	290	277
Qwen-2.5-1.5B	Llama-3.1-3B	Exp. & Det.	87.71	1.09 ↑	2.07 ↑	0.45	0.40	165.00	447	433
Qwen-2.5-3B	Phi-mini-3.8B	Default	95.24	2.18 ↑	3.28 ↑	0.15	0.14	61.00	135	132
Qwen-2.5-3B	Phi-mini-3.8B	Deterministic	94.91	1.85 ↑	2.95 ↑	0.14	0.12	72.00	106	102
Qwen-2.5-3B	Phi-mini-3.8B	Exploratory	95.24	2.18 ↑	3.28 ↑	0.17	0.16	57.00	184	178
Qwen-2.5-3B	Phi-mini-3.8B	Det. & Exp.	94.91	1.85 ↑	2.95 ↑	0.17	0.15	68.00	148	148
Qwen-2.5-3B	Phi-mini-3.8B	Exp. & Det.	95.75	2.69 ↑	3.79 ↑	0.15	0.14	58.00	146	139
Qwen-2.5-1.5B	Qwen-2.5-3B	Default	91.88	5.26 ↑	1.18 ↓	0.33	0.29	112.00	363	359
Qwen-2.5-1.5B	Qwen-2.5-3B	Deterministic	92.59	5.97 ↑	0.47 ↓	0.24	0.23	94.00	263	254
Qwen-2.5-1.5B	Qwen-2.5-3B	Exploratory	91.79	5.17 ↑	1.27 ↓	0.42	0.38	95.00	498	487
Qwen-2.5-1.5B	Qwen-2.5-3B	Det. & Exp.	92.76	6.14 ↑	0.20 ↓	0.27	0.27	81.00	294	286
Qwen-2.5-1.5B	Qwen-2.5-3B	Exp. & Det.	92.51	5.89 ↑	0.45 ↓	0.39	0.32	96.00	469	466
Llama-3.1-3B	Llama-3.1-8B	Default	91.79	7.15 ↑	2.69 ↑	0.24	0.22	110.00	184	179
Llama-3.1-3B	Llama-3.1-8B	Deterministic	91.12	6.48 ↑	2.02 ↑	0.22	0.16	113.00	138	133
Llama-3.1-3B	Llama-3.1-8B	Exploratory	90.61	5.97 ↑	1.51 ↑	0.28	0.27	115.00	202	192
Llama-3.1-3B	Llama-3.1-8B	Det. & Exp.	90.99	6.35 ↑	1.89 ↑	0.24	0.18	108.00	152	149
Llama-3.1-3B	Llama-3.1-8B	Exp. & Det.	91.96	7.32 ↑	2.86 ↑	0.28	0.26	99.00	229	222
Qwen-2.5-7B	Qwen-2.5-14B	Default	97.94	3.25 ↑	2.28 ↑	0.05	0.05	21.00	55	55
Qwen-2.5-7B	Qwen-2.5-14B	Deterministic	97.64	2.95 ↑	1.98 ↑	0.07	0.04	20.00	48	47
Qwen-2.5-7B	Qwen-2.5-14B	Exploratory	97.39	2.70 ↑	1.73 ↑	0.08	0.07	32.00	67	66
Qwen-2.5-7B	Qwen-2.5-14B	Det. & Exp.	97.43	2.74 ↑	1.77 ↑	0.06	0.05	33.00	49	48
Qwen-2.5-7B	Qwen-2.5-14B	Exp. & Det.	97.47	2.78 ↑	1.81 ↑	0.07	0.04	27.00	49	48

Table 22: Comparative Analysis of Different Language Model Pairs in Multi-Agent Debate Settings on the **ARC-Easy** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters) on the **Accuracy** of various model pairs. The ∆ **Lower** and ∆ **Upper** columns quantify the improvement (or decline) over each individual model's single-agent performance. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 2376 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics between different model pairings.

Agent 1	Agent 2	Agent Settings	MAD Accuracy (RCR Prompting)	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 2376)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Default	51.30	3.33 ↓	2.18	2.67	1046.00	990	642
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Deterministic	53.24	1.39 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Exploratory	46.80	7.83 ↓	2.78	3.22	1228.00	1099	655
Qwen-2.5-0.5B	Qwen-2.5-0.5B	1 Det. & 2 Exp.	48.99	5.64 ↓	2.47	2.82	1136.00	1053	658
Qwen-2.5-0.5B	Qwen-2.5-0.5B	2 Det. & 1 Exp.	50.80	3.83 ↓	1.34	1.60	794.00	793	495
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Default	87.37	0.75 ↑	0.63	0.84	232.00	717	573
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Deterministic	84.60	2.02 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Exploratory	85.61	1.01 ↓	0.90	1.17	279.00	1011	795
Qwen-2.5-1.5B	Qwen-2.5-1.5B	1 Det. & 2 Exp.	86.32	0.30 ↓	0.76	0.98	275.00	834	672
Qwen-2.5-1.5B	Qwen-2.5-1.5B	2 Det. & 1 Exp.	86.53	0.09 ↓	0.43	0.62	198.00	587	451
Qwen-2.5-3B	Qwen-2.5-3B	Both: Default	94.87	1.81 ↑	0.19	0.19	80.00	196	165
Qwen-2.5-3B	Qwen-2.5-3B	Both: Deterministic	93.35	0.29 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Both: Exploratory	94.28	1.22 ↑	0.25	0.28	102.00	252	206
Qwen-2.5-3B	Qwen-2.5-3B	1 Det. & 2 Exp.	94.70	1.64 ↑	0.25	0.23	90.00	238	195
Qwen-2.5-3B	Qwen-2.5-3B	2 Det. & 1 Exp.	93.94	0.88 ↑	0.20	0.18	94.00	162	117
Qwen-2.5-7B	Qwen-2.5-7B	Both: Default	96.21	1.52 ↑	0.08	0.08	53.00	69	58
Qwen-2.5-7B	Qwen-2.5-7B	Both: Deterministic	96.17	1.48 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Both: Exploratory	96.55	1.86 ↑	0.10	0.11	57.00	86	71
Qwen-2.5-7B	Qwen-2.5-7B	1 Det. & 2 Exp.	96.55	1.86 ↑	0.10	0.11	56.00	78	65
Qwen-2.5-7B	Qwen-2.5-7B	2 Det. & 1 Exp.	96.34	1.65 ↑	0.07	0.07	39.00	56	40
Qwen-2.5-14B	Qwen-2.5-14B	Both: Default	98.15	2.49 ↑	0.04	0.04	23.00	29	26
Qwen-2.5-14B	Qwen-2.5-14B	Both: Deterministic	97.77	2.11 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-14B	Qwen-2.5-14B	Both: Exploratory	98.19	2.53 ↑	0.04	0.05	18.00	40	36
Qwen-2.5-14B	Qwen-2.5-14B	1 Det. & 2 Exp.	98.02	2.36 ↑	0.03	0.04	28.00	40	31
Qwen-2.5-14B	Qwen-2.5-14B	2 Det. & 1 Exp.	97.81	2.15 ↑	0.03	0.03	23.00	28	25
Qwen-2.5-32B	Qwen-2.5-32B	Both: Default	98.57	0.25 ↑	0.02	0.03	16.00	15	13
Qwen-2.5-32B	Qwen-2.5-32B	Both: Deterministic	98.36	0.04 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Both: Exploratory	98.48	0.16 ↑	0.02	0.02	15.00	14	14
Qwen-2.5-32B	Qwen-2.5-32B	1 Det. & 2 Exp.	98.48	0.16 ↑	0.02	0.03	16.00	15	12
Qwen-2.5-32B	Qwen-2.5-32B	2 Det. & 1 Exp.	98.32	0.00	0.01	0.02	12.00	9	6
Phi-mini-3.8B	Phi-mini-3.8B	Both: Default	95.79	3.83 ↑	0.16	0.28	79.00	138	105
Phi-mini-3.8B	Phi-mini-3.8B	Both: Deterministic	95.37	3.41 ↑	0.00	0.00	0.00	0	0
Phi-mini-3.8B	Phi-mini-3.8B	Both: Exploratory	94.91	2.95 ↑	0.28	0.43	110.00	234	185
Phi-mini-3.8B	Phi-mini-3.8B	1 Det. & 2 Exp.	96.34	4.38 ↑	0.18	0.27	70.00	189	149
Phi-mini-3.8B	Phi-mini-3.8B	2 Det. & 1 Exp.	95.92	3.96 ↑	0.13	0.24	53.00	115	83
Llama-3.1-3B	Llama-3.1-3B	Both: Default	87.33	2.69 ↑	0.46	0.44	252.00	292	227
Llama-3.1-3B	Llama-3.1-3B	Both: Deterministic	87.63	2.99 ↑	0.00	0.00	0.00	0	0
Llama-3.1-3B	Llama-3.1-3B	Both: Exploratory	87.71	3.07 ↑	0.58	0.61	255.00	415	323
Llama-3.1-3B	Llama-3.1-3B	1 Det. & 2 Exp.	87.58	2.94 ↑	0.53	0.48	241.00	328	259
Llama-3.1-3B	Llama-3.1-3B	2 Det. & 1 Exp.	88.47	3.83 ↑	0.32	0.27	148.00	236	169
Llama-3.1-8B	Llama-3.1-8B	Both: Default	93.86	4.76 ↑	0.20	0.26	114.00	139	102
Llama-3.1-8B	Llama-3.1-8B	Both: Deterministic	93.64	4.54 ↑	0.00	0.00	0.00	0	0
Llama-3.1-8B	Llama-3.1-8B	Both: Exploratory	94.19	5.09 ↑	0.25	0.36	130.00	190	141
Llama-3.1-8B	Llama-3.1-8B	1 Det. & 2 Exp.	94.11	5.01 ↑	0.23	0.33	119.00	185	143
Llama-3.1-8B	Llama-3.1-8B	2 Det. & 1 Exp.	94.49	5.39 ↑	0.14	0.20	69.00	139	89
Mistral-7B	Mistral-7B	Both: Default	82.20	1.18 ↑	0.69	0.71	318.00	469	342
Mistral-7B	Mistral-7B	Both: Deterministic	80.43	0.59 ↓	0.00	0.00	0.00	0	0
Mistral-7B	Mistral-7B	Both: Exploratory	82.66	1.64 ↑	0.83	0.88	325.00	566	429
Mistral-7B	Mistral-7B	1 Det. & 2 Exp.	82.37	1.35 ↑	0.78	0.81	324.00	506	376
Mistral-7B	Mistral-7B	2 Det. & 1 Exp.	81.69	0.67 ↑	0.47	0.51	230.00	346	230

