Can Lessons From Human Teams Be Applied to Multi-Agent Systems? The Role of Structure, Diversity, and Interaction Dynamics

Anonymous ACL submission

Abstract

Multi-Agent Systems (MAS) with Large Language Model (LLM)-powered agents are gaining attention, yet fewer studies explore their team dynamics. Inspired by human team science, we propose a multi-agent framework to examine core aspects of team science: structure (flat vs. hierarchical teams), diversity (via demographic personas), and interaction dynamics (through pre-/post-task interviews and GPT-4o-based conversation analysis). We evaluate team performance across four tasks: CommonsenseQA, StrategyQA, Social IQa, and Latent Implicit Hate, spanning commonsense and social reasoning. Our results show that flat teams tend to perform better than hierarchical ones, while diversity has a nuanced impact. Interviews suggest agents are overconfident about their team performance, yet post-task reflections reveal both appreciation for collaboration and challenges in integration. GPT-4o analysis highlights limited conversational coordination among agents.

1 Introduction

Large Language Models (LLMs)' growing ability to process, generate, and reason with natural language has driven interest in designing multi-agent systems (MAS)—collections of AI agents collaborating on complex problems. These systems offer several advantages: supporting distributed problem-solving, representing diverse viewpoints, and simulating collaborative dynamics such as debate, negotiation, and cooperation (Du et al., 2023; Chen et al., 2024b; Li et al., 2024; Zhu et al., 2025; Zhang et al., 2024b; Wang et al., 2025). MAS allows us to explore social phenomena and study interaction dynamics that mirror human team behavior. However, few studies examine agent structures, diversity effects, and interactions, despite their potential efficiency and adaptability (Wu and Ito, 2025; Bettini et al., 2025; Li et al., 2021). Well-designed structure and diversity can also foster trust and alignment in human-AI collaboration (Stahl and Maznevski, 2021; Delice et al., 2019; Hattori and Yamada, 2023; McGrath et al., 2024).

In addition, recent work on human-AI collaboration points to the role of coordination and communication (Stahl and Maznevski, 2021; Yang et al., 2024; Agashe et al., 2025; Li et al.). To better understand collaboration in AI teams, we turn to insights from human team science. This literature emphasizes the importance of team structure, including how authority and communication are organized, and diversity in terms of members' backgrounds and perspectives. It further stresses that collaboration depends not only on outcomes but also on how team members *understand*, *coordinate*, and *reason* together.

This leads to our central question: *Can principles from team science help us design more effective AI teams?* To explore this, we ground our study in theories from organizational science. Prior work suggests that flat team structures encourage open communication and trust, while hierarchical structures can expedite decision-making through defined roles (Ji and Yan, 2020; Greer et al., 2018). Diversity-performance theory further suggests that teams with diverse backgrounds can outperform homogeneous ones by bringing in broader perspectives (Cox and Blake, 1991; Pelled et al., 1999; van Knippenberg et al., 2020). Building on these foundations, we propose three research questions:

- **RQ1 (Structure)**: How does team structure (flat versus hierarchical) affect team performance across reasoning and inference tasks?
- **RQ2 (Diversity):** How does demographic diversity, instantiated via agent personas, influence team performance, and does its impact vary by team structure?
- **RQ3 (Interaction):** How do agents perceive their roles and interactions within the team, and what do their communication patterns reveal about coordination, understanding, and reasoning?

To address these questions, we simulate flat and hierarchical teams of LLM agents, each assigned demographic personas (e.g., age, race, gender, occupation), and evaluate them on four tasks requiring reasoning, social inference, and normative judgment: CommonsenseQA (Talmor et al., 2019), StrategyQA (Geva et al., 2021), Social IQa (Sap et al., 2019), and Latent Implicit Hate Detection (ElSherief et al., 2021). These tasks are selected for their reliance on nuanced reasoning, diverse perspectives, and value-sensitive judgment, as these factors are likely influenced by team structure and diversity.

Together, this study offers a theory-driven investigation of how structure and diversity shape both performance and internal dynamics of AI teams. Our findings show that these dimensions significantly impact how agents interact, reason, and coordinate. This, in turn, offers design insights for building more interpretable, collaborative, and socially aware AI teams. Our contributions are:

- A framework for building structured multi-agent LLM teams with demographic personas.
- A comprehensive evaluation including quantitative performance and qualitative interaction analysis.
- Empirical findings on how team structure and composition affect reasoning and social inference tasks.
- Theoretical implications for MAS with LLMs design, demonstrating that communication structure and social framing mediate reasoning and coordination.

2 Background

2.1 Multi-Agent Frameworks for LLMs

MAS are collections of intelligent agents that interact in a shared environment to achieve individual and collective goals. A defining feature of MAS is interaction—the ability to communicate, coordinate, and negotiate to accomplish tasks. In Natural Language Processing (NLP), MAS enable advanced problem solving in commonsense reasoning and social understanding (Hegazy, 2024; Wang et al., 2023; Xu et al., 2023).

A prominent paradigm is multi-agent debate, where multiple LLMs engage in structured argumentation to improve factual accuracy, identify reasoning failures, and simulate consensus (Chen et al., 2024b; Du et al., 2023; Liang et al., 2024). Another line of work explores hierarchical teams, modeling organizational structures with chains of command and task delegation (Wang et al., 2025; Zhu et al., 2025). A complementary trend assigns social characteristics to agents, such as personality traits or demographics, to study emergent behaviors. Studies have shown that incorporating social characteristics in MAS with LLMs show human-like social phenomena through communication, interaction and collaboration (Park et al., 2023; Chuang et al., 2024a; Zhang et al., 2024a; Chuang et al., 2024b; Chen et al., 2024a; Jiang et al., 2024; Samuel et al., 2024; Park et al., 2024; Sahu et al., 2021).

2.2 Insights from Team Science

To guide our investigation, we turn to team science, a multidisciplinary field that examines the factors driving effective collaboration. Decades of research emphasize two key determinants of team performance: structure and diversity (Ji and Yan, 2020; Xu et al., 2022; Cooke and Hilton, 2015; Horwitz and Horwitz, 2007; Salas et al., 2008; Cox and Blake, 1991).

Team structure affects information flow, decision-making, and conflict resolution (Hackman, 2002; Salas et al., 2008). Two common structures are: (1) flat, with decentralized decision-making, which fosters openness but can lack accountability and scalability and (2) hierarchical, with clear authority layers, which improves coordination but risks communication silos (Greer et al., 2018).

Team diversity, encompassed in demographic, cognitive, and functional differences, can enhance team efficacy. Diverse teams are often more innovative (Horwitz and Horwitz, 2007; van Knippenberg et al., 2020), avoid cognitive traps, and excel in logical reasoning and social inference (Roberge and van Dick, 2010). Yet, they may also face communication barriers and increased conflict (Cox and Blake, 1991).

Interaction dynamics, such as communication patterns, coordination mechanisms, and leadership styles, are essential for team success. Research shows that trust calibration, role negotiation, and adaptive communication significantly shape outcomes in both human and human-AI teams (Stahl and Maznevski, 2021).

LLM-based MAS provide a promising testbed to explore how structure, diversity, and interaction dynamics affect performance on NLP tasks. While MAS have been used in commonsense and social reasoning, few studies systemically compare the effects of structure and diversity on team outcomes.

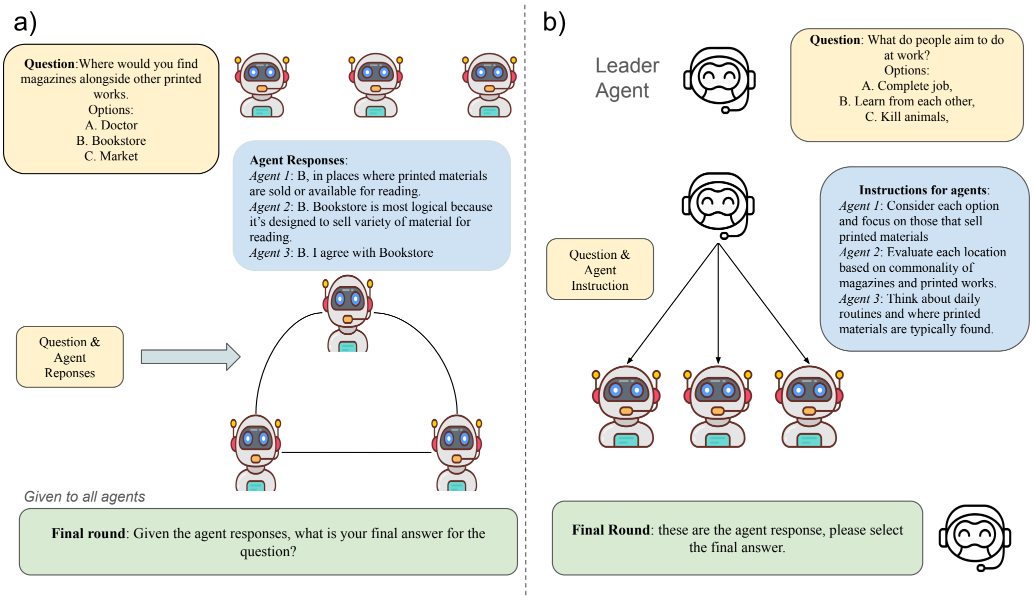

Figure 1: Conversation flows in (a) flat and (b) hierarchical teams. In flat teams, agents respond independently and iteratively refine their answers. In hierarchical teams, leader agents issue instructions and determine the final answer based on others' responses.

3 Multi-Agent Team Design

Team science identifies structure and diversity as key to human collaboration. We operationalize these theoretical constructs into multi-agent design, examining how structure, diversity, and communication shape AI team behavior and effectiveness.

3.1 Team Structure (RQ1)

Flat and hierarchical structures are two central organization forms in team science. In our study, flat teams consist of 3, 5, or 7 agents, odd numbers to enable majority voting without ties. Teams engage in a 2-4 round debate. As in Fig. 1 a), in Round 0, agents answer independently, storing responses in shared memory. In subsequent rounds, agents review previous responses and revise or reaffirm their answers while acknowledging others. In the final round, each agent submits a final judgment, and the team's decision is made by majority vote.

Hierarchical teams follow a top-down communication structure, with designated leaders responsible for delegating tasks and synthesizing responses. We design two variants: 1) a 4-agent team with one leader and three subordinates; 2) a 7-agent team with one leader, two managers, and four subordinates (two under each manager).

As shown in Fig 1 b), in Round 0, the leader receives the question and issues tailored instructions to each agent, simulating division of labor, specialization, and perspective diversification. These tailored instructions guide how agents interpret the question, which aspect to focus on, and what reasoning strategy to use (e.g., "focus on edge cases," "consider the most probable answer first," and "identify counterexamples or contradictions"). In the 7-agent setting, the leader sends meta-instructions to managers (e.g., "gather diverse reasoning paths" or "probe conflicting assumptions"), who relay specific directives to subordinates. Agents respond based on these instructions, and their outputs are routed back to the leader. In later rounds, the leader refines guidance or resolves inconsistencies. In the final round, the leader reviews all inputs and makes the team's final decision, potentially overriding the majority to reflect hierarchical veto power.

3.2 Team Diversity (RQ2)

Team science emphasizes the role of diversity, particularly in demographic and experiential attributes, as a key determinant of team performance. To examine its effect on reasoning and coordination in LLM-based teams, we assign each agent a persona that reflects human demographics and systematically test teams with varying compositions.

Each persona is defined along four dimensions: gender (male, female), age (young adult, young working professional, working professional, senior), ethnicity (White, Black, Asian), and occupation (white- or blue-collar). These dimensions are well-established markers of social identity known to influence communication, authority, and decision-making in human teams (Kunze and Hampel, 2022; Joshi and Roh, 2009; Song and Li, 2020).

3.3 Interaction Dynamics (RQ3)

Beyond structure and diversity, team science highlights the critical role of interaction dynamics, including how members communicate, coordinate, and reflect on their roles. To capture these aspects in multi-agent settings, we incorporate pre- and post-task self-assessments and adopt an *LLM-as-judge* approach, using GPT-4o to score team comprehension and conversation quality. Specifically, we assess understanding of team goals, perceived role clarity, and reasoning process. The qualitative feedback complements our quantitative measures and offers deeper insights into intra-team coordination.

Each agent is asked to answer the following questions before the task. Q_1^{pre} and Q_2^{pre} are open-ended, while Q_3^{pre}–Q_5^{pre} use a 1–5 scale (5 = highest):

Q_1^{pre}. What do you think is the primary goal of the team?

Q_2^{pre}. What is your role in the team?

Q_3^{pre}. How confident are you about executing the role?

Q_4^{pre}. How confident are you in your team executing the task?

Q_5^{pre}. How confident are you in the team's ability to integrate diverse perspectives during the task?

These questions gauge initial expectations about team goals, individual readiness, and perceived inclusiveness.

After the task, we conduct a follow-up interview to assess how the team experience may have shifted perceptions. Agents respond to the following on a 1-5 scale (5 = highest):

Q_1^{post}. How do you think your team performed to achieve the goal?

Q_2^{post}. How well do you think you contributed to the team?

Q_3^{post}. How well do you think your team members contributed to the team?

Q_4^{post}. Were you able to understand your team members?

Q_5^{post}. Do you think your team members understood you?

Q_6^{post}. Do you think you could come up with these solutions that the group came with?

Together, these interviews offer a window into internal team by measuring confidence, role clarity, and perceived synergy. This reflective process helps assess how well agents align in understanding and coordination.

While interviews offer some insight into agent interaction, they do not fully capture the quality of agent-to-agent communication. To better evaluate these dynamics, we adopt an *LLM-as-judge* approach, using GPT-4o to score sample team conversations across five dimensions. Each conversation is rated on a 1-5 scale (5 = highest):

Q_1^{judge}. How well do the agents understand each other and collectively complete the task?