Table 23: Comparative Analysis of Language Model Performance in Multi-Agent Debate Settings on the **ARC-Easy** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters like Default, Deterministic, Exploratory, and combinations) on the **MAD Accuracy (RCR Prompting)** of various language models. The Δ column quantifies the **improvement (or decline) over the single base model performance**. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 2376 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent 3	MAD Accuracy (RCR Prompting)	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 2376)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-3B	92.72	0.34 ↓	1.00	0.95	145	1377	1153
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Llama-3.1-3B	84.64	0.00	1.18	1.27	387	1223	1006
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Phi-mini-3.8B	92.93	0.97 ↑	1.03	1.04	184	1379	1156
Qwen-2.5-0.5B	Qwen-2.5-3B	Llama-3.1-3B	91.20	1.86 ↓	1.13	0.99	213	1221	1070
Qwen-2.5-0.5B	Qwen-2.5-3B	Phi-mini-3.8B	89.48	3.58 ↓	1.09	1.12	299	1157	1024
Qwen-2.5-0.5B	Llama-3.1-3B	Phi-mini-3.8B	91.79	0.17 ↓	0.58	0.72	238	559	479
Qwen-2.5-1.5B	Qwen-2.5-3B	Llama-3.1-3B	91.84	1.22 ↓	0.56	0.60	189	560	479
Qwen-2.5-1.5B	Qwen-2.5-3B	Phi-mini-3.8B	95.54	2.48 ↑	0.39	0.45	103	509	449
Qwen-2.5-1.5B	Llama-3.1-3B	Phi-mini-3.8B	91.79	0.17 ↓	0.58	0.72	238	559	479
Qwen-2.5-3B	Phi-mini-3.8B	Llama-3.1-3B	94.07	1.01 ↑	0.41	0.43	162	332	283
Qwen-2.5-3B	Qwen-2.5-3B	Phi-mini-3.8B	95.88	2.82 ↑	0.26	0.26	86	253	214
Qwen-2.5-3B	Phi-mini-3.8B	Phi-mini-3.8B	96.34	3.28 ↑	0.26	0.31	71	227	180
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	84.64	2.00 ↓	1.22	1.22	300	1229	1012
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-1.5B	72.43	14.19 ↓	1.86	2.11	616	1400	982

Table 24: Comparative Analysis of Multi-Model Combinations in Agent Debate Settings on the **ARC-Easy** Dataset. This table showcases the performance of heterogeneous agent teams consisting of different language models. The **MAD Accuracy (RCR Prompting)** reflects the team performance, while the Δ column quantifies the **improvement (or decline) relative to the best single model** in each combination. Additional metrics include average **Debate Rounds**, normalized **Sycophancy** (per 2376 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), revealing how diverse model combinations affect debate dynamics and overall helpfulness.