Q_2^{judge}. How well do the agents coordinate, delegate tasks and integrate ideas?

Q_3^{judge}. How strong is the team's reasoning compared to what an individual agent might produce?

Q_4^{judge}. How clear, coherent and logically structure is the conversation?

Q_5^{judge}. How confident are you in the team's final answer based on their reasoning?

This provides a complementary view of how agents engage, reason together, and coordinate toward shared goals, beyond what is captured in interviews or performance metrics.

4 Experiment settings

Implementation details, including prompt designs, are provided in the Appendix §A.1. To ensure reproducibility, we use four open-source LLMs: Meta's LLaMA-8B Instruct, Alibaba's Qwen-7.5B Instruct, Mistral-7B v0.3 Instruct, and DeepSeek R1-8B.

4.1 Datasets

Our evaluation leverages four datasets. CommonsenseQA (Talmor et al., 2019) (CS), a multiple-choice dataset testing general common sense, and StrategyQA (Geva et al., 2021) (ST), which requires strategic reasoning over a knowledge graph, assess agents' commonsense understanding. In contrast, Social-IQa (Sap et al., 2019) (SQA), which focuses on reasoning about social interactions and motivations, and Implicit Hate dataset (ElSherief et al., 2021) (IH), designed to identify subtle forms of hate speech, evaluate agents' social reasoning in nuanced contexts. For brevity, we refer to these datasets using their abbreviations (CS, ST, SQA, IH) in all subsequent tables and figures.

4.2 Team Structure Experiments

We evaluate team structure by comparing bootstrapped accuracy between flat and hierarchical teams across datasets. For this comparison, we use the full test or validation sets of CommonsenseQA, StrategyQA, and Social IQa. For Implicit Hate, we use the stage 1 set of data, which labels each post as 'implicit hate,' 'explicit hate,' or 'non-hate.' We sample 500 from each class to ensure balance and match the overall scale of the other three datasets.

4.3 Team Diversity Experiments

To evaluate the impact of demographic diversity on team performance, we compare persona-based teams, where diversity is introduced through assigned personas, with matched no-persona teams, across both flat and hierarchical structures. In persona-based teams, each agent is assigned a persona along four demographic dimensions (e.g., age, gender, ethnicity, occupation), introducing controlled diversity into the team composition.

For each experimental configuration (model, task, rounds), we match team size and structure between conditions. We then conduct paired statistical tests (paired t-tests and Wilcoxon signed-rank tests) and compute Cohen's d and mean accuracy deltas to assess significance and effect size.

To quantify team diversity, we use Gini's Index (Farris, 2010), which captures variation across demographic dimensions. As exhaustively testing all persona combinations across team sizes is infeasible, we apply stratified sampling to generate teams and select 15 combinations per setting, with an equal number of high, medium, and low diversity teams. Intuitively, high-diversity teams feature agents with maximal differences across the four persona dimensions, while low-diversity teams consist of agents with mostly overlapping demographic traits. This allows us to systematically study the impact of team diversity on task performance. Diversity experiments are run on a 10–20% random subsample of the test or validation sets. Teams without personas are evaluated on the same subsample for consistency. To confirm robustness, we additionally test the best-performing model on the full test or validation datasets. For the Implicit Hate dataset, we sample 100 examples each from the 'implicit hate,' 'explicit hate,' and 'non-hate' categories to maintain class balance.

5 Results

5.1 Effect of Team Structure on Performance

Across all models and tasks, flat teams consistently outperform hierarchical ones, as shown in Table 1.

Table 1: Accuracy comparison of flat/hierarchical teams across models and tasks.

Model	CS	ST	SQA	IH
DeepSeek	66 / 50	61 / 55	49 / 42	38 / 32
LLaMA	79 / 69	67 / 51	54 / 44	44 / 39
Mistral	71 / 64	63 / 57	52 / 45	36 / 41
Qwen	85 / 75	61 / 52	68 / 54	49 / 42

Table 2: Paired t-test comparing flat vs. hierarchical team performance (no diversity condition) across tasks. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Task	t-stat	Mean Diff.	Cohen's d
CS	2.69*	9.54	1.35
ST	5.13*	5.89	2.18
SQA	0.53	0.89	0.26
IH	-0.35	-1.38	-0.18

A paired t-test over all comparisons confirms the significance of this difference ($t = 2.6230$, $p = 0.0192$), with an average performance gain of 5.26 points in favor of flat teams.

To assess whether this structural advantage varies by task, we conduct paired t-tests on each dataset individually (Table 2). Flat teams significantly outperform hierarchical teams on StrategyQA ($t = 4.36^*$, $d = 2.18$) and CommonsenseQA ($t = 2.70^*$, $d = 1.35$). In contrast, the differences for Social IQa and Implicit Hate are small and not statistically significant, suggesting that the effect of team structure may be task-dependent.

These findings indicate that flat teams are especially well-suited for tasks requiring procedural reasoning or multi-step inference, such as strategy problems. In such settings, the peer-to-peer nature of flat communication likely enables more efficient information exchange and decision convergence. Conversely, hierarchical structures may introduce information bottlenecks or distortion as messages propagate across layers, diminishing responsiveness and fidelity, particularly detrimental in tasks where contextual nuance is crucial.

5.2 Effect of Team Diversity on Performance

This section analyzes how task accuracy is affected by demographic diversity, comparing persona-based and no-persona teams, and examining performance variation by Gini-based diversity levels across four tasks.

Across all experimental pairs, we observe a statistically significant performance decline in flat teams when diversity is introduced (t-test = -14.86, Cohen's $d = -0.21$, $p < 0.05$) , with an average drop of 1.35% points. This may stem from increased conflict or misalignment in communication, as agents reason from different demographic perspectives via assigned personas. Hierarchical teams also show a small but significant decline (t = -2.76, Cohen's $d = -0.06$, $p < 0.001$), suggesting that structured communication may limit the effective use of demographic cues. On average, hierarchical teams experience a 0.3% point drop in performance with the addition of personas.

Table 3: Paired t-test values comparing diversity vs. no-diversity (Flat = Flat (3 agents), r. = rounds, Hier. = Hierarchical. All results are significant at $p < 0.05$)

Setting	CS	ST	SQA	IH
Flat, 2 r.	-15.43*	-3.13*	-15.01*	8.58*
Flat, 3 r.	-12.6*	-1.16*	-15.46*	2.01*
Flat, 4 r.	-13.04*	-2.3*	-15.17*	6.96*
Hier., 2 r.	-1.46*	-0.25*	3.5*	-4.3*
Hier., 3 r.	1.25*	-2.9*	4.18*	-3.48*
Hier., 4 r.	-0.78*	-2.81*	3.98*	-7.68*

We further investigate the relationship between diversity and team performance across different team settings, as shown in Table 3. Flat teams consistently show significant performance declines with diversity, with large negative effect sizes (Cohen's d ranging from -0.56 to -0.84) and t-statistics between 13 and 16. In contrast, hierarchical teams exhibit weaker and more inconsistent effects, though the overall trend remains negative.