Agent 1	Agent 2	Agent Settings	MAD Accuracy (ARC-Challenge)	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 1172)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Default	39.51	1.54 ↑	1.32	1.10	253.00	265	228
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Deterministic	40.78	2.81 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Exploratory	37.54	0.43 ↓	1.51	1.14	266.00	309	245
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Det. & Exp.	39.85	1.88 ↑	1.34	1.12	247.00	259	227
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Default	70.90	1.69 ↑	0.57	0.58	115.00	249	242
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Deterministic	67.58	1.63 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Exploratory	68.52	0.69 ↓	0.75	0.70	133.00	296	275
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Det. & Exp.	69.88	0.67 ↑	0.60	0.61	101.00	262	252
Qwen-2.5-3B	Qwen-2.5-3B	Both: Default	85.41	1.88 ↑	0.29	0.29	53.00	114	111
Qwen-2.5-3B	Qwen-2.5-3B	Both: Deterministic	84.13	0.60 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Both: Exploratory	84.64	1.11 ↑	0.30	0.27	56.00	116	109
Qwen-2.5-3B	Qwen-2.5-3B	Both: Det. & Exp.	83.70	0.17 ↑	0.28	0.23	70.00	79	73
Qwen-2.5-7B	Qwen-2.5-7B	Both: Default	91.55	4.33 ↑	0.11	0.11	29.00	46	45
Qwen-2.5-7B	Qwen-2.5-7B	Both: Deterministic	91.21	3.99 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Both: Exploratory	91.64	4.42 ↑	0.12	0.11	23.00	53	51
Qwen-2.5-7B	Qwen-2.5-7B	Both: Det. & Exp.	92.06	4.84 ↑	0.13	0.12	30.00	48	43
Qwen-2.5-14B	Qwen-2.5-14B	Both: Default	94.54	4.27 ↑	0.06	0.05	13.00	24	24
Qwen-2.5-14B	Qwen-2.5-14B	Both: Deterministic	94.37	4.10 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-14B	Qwen-2.5-14B	Both: Exploratory	93.77	3.50 ↑	0.06	0.07	23.00	24	24
Qwen-2.5-14B	Qwen-2.5-14B	Both: Det. & Exp.	94.71	4.44 ↑	0.06	0.06	11.00	22	21
Qwen-2.5-32B	Qwen-2.5-32B	Both: Default	98.53	3.25 ↑	0.02	0.06	10.00	14	13
Qwen-2.5-32B	Qwen-2.5-32B	Both: Deterministic	98.36	3.08 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Both: Exploratory	98.53	3.25 ↑	0.02	0.06	8.00	14	14
Qwen-2.5-32B	Qwen-2.5-32B	Both: Det. & Exp.	98.36	3.08 ↑	0.02	0.04	9.00	10	8
Phi-mini-3.8B	Phi-mini-3.8B	Both: Default	90.10	5.37 ↑	0.24	0.34	42.00	75	66
Phi-mini-3.8B	Phi-mini-3.8B	Both: Deterministic	88.91	4.18 ↑	0.00	0.00	0.00	0	0
Phi-mini-3.8B	Phi-mini-3.8B	Both: Exploratory	87.03	2.30 ↑	0.31	0.40	58.00	107	100
Phi-mini-3.8B	Phi-mini-3.8B	Both: Det. & Exp.	88.05	3.32 ↑	0.23	0.31	46.00	69	62
Llama-3.1-3B	Llama-3.1-3B	Both: Default	75.77	2.65 ↑	0.46	0.37	93.00	130	126
Llama-3.1-3B	Llama-3.1-3B	Both: Deterministic	74.66	1.54 ↑	0.00	0.00	0.00	0	0
Llama-3.1-3B	Llama-3.1-3B	Both: Exploratory	76.19	3.07 ↑	0.50	0.43	89.00	166	149
Llama-3.1-3B	Llama-3.1-3B	Both: Det. & Exp.	75.60	2.48 ↑	0.45	0.34	108.00	129	124
Llama-3.1-8B	Llama-3.1-8B	Both: Default	87.20	9.55 ↑	0.26	0.30	45.00	91	88
Llama-3.1-8B	Llama-3.1-8B	Both: Deterministic	85.75	8.10 ↑	0.00	0.00	0.00	0	0
Llama-3.1-8B	Llama-3.1-8B	Both: Exploratory	85.07	7.42 ↑	0.28	0.32	58.00	96	94
Llama-3.1-8B	Llama-3.1-8B	Both: Det. & Exp.	86.86	9.21 ↑	0.23	0.27	56.00	84	80
Mistral-7B	Mistral-7B	Both: Default	70.48	1.71 ↑	0.51	0.37	99.00	145	137
Mistral-7B	Mistral-7B	Both: Deterministic	68.26	0.51 ↓	0.00	0.00	0.00	0	0
Mistral-7B	Mistral-7B	Both: Exploratory	72.78	4.01 ↑	0.58	0.44	106.00	185	177
Mistral-7B	Mistral-7B	Both: Det. & Exp.	70.82	2.05 ↑	0.50	0.34	84.00	151	142

Table 25: Comparative Analysis of Language Model Performance in Multi-Agent Debate Settings on the **ARC-Challenge** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters like Default, Deterministic, Exploratory, and a combination) on the **MAD Accuracy** of various language models. The Δ column quantifies the **improvement (or decline) over the single base model performance** shown in parentheses next to each model name. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 1172 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent Settings	MAD Accuracy (ARC-Challenge)	Δ_1	Δ_2	Debate Rounds (Avg)	Sycophancy (Avg / 1172)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Default	58.28	20.31 ↑	10.93 ↓	1.27	0.97	193	401	369
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Deterministic	63.57	25.60 ↑	5.64 ↓	1.09	0.81	169	375	357
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Exploratory	55.80	17.83 ↑	13.41 ↓	1.46	1.07	211	418	368
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Det. & Exp.	60.32	22.35 ↑	8.89 ↓	1.10	0.94	181	397	360
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Exp. & Det.	61.43	23.46 ↑	7.78 ↓	1.39	0.95	197	409	387
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Default	72.35	3.14 ↑	0.77 ↓	0.67	0.66	143	216	207
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Deterministic	74.91	5.70 ↑	1.79 ↑	0.51	0.51	135	191	185
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Exploratory	73.12	3.91 ↑	0.00	0.78	0.78	153	281	265
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Det. & Exp.	76.02	6.81 ↑	2.90 ↑	0.60	0.66	127	219	205
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Exp. & Det.	74.15	4.94 ↑	1.03 ↑	0.71	0.61	135	291	274
Qwen-2.5-3B	Phi-mini-3.8B	Both: Default	87.97	4.44 ↑	3.24 ↑	0.32	0.31	59	133	130
Qwen-2.5-3B	Phi-mini-3.8B	Both: Deterministic	88.57	5.04 ↑	3.84 ↑	0.31	0.25	58	110	107
Qwen-2.5-3B	Phi-mini-3.8B	Both: Exploratory	87.03	3.50 ↑	2.30 ↑	0.38	0.37	72	173	160
Qwen-2.5-3B	Phi-mini-3.8B	Both: Det. & Exp.	87.80	4.27 ↑	3.07 ↑	0.33	0.30	59	141	139
Qwen-2.5-3B	Phi-mini-3.8B	Both: Exp. & Det.	89.85	6.32 ↑	5.12 ↑	0.34	0.30	50	143	137
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Default	82.25	13.04 ↑	1.28 ↓	0.51	0.45	80	247	243
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Deterministic	82.59	13.38 ↑	0.94 ↓	0.42	0.40	80	205	200
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Exploratory	81.91	12.70 ↑	1.62 ↓	0.66	0.56	94	317	310
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Det. & Exp.	83.45	14.24 ↑	0.08 ↓	0.47	0.46	66	227	219
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Exp. & Det.	83.62	14.41 ↑	0.09 ↑	0.62	0.51	67	328	320
Llama-3.1-3B	Llama-3.1-8B	Both: Default	81.66	8.54 ↑	4.01 ↑	0.47	0.41	114	141	133
Llama-3.1-3B	Llama-3.1-8B	Both: Deterministic	80.46	7.34 ↑	2.81 ↑	0.51	0.36	120	135	124
Llama-3.1-3B	Llama-3.1-8B	Both: Exploratory	75.68	2.56 ↑	1.97 ↓	0.48	0.43	107	160	151
Llama-3.1-3B	Llama-3.1-8B	Both: Det. & Exp.	80.12	7.00 ↑	2.47 ↑	0.46	0.37	117	138	132
Llama-3.1-3B	Llama-3.1-8B	Both: Exp. & Det.	80.97	7.85 ↑	3.32 ↑	0.49	0.43	109	159	154
Qwen-2.5-7B	Qwen-2.5-14B	Both: Default	93.43	6.21 ↑	3.16 ↑	0.14	0.11	35	54	53
Qwen-2.5-7B	Qwen-2.5-14B	Both: Deterministic	93.60	6.38 ↑	3.33 ↑	0.13	0.10	24	59	58
Qwen-2.5-7B	Qwen-2.5-14B	Both: Exploratory	94.45	7.23 ↑	4.18 ↑	0.15	0.14	27	67	65
Qwen-2.5-7B	Qwen-2.5-14B	Both: Det. & Exp.	93.00	5.78 ↑	2.73 ↑	0.16	0.13	37	50	49
Qwen-2.5-7B	Qwen-2.5-14B	Both: Exp. & Det.	93.77	6.55 ↑	3.50 ↑	0.15	0.12	26	58	58

Table 26: Comparative Analysis of Mixed Model Pairs in Multi-Agent Debate Settings on the **ARC-Challenge** Dataset. This table showcases different model combinations and the impact of various **Agent Settings** on accuracy. Δ_1 represents the improvement over the lower-capability model (the first agent), while Δ_2 represents the improvement or decline relative to the higher-capability model (the second agent). Values in parentheses next to each model name indicate the single-agent baseline performance. The table also shows average **Debate Rounds**, normalized **Sycophancy** (per 1172 data points), and transitions between correct (C) and incorrect (I) states, demonstrating how mixed-capability agents interact in debate scenarios.

Agent 1	Agent 2	Agent 3	Agent Settings	Accuracy	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 1172)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Default	35.15	2.82 ↓	2.54	3.14	535	484	283
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Deterministic	40.78	2.81 ↑	0.00	0.00	0	0	0
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Exploratory	35.32	2.65 ↓	3.12	3.54	587	528	303
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	1 Det. & 2 Exp.	37.20	0.77 ↓	2.78	3.19	523	503	306
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	2 Det. & 1 Exp.	38.23	0.26 ↑	1.49	1.75	404	353	219
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Default	72.53	3.32 ↑	0.98	1.29	206	454	343
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Deterministic	67.58	1.63 ↓	0.00	0.00	0	0	0
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Exploratory	72.10	2.89 ↑	1.37	1.85	235	611	433
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	1 Det. & 2 Exp.	71.93	2.72 ↑	1.12	1.53	229	520	386
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	2 Det. & 1 Exp.	70.82	1.61 ↑	0.63	0.93	163	345	245
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Default	85.75	2.22 ↑	0.43	0.43	79	197	156
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Deterministic	84.13	0.60 ↑	0.00	0.00	0	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Exploratory	86.26	2.73 ↑	0.50	0.57	96	229	167
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	1 Det. & 2 Exp.	86.26	2.73 ↑	0.51	0.48	106	193	149
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	2 Det. & 1 Exp.	84.73	1.20 ↑	0.33	0.31	71	131	101
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Default	91.81	4.59 ↑	0.19	0.22	56	84	66
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Deterministic	90.61	3.39 ↑	0.00	0.00	0	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Exploratory	91.72	4.50 ↑	0.23	0.29	66	85	65
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	1 Det. & 2 Exp.	91.04	3.82 ↑	0.22	0.24	60	80	68
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	2 Det. & 1 Exp.	91.30	4.08 ↑	0.14	0.15	40	57	40
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Default	94.20	3.93 ↑	0.12	0.13	27	54	45
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Deterministic	94.37	4.10 ↑	0.00	0.00	0	0	0
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Exploratory	94.80	4.53 ↑	0.10	0.12	28	50	39
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	1 Det. & 2 Exp.	94.54	4.27 ↑	0.09	0.09	22	41	33
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	2 Det. & 1 Exp.	94.71	4.44 ↑	0.06	0.06	10	32	26
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Default	95.82	0.54 ↑	0.07	0.11	22	36	28
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Deterministic	95.73	0.45 ↑	0.00	0.00	0	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Exploratory	95.56	0.28 ↑	0.08	0.12	28	35	32
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	1 Det. & 2 Exp.	95.56	0.28 ↑	0.07	0.10	30	29	25
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	2 Det. & 1 Exp.	95.99	0.71 ↑	0.03	0.04	13	18	14
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Default	88.91	4.18 ↑	0.35	0.61	69	130	104
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Deterministic	88.91	4.18 ↑	0.00	0.00	0	0	0
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Exploratory	88.74	4.01 ↑	0.50	0.83	85	196	151
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	1 Det. & 2 Exp.	88.74	4.01 ↑	0.37	0.61	74	155	121
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	2 Det. & 1 Exp.	89.08	4.35 ↑	0.30	0.52	54	109	81
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Default	75.77	2.65 ↑	0.81	0.80	177	244	190
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Deterministic	74.83	1.71 ↑	0.00	0.00	0	0	0
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Exploratory	75.51	2.39 ↑	0.90	1.00	196	303	210
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	1 Det. & 2 Exp.	75.17	2.05 ↑	0.99	0.91	223	262	192
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	2 Det. & 1 Exp.	75.26	2.14 ↑	0.53	0.43	118	162	117
Mistral-7B	Mistral-7B	Mistral-7B	Default	70.73	1.96 ↑	0.97	0.94	213	292	207
Mistral-7B	Mistral-7B	Mistral-7B	Deterministic	68.26	0.51 ↓	0.00	0.00	0	0	0
Mistral-7B	Mistral-7B	Mistral-7B	Exploratory	71.67	2.90 ↑	1.14	1.20	232	360	249
Mistral-7B	Mistral-7B	Mistral-7B	1 Det. & 2 Exp.	71.25	2.48 ↑	1.03	1.03	209	317	227
Mistral-7B	Mistral-7B	Mistral-7B	2 Det. & 1 Exp.	70.48	1.71 ↑	0.62	0.66	142	214	136
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Default	87.46	9.81 ↑	0.40	0.56	98	145	107
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Deterministic	86.43	8.78 ↑	0.00	0.00	0	0	0
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Exploratory	86.01	8.36 ↑	0.52	0.77	127	187	150
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	1 Det. & 2 Exp.	86.69	9.04 ↑	0.50	0.72	114	174	128
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	2 Det. & 1 Exp.	85.67	8.02 ↑	0.30	0.46	115	119	73

Table 27: Comparative Analysis of Language Model Performance in Multi-Agent Debate Settings on the **ARC-Challenge** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters) on the **Accuracy** of various language models in a three-agent configuration. The Δ column quantifies the **improvement (or decline) over the single base model performance** (shown in parentheses after model names). Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 1172 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent 3	Accuracy	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 1172)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-3B	82.59	0.94 ↓	1.41	1.40	148	820	629
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Llama-3.1-3B	68.00	5.12 ↓	1.66	1.85	311	641	489
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Phi-mini-3.8B	82.76	1.97 ↓	1.48	1.60	170	804	621
Qwen-2.5-0.5B	Qwen-2.5-3B	Llama-3.1-3B	79.69	3.84 ↓	1.62	1.50	208	699	581
Qwen-2.5-0.5B	Qwen-2.5-3B	Phi-mini-3.8B	86.95	2.22 ↑	1.34	1.23	133	722	631
Qwen-2.5-0.5B	Llama-3.1-3B	Phi-mini-3.8B	78.41	6.32 ↓	1.54	1.72	238	683	559
Qwen-2.5-1.5B	Qwen-2.5-3B	Llama-3.1-3B	82.34	1.19 ↓	0.98	1.10	180	447	358
Qwen-2.5-1.5B	Qwen-2.5-3B	Phi-mini-3.8B	87.37	2.64 ↑	0.71	0.81	105	423	358
Qwen-2.5-1.5B	Llama-3.1-3B	Phi-mini-3.8B	81.74	3.00 ↓	0.93	1.19	195	412	341
Qwen-2.5-3B	Phi-mini-3.8B	Llama-3.1-3B	85.67	2.14 ↑	0.84	0.89	143	319	244
Qwen-2.5-3B	Qwen-2.5-3B	Phi-mini-3.8B	87.88	3.15 ↑	0.50	0.52	110	225	170
Qwen-2.5-3B	Phi-mini-3.8B	Phi-mini-3.8B	89.33	4.60 ↑	0.52	0.61	81	214	174
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	69.80	0.59 ↑	1.66	1.77	231	686	523
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-1.5B	55.97	13.24 ↓	2.33	2.69	393	680	451

Table 28: Analysis of Mixed-Model Configurations in Multi-Agent Debate Settings on the **ARC-Challenge** Dataset. This table examines various heterogeneous model combinations in three-agent debate setups. The Δ column quantifies the **improvement (or decline) compared to the best single model performance** among the three agents used in each configuration. All agent combinations use the default settings for temperature and top_p. Metrics include average **Debate Rounds**, normalized **Sycophancy** (per 1172 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C). Results demonstrate that certain model combinations can achieve higher accuracy than their constituent models when debating together.