These results highlight that the impact of demographic personas varies by team structure: flat teams are more sensitive to composition, showing both stronger gains and sharper declines. While diversity often hinders performance, tasks requiring social reasoning and normative understanding may benefit from aligned persona perspectives.

To explore this further, we examine how performance varies with diversity level within a specific task. Figure 2 visualizes team performance on the Implicit Hate task as a function of diversity level, measured by the Gini index. A key pattern emerges: demographic diversity amplifies variance in team performance, with some diverse teams outperforming the baseline and others falling well below it. This variance-amplifying effect echoes findings from human team science (Van Knippenberg et al., 2004), which suggest that diversity tends to increase the spread of outcomes rather than ensuring improvement.

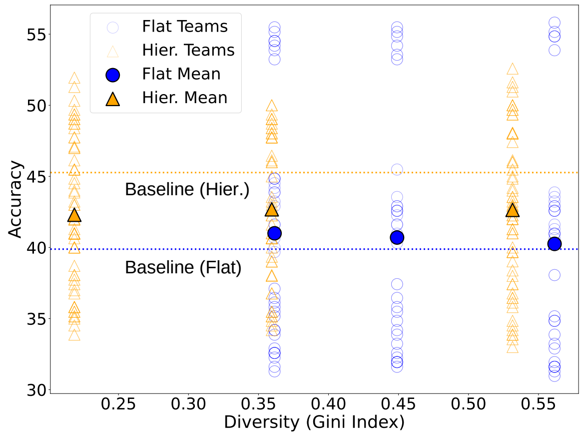

Figure 2: Trend of team diversity and performance in flat teams and hierarchical teams for Implicit Hate.

Flat teams show both larger performance drops and more pronounced outlier gains, while hierarchical teams exhibit weaker and less consistent effects. This suggests that open communication structures may magnify the influence of diversity, depending on how well team members align. Similar trends are observed across other datasets (see Appendix B.4), indicating that diversity's impact is shaped by both team composition and task characteristics.

Task-specific trends further support this interpretation. For example, CommonsenseQA exhibits a modest but steady increase in average accuracy for flat teams as diversity rises, suggesting a consistent benefit from diverse perspectives. In contrast, Implicit Hate task demonstrates increased variance, especially in hierarchical teams, where some configurations excel while others fail to coordinate. These sensitivities highlight the need for further investigation into the interaction between diversity, structure, and task type.

In summary, our results caution against treating diversity as inherently beneficial or harmful. While it can enrich reasoning, its impact depends on team composition and the alignment between personas, task demands, and communication structure. Future work should explore how to select or design persona combinations that are both diverse and cohesive, maximizing the benefits of diversity while mitigating its risks.

5.3 Evaluating Team Comprehension and Coordination

Pre-task expectations As outlined in §3.3, Q_1^{pre} and Q_2^{pre} assess agents' understanding of the shared team goal and their individual roles. We use log odds to compare word usage across groups by computing the logarithm of the odds ratio (Barnard, 2018). It highlights words disproportionately more likely to appear in one group than another, revealing how agents internalize team goals and roles.

Top log-odds words in responses to Q_1^{pre} show that flat teams emphasize efficiency and coordination (e.g.,“wellorganized,” “guide,” “facilitate”), while hierarchical teams highlight structured, task-oriented language (e.g., “brainstorming,” “development,” “provided”). When assessing effect of diversity, we observe subtle shifts: flat teams reference “members” and “finding,” suggesting greater awareness of group dynamics, whereas hierarchical teams remain consistent, continuing to use structural terms like “wellstructured” and “provided.” These lexical patterns reflect how both team structure and demographic framing influence how agents conceptualize their roles and team objectives. A complete list of top log-odds words is provided in Table 12 in the Appendix B.5.2.

Likewise, the top log-odds words for Q_2^{pre} show that flat teams emphasize collective action and coordination (e.g., “facilitate,” “collective,” “wellorganized”), while hierarchical teams reference structured processes and delegation (e.g., “provided,” “decisionmaking,” “wellstructured”). Comparing teams with and without diversity further reveals how social characteristics influence agents’ role perception. Flat teams with diversity mention socially grounded terms like “members,” “finding,” and “methodical,” suggesting role awareness shaped by demographic cues. Meanwhile, hierarchical teams show minimal lexical change, reflecting the dominant role of structural hierarchy. Overall, the word distributions reveal how team structure and diversity framing influence how agents conceptualize their roles. A full list of top log-odds words is available in Table 13 in the Appendix B.5.2.

We analyze average scores for Q_3^{pre} to Q_5^{pre}, which assess agents’ confidence in self, confidence in team, and expected team comprehension, across different team settings. Hierarchical teams report greater confidence in their team’s ability to perform and integrate, whereas flat teams exhibit higher individual confidence. Team diversity has minimal effect on perceived self or team efficacy in flat teams, but it leads to a decline in overall confidence when introduced in hierarchical teams. When examining teams by their level of diversity, we find that highly diverse teams tend to show greater individual con-

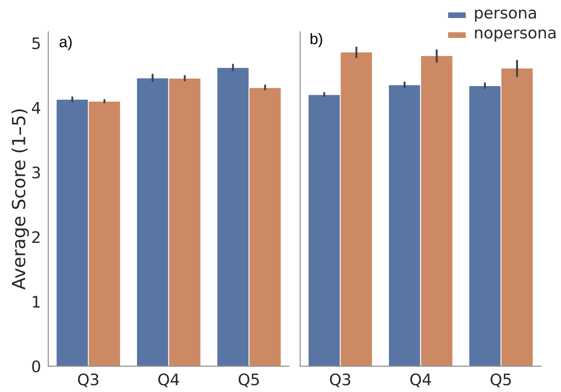

Figure 3: Average score for Q_3^{pre}, Q_4^{pre}, Q_5^{pre}. a) flat structure. b) hierarchical structure.

fidence, while low-diversity teams express more confidence in team comprehension. In hierarchical structures, high diversity improves agents’ confidence in their individual role but slightly reduces perceived team cohesion, although expected team comprehension still increases. These findings suggest that both team structure and diversity shape how agents anticipate their collaborative dynamics before task execution. Additional detailed results are provided in Table 14-17 in the Appendix B.5.3.

Post-task reflections In the post-task interview analysis, we observe distinct patterns across team structures and diversity conditions. Flat teams tend to foster a stronger sense of individual achievement and contribution, whereas hierarchical teams elicit higher ratings for team comprehension and perceived reliance on others. Team diversity leads to a general decline in post-task scores, suggesting that demographic framing may introduce challenges to integration and coordination. While team diversity does not seem to affect perceptions of team dependence or understanding, teams with higher diversity report a greater sense of accomplishment and role-specific satisfaction. These findings point to nuanced effects of structure and diversity on how agents perceive their contributions and collective outcomes after collaboration. Additional detailed results are provided in Table 18-21 in the Appendix B.5.3.