Agent 1	Agent 2	Agent Settings	Accuracy	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 1221)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Default	39.80	3.31 ↑	1.47	1.11	239.00	306	240
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Deterministic	40.87	4.38 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Exploratory	33.50	2.99 ↓	1.90	1.17	279.00	338	257
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Both: Det. & Exp.	41.93	5.44 ↑	1.64	1.08	251.00	355	289
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Default	67.40	0.88 ↑	0.44	0.34	110.00	154	154
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Deterministic	68.14	1.62 ↑	0.00	0.00	0.00	2	1
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Exploratory	67.24	0.72 ↑	0.60	0.51	143.00	217	201
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Both: Det. & Exp.	66.67	0.15 ↑	0.47	0.41	111.00	166	158
Qwen-2.5-3B	Qwen-2.5-3B	Both: Default	74.37	1.71 ↑	0.37	0.33	85.00	128	123
Qwen-2.5-3B	Qwen-2.5-3B	Both: Deterministic	74.77	2.11 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Both: Exploratory	73.87	1.21 ↑	0.39	0.37	93.00	127	120
Qwen-2.5-3B	Qwen-2.5-3B	Both: Det. & Exp.	75.51	2.85 ↑	0.35	0.25	73.00	127	123
Qwen-2.5-7B	Qwen-2.5-7B	Both: Default	81.57	2.01 ↑	0.15	0.14	38.00	66	64
Qwen-2.5-7B	Qwen-2.5-7B	Both: Deterministic	81.65	2.09 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Both: Exploratory	81.90	2.34 ↑	0.19	0.19	46.00	78	75
Qwen-2.5-7B	Qwen-2.5-7B	Both: Det. & Exp.	82.56	3.00 ↑	0.20	0.19	54.00	62	61
Qwen-2.5-14B	Qwen-2.5-14B	Both: Default	83.37	1.00 ↑	0.15	0.15	34.00	43	41
Qwen-2.5-14B	Qwen-2.5-14B	Both: Deterministic	83.70	1.33 ↑	0.00	0.00	0.00	0	0
Qwen-2.5-14B	Qwen-2.5-14B	Both: Exploratory	83.21	0.84 ↑	0.18	0.19	44.00	66	62
Qwen-2.5-14B	Qwen-2.5-14B	Both: Det. & Exp.	83.87	1.50 ↑	0.16	0.15	40.00	59	54
Qwen-2.5-32B	Qwen-2.5-32B	Both: Default	86.24	0.48 ↑	0.12	0.17	28.00	47	46
Qwen-2.5-32B	Qwen-2.5-32B	Both: Deterministic	85.75	0.01 ↓	0.00	0.00	0.00	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Both: Exploratory	86.24	0.48 ↑	0.14	0.20	34.00	46	43
Qwen-2.5-32B	Qwen-2.5-32B	Both: Det. & Exp.	86.57	0.81 ↑	0.16	0.24	32.00	55	46
Phi-mini-3.8B	Phi-mini-3.8B	Both: Default	71.66	1.78 ↑	0.46	0.68	108.00	100	79
Phi-mini-3.8B	Phi-mini-3.8B	Both: Deterministic	72.24	2.36 ↑	0.00	0.00	0.00	0	0
Phi-mini-3.8B	Phi-mini-3.8B	Both: Exploratory	73.87	3.99 ↑	0.50	0.70	85.00	141	121
Phi-mini-3.8B	Phi-mini-3.8B	Both: Det. & Exp.	73.22	3.34 ↑	0.47	0.66	91.00	124	105
Llama-3.1-3B	Llama-3.1-3B	Both: Default	68.55	3.51 ↑	0.44	0.40	107.00	117	110
Llama-3.1-3B	Llama-3.1-3B	Both: Deterministic	67.40	2.36 ↑	0.00	0.00	0.00	0	0
Llama-3.1-3B	Llama-3.1-3B	Both: Exploratory	66.75	1.71 ↑	0.53	0.48	116.00	131	122
Llama-3.1-3B	Llama-3.1-3B	Both: Det. & Exp.	67.73	2.69 ↑	0.47	0.45	105.00	113	109
Mistral-7B	Mistral-7B	Both: Default	66.34	1.79 ↑	0.30	0.22	57.00	64	57
Mistral-7B	Mistral-7B	Both: Deterministic	66.99	2.44 ↑	0.00	0.00	0.00	0	0
Mistral-7B	Mistral-7B	Both: Exploratory	65.11	0.56 ↑	0.38	0.30	81.00	85	80
Mistral-7B	Mistral-7B	Both: Det. & Exp.	66.42	1.87 ↑	0.34	0.25	62.00	89	81
Llama-3.1-8B	Llama-3.1-8B	Both: Default	74.28	1.26 ↑	0.41	0.47	79.00	114	106
Llama-3.1-8B	Llama-3.1-8B	Both: Deterministic	75.43	2.41 ↑	0.00	0.00	0.00	2	1
Llama-3.1-8B	Llama-3.1-8B	Both: Exploratory	74.86	1.84 ↑	0.46	0.54	95.00	139	130
Llama-3.1-8B	Llama-3.1-8B	Both: Det. & Exp.	74.45	1.43 ↑	0.41	0.48	99.00	112	102

Table 29: Comparative Analysis of Language Model Performance in Multi-Agent Debate Settings on the **CommonsenseQA** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters like Default, Deterministic, Exploratory, and a combination) on the **Accuracy** of various language models. The Δ column quantifies the **improvement (or decline) over the single base model performance**. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 1221 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent Settings	Accuracy	Δ_1	Δ_2	Debate Rounds (Avg)	Sycophancy (Avg / 1221)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Default	56.92	20.43 ↑	9.60 ↓	1.34	0.84	237.00	370	345
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Deterministic	58.39	21.90 ↑	8.13 ↓	1.26	0.63	148.00	326	295
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Exploratory	56.91	20.42 ↑	9.61 ↓	1.63	0.99	216.00	430	377
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Det. & Exp.	57.08	20.59 ↑	9.44 ↓	1.28	0.82	177.00	371	332
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Both: Exp. & Det.	57.49	21.00 ↑	9.03 ↓	1.51	0.87	206.00	407	379
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Default	66.83	0.31 ↑	1.79 ↑	0.59	0.63	168.00	170	165
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Deterministic	68.63	2.11 ↑	3.59 ↑	0.66	0.80	160.00	198	184
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Exploratory	67.08	0.56 ↑	2.04 ↑	0.82	0.90	164.00	237	223
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Det. & Exp.	69.78	3.26 ↑	4.74 ↑	0.61	0.69	140.00	203	193
Qwen-2.5-1.5B	Llama-3.1-3B	Both: Exp. & Det.	67.73	1.21 ↑	2.69 ↑	0.66	0.72	160.00	219	200
Qwen-2.5-3B	Phi-mini-3.8B	Both: Default	75.02	2.36 ↑	5.14 ↑	0.44	0.39	100.00	158	150
Qwen-2.5-3B	Phi-mini-3.8B	Both: Deterministic	76.09	3.43 ↑	6.21 ↑	0.50	0.37	104.00	161	154
Qwen-2.5-3B	Phi-mini-3.8B	Both: Exploratory	74.69	2.03 ↑	4.81 ↑	0.50	0.52	85.00	177	167
Qwen-2.5-3B	Phi-mini-3.8B	Both: Det. & Exp.	75.76	3.10 ↑	5.88 ↑	0.52	0.40	114.00	191	179
Qwen-2.5-3B	Phi-mini-3.8B	Both: Exp. & Det.	75.10	2.44 ↑	5.22 ↑	0.49	0.49	106.00	162	156
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Default	73.87	7.35 ↑	1.21 ↑	0.51	0.47	100.00	225	217
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Deterministic	74.94	8.42 ↑	2.28 ↑	0.48	0.40	108.00	191	187
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Exploratory	74.12	7.60 ↑	1.46 ↑	0.60	0.55	115.00	279	264
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Det. & Exp.	74.04	7.52 ↑	1.38 ↑	0.51	0.52	106.00	208	204
Qwen-2.5-1.5B	Qwen-2.5-3B	Both: Exp. & Det.	74.94	8.42 ↑	2.28 ↑	0.57	0.42	108.00	251	246
Llama-3.1-3B	Llama-3.1-8B	Both: Default	72.24	7.20 ↑	0.78 ↓	0.54	0.52	119.00	165	153
Llama-3.1-3B	Llama-3.1-8B	Both: Deterministic	73.79	8.75 ↑	0.77 ↑	0.57	0.57	118.00	190	183
Llama-3.1-3B	Llama-3.1-8B	Both: Exploratory	72.15	7.11 ↑	0.87 ↓	0.59	0.58	112.00	167	157
Llama-3.1-3B	Llama-3.1-8B	Both: Det. & Exp.	70.68	5.64 ↑	2.34 ↓	0.60	0.58	131.00	162	154
Llama-3.1-3B	Llama-3.1-8B	Both: Exp. & Det.	73.96	8.92 ↑	0.94 ↑	0.60	0.61	120.00	200	193
Qwen-2.5-7B	Qwen-2.5-14B	Both: Default	83.37	3.81 ↑	1.00 ↑	0.28	0.26	62.00	98	96
Qwen-2.5-7B	Qwen-2.5-14B	Both: Deterministic	83.78	4.22 ↑	1.41 ↑	0.33	0.21	71.00	101	95
Qwen-2.5-7B	Qwen-2.5-14B	Both: Exploratory	84.19	4.63 ↑	1.82 ↑	0.28	0.27	60.00	112	110
Qwen-2.5-7B	Qwen-2.5-14B	Both: Det. & Exp.	83.37	3.81 ↑	1.00 ↑	0.29	0.24	66.00	103	99
Qwen-2.5-7B	Qwen-2.5-14B	Both: Exp. & Det.	83.29	3.73 ↑	0.92 ↑	0.28	0.21	66.00	95	93

Table 30: Comparative Analysis of Mixed Language Model Performance in Multi-Agent Debate Settings on the **CommonsenseQA** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters) on the **Accuracy** when pairing different language models. The Δ_1 column shows the improvement over the weaker model's performance, while Δ_2 shows comparison to the stronger model. This highlights whether mixed-agent debates benefit from model complementarity or are constrained by the weaker model's capabilities. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 1221 data points), and transitions between correct (C) and incorrect (I) states.

Agent 1	Agent 2	Agent 3	Agent Settings	Accuracy	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 1221)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Default	37.76	1.27 ↑	2.69	3.02	545	538	327
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Deterministic	39.80	3.31 ↑	0.00	0.00	0	0	0
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	Exploratory	32.60	3.89 ↓	3.45	3.66	580	604	336
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	1 Det. & 2 Exp.	36.77	0.28 ↑	3.05	3.11	569	558	317
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-0.5B	2 Det. & 1 Exp.	37.51	1.02 ↑	1.76	1.84	433	420	237
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Default	68.80	2.28 ↑	0.77	0.83	193	333	264
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Deterministic	67.90	1.38 ↑	0.00	0.00	0	3	1
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	Exploratory	67.57	1.05 ↑	1.14	1.34	256	429	315
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	1 Det. & 2 Exp.	68.55	2.03 ↑	0.92	1.01	211	346	270
Qwen-2.5-1.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	2 Det. & 1 Exp.	68.55	2.03 ↑	0.57	0.57	172	244	179
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Default	75.18	2.52 ↑	0.63	0.68	147	225	180
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Deterministic	74.28	1.62 ↑	0.00	0.00	0	0	0
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	Exploratory	74.37	1.71 ↑	0.66	0.82	164	248	196
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	1 Det. & 2 Exp.	75.02	2.36 ↑	0.67	0.66	166	211	163
Qwen-2.5-3B	Qwen-2.5-3B	Qwen-2.5-3B	2 Det. & 1 Exp.	75.76	3.10 ↑	0.45	0.44	116	163	115
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Default	81.90	2.34 ↑	0.31	0.38	85	122	96
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Deterministic	81.57	2.01 ↑	0.00	0.00	0	0	0
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	Exploratory	81.98	2.42 ↑	0.38	0.47	99	147	117
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	1 Det. & 2 Exp.	81.41	1.85 ↑	0.32	0.38	98	124	99
Qwen-2.5-7B	Qwen-2.5-7B	Qwen-2.5-7B	2 Det. & 1 Exp.	81.74	2.18 ↑	0.25	0.26	84	89	65
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Default	83.05	0.68 ↑	0.27	0.28	84	85	69
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Deterministic	83.87	1.50 ↑	0.00	0.00	0	0	0
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	Exploratory	83.13	0.76 ↑	0.28	0.33	76	100	75
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	1 Det. & 2 Exp.	83.54	1.17 ↑	0.25	0.25	74	93	77
Qwen-2.5-14B	Qwen-2.5-14B	Qwen-2.5-14B	2 Det. & 1 Exp.	83.95	1.58 ↑	0.14	0.12	45	56	46
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Default	86.00	0.24 ↑	0.18	0.26	61	80	67
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Deterministic	85.75	0.01 ↓	0.00	0.00	0	0	0
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	Exploratory	86.57	0.81 ↑	0.18	0.25	56	87	74
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	1 Det. & 2 Exp.	86.00	0.24 ↑	0.16	0.21	61	71	57
Qwen-2.5-32B	Qwen-2.5-32B	Qwen-2.5-32B	2 Det. & 1 Exp.	86.08	0.32 ↑	0.11	0.14	35	50	41
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Default	73.22	3.34 ↑	0.62	1.12	170	171	121
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Deterministic	73.71	3.83 ↑	0.00	0.00	0	0	0
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	Exploratory	73.96	4.08 ↑	0.74	1.24	161	231	170
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	1 Det. & 2 Exp.	75.18	5.30 ↑	0.69	1.21	134	217	159
Phi-mini-3.8B	Phi-mini-3.8B	Phi-mini-3.8B	2 Det. & 1 Exp.	73.71	3.83 ↑	0.47	0.86	107	137	97
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Default	68.39	3.35 ↑	0.87	0.92	210	237	169
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Deterministic	68.06	3.02 ↑	0.00	0.00	0	0	0
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	Exploratory	67.65	2.61 ↑	1.04	1.16	250	261	190
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	1 Det. & 2 Exp.	67.08	2.04 ↑	0.89	0.95	213	225	165
Llama-3.1-3B	Llama-3.1-3B	Llama-3.1-3B	2 Det. & 1 Exp.	67.73	2.69 ↑	0.58	0.58	132	149	105
Mistral-7B	Mistral-7B	Mistral-7B	Default	66.83	2.28 ↑	0.53	0.57	121	137	99
Mistral-7B	Mistral-7B	Mistral-7B	Deterministic	66.75	2.20 ↑	0.00	0.00	0	0	0
Mistral-7B	Mistral-7B	Mistral-7B	Exploratory	65.60	1.05 ↑	0.79	0.83	179	167	119
Mistral-7B	Mistral-7B	Mistral-7B	1 Det. & 2 Exp.	65.44	0.89 ↑	0.64	0.70	157	144	97
Mistral-7B	Mistral-7B	Mistral-7B	2 Det. & 1 Exp.	66.75	2.20 ↑	0.32	0.35	81	98	68
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Default	75.92	2.90 ↑	0.62	0.83	147	211	148
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Deterministic	75.84	2.82 ↑	0.00	0.00	0	9	3
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	Exploratory	74.12	1.10 ↑	0.79	1.13	203	246	168
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	1 Det. & 2 Exp.	75.51	2.49 ↑	0.71	0.94	173	233	161
Llama-3.1-8B	Llama-3.1-8B	Llama-3.1-8B	2 Det. & 1 Exp.	75.51	2.49 ↑	0.44	0.60	118	150	92