Figure 4 shows the post-task interview scores across all agents in all team settings. Post-task interviews reveal a slight decline in perceived team comprehension. However, agents strongly feel that they cannot perform well-enough without the team as shown in Q_6^{post}. Agents in flat teams generally report more positive collaboration experiences

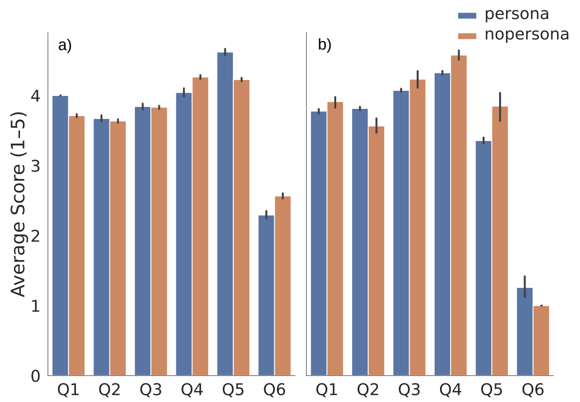

Figure 4: Average score all post-Interview questions. a) flat structure. b) hierarchical structure.

and higher mutual comprehension-reinforcing the idea that open, symmetric communication supports effective reasoning, especially when diversity is present. When comparing teams with and without diversity, we find that hierarchical teams with diversity report greater team comprehension. They also express more positive views of both their contributions and those of their team members, and strongly believe that the team setting is vital to overall performance.

We map each confidence-related pre-task item to its post-task counterpart ($Q_3^{pre} \leftrightarrow Q_2^{post}$, $Q_4^{pre} \leftrightarrow Q_3^{post}$, $Q_5^{pre} \leftrightarrow Q_4^{post}$) to assess change in perceived collaboration over time. Each pair understands the relationship between individual contribution and confidence, team contribution and confidence and team comprehension. We find that in all team settings, we see pre-interview questions had higher scores, indicating that post-task, there is a decrease in the confidence and perception of agents and team performance, team comprehension.

LLM-judged conversation quality GPT-4o evaluation, following the LLM-as-judge approach, suggests that flat teams outperform hierarchical teams across all dimensions. On average, flat teams receive higher scores in *Team Comprehension* (3.91 vs. 3.61), *Collaboration* (4.00 vs. 3.74), *Coherence* (3.83 vs. 3.52), *Reasoning Strength* (3.78 vs. 3.61), *Confidence in Final Answer* (3.70 vs. 3.57), and *Structure Score* (3.87 vs. 3.61). This may be because flat teams exhibit more balanced reasoning and coordination, which align better with GPT-4o's evaluation preferences. Notably, team diversity further boosts GPT-4o evaluations in flat teams across all metrics, while in hierarchical teams, teams with diversity show only marginal improvements—or even slightly lower ratings in some dimensions—indicating that diversity aids collaboration primarily in settings with open, peer-based communication. Additional detailed results are provided in Table 22 in the Appendix B.6.

6 Conclusion

Can lessons from human team science inform the design of multi-agent LLM systems? Our findings suggest that team structure, diversity, and interaction dynamics each play a critical role in shaping team outcomes.

Flat teams consistently outperform hierarchical teams across reasoning tasks, particularly in collaborative or multi-step problems where decentralized communication enables more effective coordination. In contrast, the impact of team diversity is more complex. While diversity often lowers performance, it can improve social reasoning tasks and enhance agents' perceptions of interaction quality. This suggests that diversity plays a positive role in shaping team dynamics, even if it does not always translate into higher accuracy.

Interestingly, agents report high confidence in their team's ability before the task, but post-task reflections reveal difficulties in integrating diverse perspectives. This gap is especially pronounced in hierarchical teams, where constrained communication may limit mutual understanding.

Crucially, our results highlight team diversity as a double-edged sword: it may hinder accuracy, but it also fosters more reflective, calibrated teams that are aware of their limitations. Future work should explore how to better align structural design with diversity-aware coordination strategies to harness the social benefits of diversity without compromising task performance.

Future work should explore adaptive team structures that dynamically adjust roles, delegation, and communication patterns based on task complexity and team composition. Learning-based coordination strategies, such as reinforcement learning or meta-optimization, may help align structure with agent capabilities and diversity profiles. Evaluating these approaches in multilingual, cross-cultural, and real-world tasks would further test the generalizability of our findings. Finally, further research is needed on interpretability and accountability in multi-agent systems, particularly in understanding how teams reason, disagree, and converge on decisions over time.

Limitations

We highlight five key limitations of this study. First, our experiments are constrained to the English-language prompts and may not generalize to multilingual settings. Second, we operationalize diversity using demographic attributes, which serves only as surface-level proxies for deeper experiential and cultural variation. Third, we use relatively small open-source models (7–8B), which may constrain agents' interaction capabilities and reasoning depth. Fourth, our team design does not incorporate dynamic or adaptive strategies, which is a fundamental aspect of interactions. This may have exacerbated communication bottlenecks and misalignments between team members. Lastly, our post-task reflections and confidence measures offer only coarse-grained approximations of agent metacognition. Any interpretations of agent "perception" or "awareness" should thus be made cautiously.

These limitations point to broader challenges in deploying multi-agent systems that simulate human-like teams. Naively implementing diversity or structure without sensitivity to coordination dynamics may result in degraded performance, tokenistic representation, or unintended social consequences. In high-stakes domains (e.g., education, healthcare, policy deliberation), misaligned agent teams may reinforce existing biases or produce misleading outcomes under the appearance of deliberative reasoning.

References

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2025. LLM-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8038–8057, Albuquerque, New Mexico. Association for Computational Linguistics.

G. A. Barnard. 2018. Statistical inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 11(2):115–139. _eprint: https://academic.oup.com/jrsssb/article-pdf/11/2/115/49093577/jrsssb_11_2_115.pdf.

Matteo Bettini, Ryan Kortvelesy, and Amanda Prorok. 2025. The impact of behavioral diversity in multi-agent reinforcement learning.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024a. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024a. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346. Association for Computational Linguistics.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan V. Shah, Junjie Hu, and Timothy T. Rogers. 2024b. Beyond demographics: Aligning role-playing LLM-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14010–14026, Miami, Florida, USA. Association for Computational Linguistics.

Nancy J. Cooke and Margaret L. Hilton. 2015. *Enhancing the Effectiveness of Team Science*. National Academies Press, Washington, DC.

Taylor H. Cox and Stacy Blake. 1991. Managing cultural diversity: Implications for organizational competitiveness. *JSTOR*, 5(3):45–56. Publisher: Academy of Management.

Fatih Delice, Véronique Rousseau, and Juliana Feitosa. 2019. Advancing teams research: What, when, and how to measure team dynamics over time. *Frontiers in Psychology*, 10:1324.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Frank A. Farris. 2010. The gini index and measures of inequality. *The American Mathematical Monthly*, 117(10):pp. 851–864.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Lindred Greer, Bart de Jong, Maartje Schouten, and Jennifer Dannals. 2018. Why and when hierarchy impacts team effectiveness: A meta-analytic integration. *Journal of Applied Psychology*, 103:591–613.

J. Richard Hackman. 2002. *Leading Teams: Setting the Stage for Great Performances*. Harvard Business Press, Brighton, MA.

Keisuke Hattori and Mai Yamada. 2023. Closing the psychological distance: The effect of social interactions on team performance. *Unpublished manuscript or preprint*. Retrieved from title metadata; please update journal and DOI if available.

Mahmood Hegazy. 2024. Diversity of thought elicits stronger reasoning capabilities in multi-agent debate frameworks. *International Journal of Computer Science and Mobile Applications*, 12(10):16–29.

Sujin K. Horwitz and Irwin B. Horwitz. 2007. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, 33(6):987–1015.

Hao Ji and Jin Yan. 2020. How team structure can enhance performance: Team longevity's moderating effect and team coordination's mediating effect. *Frontiers in Psychology*, 11:1873.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627. Association for Computational Linguistics.

Aparna Joshi and Hyojin Roh. 2009. The role of context in work team diversity research: A meta-analytic review. *Academy of Management Journal*, 52(3):599–627.

Florian Kunze and Katharina Hampel. 2022. Age at the team and organizational levels. In Hannes Zacher and Christian W. Rudolph, editors, *Age and Work: Advances in Theory, Methods, and Practice*, pages 152–168. Routledge.

Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. 2021. Celebrating diversity in shared multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More agents is all you need.

Yuan Li, Yixuan Zhang, and Lichao Sun. MetaAgents: Simulating interactions of human behaviors for LLM-based task-oriented coordination via collaborative generative agents.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate.

Melanie J. McGrath, Andreas Duenser, Justine Lacey, and Cecile Paris. 2024. Collaborative human-ai trust (chai-t): A process framework for active management of trust in human-ai collaboration.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. ACM.

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

Lisa Hope Pelled, Kathleen M. Eisenhardt, and Katherine R. Xin. 1999. Exploring the black box: An analysis of work group diversity, conflict and performance. *Administrative Science Quarterly*, 44(1):1–28.

Marie-Élène Roberge and Rolf van Dick. 2010. Recognizing the benefits of diversity: When and how does diversity increase group performance? *Human Resource Management Review*, 20(4):295–308.

Gaurav Sahu, Robin Cohen, and Olga Vechtomova. 2021. Towards a multi-agent system for online hate speech detection.

Eduardo Salas, Nancy J. Cooke, and Michael A. Rosen. 2008. On teams, teamwork, and team performance: Discoveries and developments. *Human Factors*, 50(3):540–547. PMID: 18689065.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. PersonaGym: Evaluating persona agents and LLMs.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Ji Song and Lei Li. 2020. Comparing race, gender, age, and career categories in recognizing and grouping tasks. *PeerJ*, 8:e9156.

Günter K. Stahl and Martha L. Maznevski. 2021. Unraveling the effects of cultural diversity in teams: A retrospective of research on multicultural work groups and an agenda for future research. *Journal of International Business Studies*, 52(1):4–22.

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.

Daan Van Knippenberg, Carsten KW De Dreu, and Astrid C Homan. 2004. Work group diversity and group performance: an integrative model and research agenda. *Journal of applied psychology*, 89(6):1008.

Daan van Knippenberg, Lisa H Nishii, and David J G Dwertmann. 2020. Synergy from diversity: Managing team diversity to enhance performance. *behavioral science*, 6(1).

Hecong Wang, Erqian Xu, Pinxin Liu, Zijian Meng, and Zhen Bai. 2023. Contextualized multi-step commonsense reasoning through context extension. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. 2025. Talk structurally, act hierarchically: A collaborative framework for LLM multi-agent systems.

Zengqing Wu and Takayuki Ito. 2025. The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems.

Fengli Xu, Lingfei Wu, and James Evans. 2022. Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences*, 119(23):e2200927119.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023. Towards reasoning in large language models via multi-agent peer review collaboration.

Yingxuan Yang, Qiuying Peng, Jun Wang, Ying Wen, and Weinan Zhang. 2024. Llm-based multi-agent systems: Techniques and business perspectives.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024a. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2024b. Cumulative reasoning with large language models.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025. MultiAgentBench: Evaluating the collaboration and competition of LLM agents.

A Appendix

A.1 Implementation Details

We provide further implementation details of our study in this section. The temperature for all models—Llama-8b-Instruct, Qwen, Mistral, and Deepseek—are set to 0.7 as though are the defaults. We provide the algorithms of our workflows in Algorithm 1 and Algorithm 2. For the evaluation of conversations by GPT-4o, we provided a temperature of 0.7. The final output for both workflow is one of the multiple options provided for the question. For testing our framework, we use A800 GPUs with 80GB.

Algorithm 1: FLATTEAMDEBATE: Multi-Round Discuss-and-Vote Framework

Input: Question Q; maximum rounds R; agents $\mathcal{A} = \{A_i\}_{i=1}^{n}$ (odd n);
Output: Team answer $\hat{a}$;

$r \leftarrow 0$;
while $r \leq R$ ***and*** CONSENSUS($\{a_i^{(r-1)}\}$) = `false` **do**
  foreach *agent* $A_i \in \mathcal{A}$ **do**
    if $r = 0$ **then**
      $P \leftarrow Q$;
    else
      $P \leftarrow (Q,\ \{a_j^{(r-1)}\}_{j=1}^{n})$;
    $(a_i^{(r)}, e_i^{(r)}, p_i^{(r)}) \leftarrow A_i(P)$;
  $r \leftarrow r + 1$;
$\hat{a} \leftarrow$ MAJORITYVOTE($\{a_i^{(r-1)}\}$);
return $\hat{a}$;

We use the following prompts for Flat Team. For tests with diversity, we provide the demographic diversity of agents in each round.

Prompt Design for Flat Team Structure

Round 0 (Initial Answer)
You are a reasoning agent `agent_id`*. You are here to answer multiple choice reasoning questions.*
Please answer the following question by selecting only one option.
Question: `q`
Answer: ___

Rounds 1 to N (Refinement Phase)
You are a reasoning agent `agent_id`*. You are here to answer multiple choice reasoning questions.*
You are part of a team of agents. You are expected to help your team get to the correct answer.
You will be given the question and your previous response and your team members' previous responses.
Here are your previous answers from your team: `context`
Take a moment to reflect on the responses and then engage in conversation to come to the right answer.
Question: `q`
Answer: ___

Final Round (Consensus Prompt)
Agents, review the conversation: `final_context`*.*
Come to a consensus on the best final answer for the question:
Question: `q`
Answer: ___

Algorithm 2: HIERTEAM: Leader–Subordinate Delegation Framework

Input: Question Q; maximum rounds R; leader L; subordinates $\mathcal{S} = \{S_k\}_{k=1}^{m}$;
Output: Final team answer $\hat{a}$;

$r \leftarrow 0$;
while $r \leq R$ **do**
  if $r = 0$ **then**
    $P_L \leftarrow Q$;
  else
    $P_L \leftarrow (Q,\ \{a_k^{(r-1)}\}_{k=1}^{m})$;
  $\{I_k^{(r)}\}_{k=1}^{m} \leftarrow L(P_L)$;   `// Leader generates instructions`
  foreach $S_k \in \mathcal{S}$ **do**
    $(a_k^{(r)}, e_k^{(r)}) \leftarrow S_k(I_k^{(r)})$;
  if $r = R$ **then**
    $\hat{a} \leftarrow L(\{a_k^{(r)}\}_{k=1}^{m})$;
    return $\hat{a}$;
  $r \leftarrow r + 1$;

Similarly, based on Algorithm 2, we created the following prompt flow for testing hierarhical teams.