Table 31: Comparative Analysis of Language Model Performance in Multi-Agent Debate Settings on the **CommonsenseQA** Dataset. This table showcases the impact of different **Agent Settings** (controlling temperature and top_p parameters like Default, Deterministic, Exploratory, and combinations) on the **Accuracy** of various language models. The Δ column quantifies the **improvement (or decline) over the single base model performance**. Further metrics include average **Debate Rounds**, normalized **Sycophancy** (per 1221 data points), and transitions between correct (C) and incorrect (I) states (C→I, I→C), highlighting the nuanced effects of debate dynamics.

Agent 1	Agent 2	Agent 3	Accuracy	Δ	Debate Rounds (Avg)	Sycophancy (Avg / 1221)	C→I	I→C	Debate Helped (Overall)
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-3B	72.48	35.99 ↑	1.64	1.51	228	748	563
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Llama-3.1-3B	65.03	28.54 ↑	1.81	1.89	343	622	480
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Phi-mini-3.8B	70.60	34.11 ↑	1.68	1.73	246	691	537
Qwen-2.5-0.5B	Qwen-2.5-3B	Llama-3.1-3B	72.56	36.07 ↑	1.81	1.59	234	697	544
Qwen-2.5-0.5B	Qwen-2.5-3B	Phi-mini-3.8B	72.15	35.66 ↑	1.66	1.59	243	629	517
Qwen-2.5-0.5B	Llama-3.1-3B	Phi-mini-3.8B	69.12	32.63 ↑	1.76	1.91	298	617	483
Qwen-2.5-1.5B	Qwen-2.5-3B	Llama-3.1-3B	73.38	6.86 ↑	1.08	1.22	230	399	305
Qwen-2.5-1.5B	Qwen-2.5-3B	Phi-mini-3.8B	75.68	9.16 ↑	0.95	1.17	202	382	303
Qwen-2.5-1.5B	Llama-3.1-3B	Phi-mini-3.8B	71.09	4.57 ↑	1.04	1.42	260	347	273
Qwen-2.5-3B	Phi-mini-3.8B	Llama-3.1-3B	74.20	1.54 ↑	1.00	1.15	222	334	253
Qwen-2.5-3B	Qwen-2.5-3B	Phi-mini-3.8B	74.77	2.11 ↑	0.73	0.84	200	256	193
Qwen-2.5-3B	Phi-mini-3.8B	Phi-mini-3.8B	76.09	3.43 ↑	0.85	1.18	183	258	186
Qwen-2.5-0.5B	Qwen-2.5-1.5B	Qwen-2.5-1.5B	64.86	28.37 ↑	1.86	1.50	267	576	447
Qwen-2.5-0.5B	Qwen-2.5-0.5B	Qwen-2.5-1.5B	55.12	18.63 ↑	2.41	2.44	384	651	438

Table 32: Comparative Analysis of Mixed Language Model Performance in Multi-Agent Debate Settings on the **CommonsenseQA** Dataset. This table presents results for heterogeneous combinations of language models in debate settings. The Δ column quantifies the improvement over the performance of the weakest model in each combination (for combinations with Qwen-2.5-0.5B, the baseline is 36.49%; for others, the baseline corresponds to the lowest-performing model). All experiments use the default debate setting. The table shows that combining models of different capacities can lead to significant performance gains, especially when smaller models are paired with larger ones.

F Additional Results and Ablations

F.1 Original MAD Results

We also report our experiments with the original Multi-Agent Debate (MAD) framework across various model sizes and architectures. Table 33 presents the results on three challenging reasoning benchmarks: GSM-Plus, GSM8K, and ARC-Challenge.

F.2 Majority Vote@3 Results

To further investigate the impact of stochastic diversity on model performance, we report results on a Majority Vote@3 approach where we sample three independent responses from each model and take a majority vote to determine the final answer. Table 34 presents these results across five benchmarks: GSM8K, GSM-Plus, ARC-Easy, ARC-Challenge, and CommonsenseQA.

The results demonstrate that simple ensemble-based approaches can significantly boost performance without requiring multi-agent debate or model fine-tuning. Across all model sizes and architectures, Majority Vote@3 consistently outperforms single-sample inference. The relative improvements are most pronounced for smaller models, with Qwen-2.5-0.5B gaining up to 4.27 percentage points on ARC-Challenge and Qwen-2.5-1.5B showing similar substantial improvements across benchmarks.

Interestingly, this pattern holds across model families. Llama-3.1-3B, Phi-3.5-mini, and Mistral-7B all exhibit significant gains when using majority voting, suggesting that the benefits of ensemble diversity transcend specific model architectures. The results also indicate diminishing returns for larger models—Qwen-2.5-14B shows more modest improvements compared to its smaller counterparts, likely because these larger models already produce more consistent answers across samples.

These findings highlight an important baseline for our research: simple ensemble methods provide strong performance improvements with minimal computational overhead during inference. However, they still require multiple forward passes for each query, motivating our DTE approach that aims to distill these benefits into a single model through training on debate traces.

F.3 Scaling Results for Multiple Agents

We investigated how performance scales with increasing numbers of debating agents (1-7) across different model sizes and reasoning benchmarks. Table 35 presents these results, revealing several important trends in multi-agent scaling behavior.

First, we observe that performance generally improves as we add more agents to the debate, but with diminishing returns. The most significant gains occur when moving from a single agent (equivalent to standard inference) to two agents, with more modest improvements as additional agents join the debate. For example, on GSM8K, Qwen-2.5-1.5B shows a substantial jump from 62.77% (1 agent) to 71.57% (2 agents), but only incremental improvements thereafter.

Second, the benefits of additional agents vary across tasks. On more complex tasks like GSM-Plus, we see continued performance improvements even with 7 agents, particularly for larger models. Qwen-2.5-14B shows its peak GSM-Plus performance with 7 agents (78.08%), suggesting that more difficult problems benefit from extended multi-agent collaboration. In contrast, on simpler tasks like ARC-Easy, performance plateaus more quickly.

Third, we find that model size influences scaling behavior. Smaller models like Qwen-2.5-1.5B show more variability in performance as agents are added, with occasional performance drops when moving from 3 to 4 agents. Larger models exhibit more stable scaling patterns, suggesting that they can more consistently integrate insights from multiple debate participants.