Prompt Design for Hierarchical Team Structure

Round 0 – Leader's Initial Instruction Prompt
You are the team leader of a reasoning team. The goal of the team is to answer reasoning questions as accurately as possible. You manage the following agents: `team_description`*.*
Your role is to delegate tasks to your team members so that they can provide you with useful information.
Create clear instructions for each agent.
Question: `q`
Team Members: `team_members`
Instructions:
Agent 1: ___
Agent 2: ___
Agent 3: ___

Round 1 – Team Member Reasoning Prompt
You are a team member of a reasoning team. `persona` *You are led by team leader Agent 1. Your role is to answer based on the leader's instruction to help solve the reasoning question.*
Question: `q`
Instruction: `instr`
Answer: ___

Round 1 to N – Leader's Refinement Instruction Prompt
You are the team leader of a reasoning team. You manage the following agents: `team_description`*.*

Your team members have submitted initial answers to the question.
Review their responses and provide each member with updated instructions to reaffirm or correct their reasoning.
Your instructions must be clear and under 10 words.
Team Members' Answers: `team_answers`
Question: `q`
Instructions:
Agent 1: ___
Agent 2: ___
Agent 3: ___

Final Round – Leader Final Reflection Prompt
You are the team leader of a reasoning team. Your team members have responded based on your updated instructions.
Reflect on their responses and provide the final correct answer. Your answer may differ from your team members'.
Team Members' Final Answers: `team_answers`
Question: `q`
Final Answer: ___

A.2 Licenses

We document the language, domain, and demographic characteristics of the datasets and models used in this study following best practices from data and model documentation toolkits (e.g., Data Statements, Model Cards, Datasheets for Datasets). All datasets—CommonsenseQA, Social IQa, StrategyQA, and Implicit Hate Detection—are in English and focus on reasoning tasks across different domains. CommonsenseQA and StrategyQA cover general knowledge and commonsense reasoning, while Social IQa focuses on social and situational commonsense, and Implicit Hate captures nuanced, often sarcastic, toxic speech from Reddit. Demographic information about dataset authors is generally unavailable, with the exception of Social IQa and Implicit Hate, which were crowd-annotated by U.S.-based workers with diverse backgrounds. The models we use include Meta's LLaMA-8B Instruct, Alibaba's Qwen-7.5B Instruct, Mistral-7B v0.3 Instruct, and DeepSeek R1 (a distilled LLaMA variant). All models are primarily trained on English, with partial multilingual capabilities in Qwen and Mistral. Their training data spans web text, code, and instruction-tuned corpora, although precise data composition is not fully disclosed for all models. None of the models guarantee demographic balancing or fairness-aware pretraining. Model licenses range from fully permissive (Apache 2.0, MIT) to research-constrained (LLaMA 2 Community License).

Table 4: Licenses for datasets used in this study.

Dataset	License
CommonsenseQA	CC BY-SA 4.0
Social IQa	MIT License
StrategyQA	Apache 2.0
Implicit Hate	MIT License

A.3 Diversity Settings

Demographic diversity of an agent is constructed using four dimensions of human demographics: age, gender, ethnicity, and occupation. Each dimension includes a range of categories——gender (male, female), age (young, young working professional, working professional, senior citizen), ethnicity (White, Asian, Black), and occupation (white-collar, blue-collar). By systematically combining these categories, we generate a total of 48 unique personas used to populate our teams. Here is an example of the persona provided to an agent:

> *You are male and of age 18 to 24. You identify as white and work a blue collar job.*

B Additional Results

B.1 Single Agent

To understand the effect of persona, we conducted an ablation study using single agents. We test the effect of the four dimensions of diversity as mentioned in Section 3.2 on the CS and SQA datasets. We examine the effect of each dimension and further the combination of these dimensions. Table 6 shows the average performance delta of diversity-based agent compared to their no-diversity counterparts, grouped by the number of demographic dimensions used. When only one or two dimensions are included (e.g., just gender or age), teams see modest performance improvements, with an average delta of 1.83 and 0.75, respectively. However, as the number of persona dimensions increases to three or four, performance begins to decline. Teams with four-dimensional personas (gender, age, ethnicity, occupation) show a negative average delta of -0.91, suggesting potential cognitive overload or misalignment introduced by more complex social cues. These results indicate that while lightweight demographic cues may support collaboration, higher-dimensional personas may hinder team effectiveness, possibly due to increased

Table 5: Licenses for models used in this study.

Model	Organization	License
LLaMA-8B Instruct	Meta	LLaMA 2 Community License
Qwen-2.5-7B Instruct	Alibaba	Apache 2.0
Mistral-7B Instruct v0.3	Mistral	Apache 2.0
DeepSeek R1 (LLaMA-8B Distil)	DeepSeek	MIT License

coordination demands or difficulty in integrating diverse perspectives.

Table 6: Effect of Persona Dimensionality on Single-Agent Accuracy (Delta from No-Persona Baseline)

# Dimensions	Avg Accuracy Delta	Std Dev
1	+1.83	2.05
2	+0.75	3.53
3	−1.62	8.70
4	−0.91	9.71

B.2 Comparing different sizes and rounds

As mentioned in Section 3.1, we create teams of varying size. In flat teams, we test across teams size of 3, 5 and 7 for 2 to 4 rounds. In Table 7, we report the average bootstrapped accuracy of all flat team settings across the four datasets. We find that the number of rounds and number of agents have very small effect on the performance of the team. These findings indicate that scaling in teams is a nuanced issue that future work can address in the context of team science for AI teams

Table 7: Average accuracy (%) across tasks by team setting (agents × rounds).

Team Setting	CS	IH	SQA	ST
3 agents, 2 rounds	71.29	43.43	54.41	63.97
3 agents, 3 rounds	71.98	38.22	55.11	63.49
3 agents, 4 rounds	71.94	41.47	54.65	63.68
5 agents, 2 rounds	72.30	38.63	54.08	64.51
5 agents, 3 rounds	73.45	37.56	54.25	64.91
5 agents, 4 rounds	72.30	37.69	54.16	64.34
7 agents, 2 rounds	67.98	42.06	55.17	64.84
7 agents, 3 rounds	69.28	42.16	54.52	65.00
7 agents, 4 rounds	68.26	42.52	54.89	64.33

As mentioned in Section 3.1, we create two hierarchical teams, one with 1 leader and 3 team members and second with 1 leader, 2 managers and 4 team members. The team settings emulate 1 level and 2 levels of hierarchy respectively. In the paired t-test comparing the performance of these two settings, we find that 1 level of hierarchy is consistently preferred, potentially indicating that more rigid structures of communications are not beneficial.