These results have important implications for our DTE framework. They demonstrate that while adding more agents generally improves performance, the computational costs may outweigh the benefits beyond 3-5 agents for most applications. This insight helped inform our design choices in balancing performance gains against computational efficiency in our final framework.

Model Configuration			Debate Performance Metrics						
Agent 1	**Agent 2**	**Debate Setting**	**Accuracy**	**Delta**	**Debate Rounds**	**Sycophancy**	**Correct→Incorrect**	**Incorrect→Correct**	**Net Benefit**
				GSM-Plus					
Qwen-2.5-0.5B	Qwen-2.5-0.5B	exploratory	28.12%	3.33 ↑	3.48	6906	261	575	432
Qwen-2.5-1.5B	Qwen-2.5-1.5B	exploratory	46.50%	4.50 ↑	2.33	5642	194	861	670
Qwen-2.5-3B	Qwen-2.5-3B	exploratory	66.79%	5.04 ↑	1.34	5315	231	373	187
Qwen-2.5-7B	Qwen-2.5-7B	exploratory	69.71%	1.09 ↑	0.76	2967	102	200	121
Qwen-2.5-14B	Qwen-2.5-14B	exploratory	76.92%	5.13 ↑	0.61	2722	119	151	47
Phi-mini-3.8B	Phi-mini-3.8B	exploratory	65.79%	2.37 ↑	1.07	3620	180	272	136
Llama-3.1-3B	Llama-3.1-3B	exploratory	42.42%	3.25 ↓	2.07	5507	379	369	238
Mistral-7B	Mistral-7B	exploratory	26.35%	11.31 ↑	1.85	4500	210	290	115
Llama-3.1-8B	Llama-3.1-8B	exploratory	57.63%	2.01 ↑	1.75	5667	273	585	351
				GSM8K					
Qwen-2.5-0.5B	Qwen-2.5-0.5B	exploratory	45.56%	3.56 ↑	2.85	3469	175	427	328
Qwen-2.5-1.5B	Qwen-2.5-1.5B	exploratory	65.81%	3.04 ↑	1.99	3471	144	650	489
Qwen-2.5-3B	Qwen-2.5-3B	exploratory	86.96%	1.82 ↑	0.63	1390	82	165	97
Qwen-2.5-7B	Qwen-2.5-7B	exploratory	91.74%	1.07 ↑	0.38	930	64	93	33
Qwen-2.5-14B	Qwen-2.5-14B	exploratory	94.39%	1.59 ↑	0.18	448	30	48	18
Phi-mini-3.8B	Phi-mini-3.8B	exploratory	88.17%	1.29 ↑	0.45	1050	65	120	65
Llama-3.1-3B	Llama-3.1-3B	exploratory	67.63%	4.92 ↓	1.51	2418	238	215	127
Mistral-7B	Mistral-7B	exploratory	43.44%	22.06 ↑	1.65	2100	175	235	95
Llama-3.1-8B	Llama-3.1-8B	exploratory	83.02%	1.29 ↑	0.94	1587	93	308	236
				ARC-Challenge					
Qwen-2.5-0.5B	Qwen-2.5-0.5B	exploratory	38.65%	0.68 ↑	1.88	2728	272	308	232
Qwen-2.5-1.5B	Qwen-2.5-1.5B	exploratory	74.15%	0.94 ↑	0.85	1671	121	231	156
Qwen-2.5-3B	Qwen-2.5-3B	exploratory	85.41%	1.88 ↑	0.57	1227	94	135	57
Qwen-2.5-7B	Qwen-2.5-7B	exploratory	91.47%	6.25 ↑	0.23	501	41	49	13
Qwen-2.5-14B	Qwen-2.5-14B	exploratory	94.54%	4.27 ↑	0.15	326	31	37	9
Phi-mini-3.8B	Phi-mini-3.8B	exploratory	87.46%	2.73 ↑	0.15	313	24	47	25
Llama-3.1-3B	Llama-3.1-3B	exploratory	76.37%	3.25 ↑	0.73	1525	111	155	69
Mistral-7B	Mistral-7B	exploratory	73.29%	4.52 ↑	0.40	795	63	114	73
Llama-3.1-8B	Llama-3.1-8B	exploratory	86.09%	8.44 ↑	0.27	514	31	84	58

Table 33: Performance of the original Multi-Agent Debate (MAD) framework across different model sizes and reasoning benchmarks. Results show accuracy, improvement over single-agent baseline (Delta), average debate rounds, and debate transition statistics. The Delta column highlights performance changes compared to individual model accuracy, with green indicating improvement and red indicating decline.

Model	**Accuracy (%) on Benchmarks**				
	GSM8K	**GSM-Plus**	**ARC-E**	**ARC-C**	**CQA**
Qwen-2.5-0.5B	49.73	30.54	58.71	42.92	42.51
Qwen-2.5-1.5B	75.82	52.08	87.12	73.55	69.62
Qwen-2.5-3B	86.28	64.08	94.19	84.13	76.90
Qwen-2.5-7B	92.19	70.46	96.46	91.21	82.88
Qwen-2.5-14B	94.09	72.54	98.44	94.20	82.15
Llama-3.1-3B	77.03	52.79	88.51	75.00	69.94
Llama-3.1-8B	85.82	60.88	93.56	83.11	74.86
Phi-3.5-mini	87.87	65.79	96.00	86.95	75.10
Mistral-7B	56.86	36.88	87.58	75.68	69.04

Table 34: Performance comparison using Majority Vote@3 approach across different benchmarks. For each model, we sample three independent responses and determine the final answer through majority voting.

Table 35: Performance scaling with increasing numbers of debating agents (1-7) across different model sizes and reasoning benchmarks. Results show accuracy percentages for each configuration.

Model	Number of Agents						
	1	**2**	**3**	**4**	**5**	**6**	**7**
			GSM8K Accuracy (%)				
Qwen-2.5-1.5B	62.77	71.57	75.13	75.89	75.13	74.98	76.50
Qwen-2.5-3B	85.14	85.52	87.64	87.11	87.04	86.66	87.11
Qwen-2.5-7B	90.67	91.21	92.42	92.49	92.57	92.34	92.72
Qwen-2.5-14B	92.80	93.33	94.84	94.31	94.69	94.62	94.24
			GSM-Plus Accuracy (%)				
Qwen-2.5-1.5B	42.00	51.62	53.33	50.62	54.21	51.50	52.67
Qwen-2.5-3B	61.75	67.79	68.00	64.21	69.71	64.88	68.54
Qwen-2.5-7B	68.62	74.17	74.96	70.88	71.08	71.38	76.00
Qwen-2.5-14B	71.79	77.25	72.29	72.83	73.29	73.38	78.08
			ARC-Challenge Accuracy (%)				
Qwen-2.5-1.5B	69.21	68.52	72.10	71.50	72.53	71.50	72.10
Qwen-2.5-3B	82.53	84.64	86.26	85.75	86.26	86.95	87.03
Qwen-2.5-7B	87.22	91.64	91.72	91.47	92.06	91.38	92.32
Qwen-2.5-14B	90.27	93.77	94.80	95.14	94.20	94.62	94.28
			ARC-Easy Accuracy (%)				
Qwen-2.5-1.5B	86.62	83.42	85.61	86.32	87.46	86.57	87.16
Qwen-2.5-3B	93.06	94.15	94.28	94.32	94.82	94.91	94.99
Qwen-2.5-7B	94.69	96.93	96.55	96.34	96.42	96.25	96.59
Qwen-2.5-14B	95.66	98.15	98.19	98.23	98.15	98.19	98.23