Table 8: Paired t-test comparing hierarchical level 1 vs. hierarchical level 2 team performance across tasks.

Task	t-stat	Mean Difference	Cohen's d
CS	1.978	21.561	0.989
IH	2.873*	7.713	1.437
SQA	5.280*	19.798	2.640
ST	3.125*	21.917	1.563

Table 9 reports t-test statistics comparing team performance under diversity versus no-diversity conditions across a variety of team configurations. Results indicate that diversity often has a statistically significant effect, but the direction and magnitude vary by task and team size. For example, in larger teams (5–7 agents), diversity has more pronounced positive effects on ST, particularly as team size and number of rounds increase. Notably, the negative impact of diversity on CS is consistent across all team sizes, suggesting that uniformity in perspective may benefit certain types of commonsense reasoning. Overall, the data underscores the nuanced and configuration-dependent impact of diversity on team reasoning dynamics.

Table 9: Paired t-test statistics for diversity vs. no-diversity comparisons across team settings. Asterisk (*) indicates $p < 0.05$.

Team Setting	CS	IH	SQA	ST
3 agents, 2 rounds	-9.63*	3.01*	-2.42*	0.91
3 agents, 3 rounds	-7.30*	0.25	0.30	-5.54*
3 agents, 4 rounds	-7.12*	7.15*	-1.89	-7.77*
5 agents, 2 rounds	-9.68*	-2.42*	-3.71*	9.02*
5 agents, 3 rounds	-10.15*	-2.56*	-4.38*	5.79*
5 agents, 4 rounds	-10.28*	-2.50*	-4.33*	10.39*
7 agents, 2 rounds	-6.04*	2.60*	-1.15	9.12*
7 agents, 3 rounds	-5.64*	2.76*	-3.12*	6.07*
7 agents, 4 rounds	-5.68*	1.38	-3.20*	8.69*

Table 10 presents paired t-test statistics comparing diversity-based teams to no-diversity teams across different hierarchical structures and reasoning rounds. The results show that diversity significantly enhances performance across all tasks in the first-level hierarchy (Hier. l1), with extremely high t-values and $p < .001$ for SQA and ST. However, the effects diminish or even reverse in the second-level hierarchy (Hier. l2), where deeper delegation and communication layers appear to hinder the effective use of persona information. Specifically, SQA shows a dramatic drop from strong positive gains in Hier. l1 to significant negative effects in Hier. l2, suggesting that tasks requiring nuanced social inference are particularly sensitive to how persona information is coordinated across levels. This pattern highlights that persona benefits are maximized in shallow hierarchies where agents can directly interpret and leverage identity cues.

Table 10: T-test statistics (t-value) for persona vs. no-persona comparisons across tasks and team settings (Team). l1 refers to hierarchical teams with 1 leader and 3 team members, l2 refers to hierarchical teams with 1 leader, 2 managers and 4 team members. Significance is denoted as: * $p < .05$, ** $p < .01$, *** $p < .001$

Team	CS	IH	SQA	ST
l1, 2 r.	7.49***	7.09***	14.46***	17.38***
l1, 3 r.	8.87***	7.31***	13.96***	19.57***
l1, 4 r.	8.07***	4.61***	13.21***	17.90***
l2, 2 r.	3.32**	0.20	-6.25***	4.18***
l2, 3 r.	2.12*	-0.37	-6.43***	3.23**
l2, 4 r.	1.22	-0.30	-5.49***	1.68

B.3 Outperforming diversity teams

Table 11 presents a comprehensive list of diversity-based team experiments that outperformed their no-diversity counterparts across various tasks, team sizes, and team structures. The results highlight the consistent advantage of persona use in flat teams, particularly those composed of 3 agents. In this configuration, CS and SQA dataset show strong gains, with experiments such as 04, 06, 07, 09, 11, and 14 repeatedly emerging across tasks and rounds. This suggests that certain team compositions—defined by the demographic personas assigned—are especially synergistic under conditions of open interaction and shared responsibility.

As team size increases to 5 or 7 agents, persona benefits remain evident, though the pattern becomes more task-specific. For instance, ST shows robust gains in larger flat teams, whereas gains for IH are sparse across all configurations. Hierarchical teams show more mixed results. While several experiments still outperform the baseline, especially in the first-level hierarchy (l1), the benefits of persona cues appear attenuated. In second-level hierarchies (l2), where communication is more constrained, persona advantage persists in SQA and ST but becomes less discriminative across configurations, likely due to reduced opportunities for mutual interpretation and integration.

These findings reinforce our broader claim that diversity does not uniformly improve performance, but rather is dependent on team structure and task.

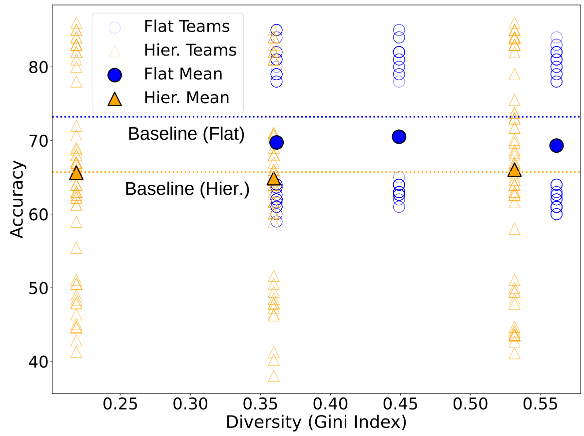

Figure 5: Trend of team diversity and performance in flat teams and hierarchical teams for CS dataset. x-axis represents the level of team diversity, calculated through Gini Index, and y-axis represents performance of teams.

B.4 Team Diversity

Figures 5, 6, 7 showcase the relationship between diversity, measured through Gini Index, against performance. Each figure shows the trend of flat and hierarchical teams with increasing diversity. We observe the trend the teams is dependent on the task. However, all three plots show high variance across the diversity teams, reaffirming that diversity does not have a universal effect on performance of teams.

B.5 Team Interview

B.5.1 Wordclouds

Figure 8 and Figure 9 presents word clouds generated from agent responses to two pre-task interview questions: Q_1^{pre} ("What is the primary goal of the team?") and Q_2^{pre}("What is your role in the team?"). These visualizations highlight the most frequently used words across different team configurations, including flat versus hierarchical structures and with versus without diversity. By comparing word usage across team settings, we can observe how team framing and diversity assignment influence how agents conceptualize collective goals and individ-

Table 11: Experiments where diversity-based teams outperformed their no-diversity counterparts.

Team	CS	IH	SQA	ST
Flat (3 agents)	[01, 04, 05, 06, 07, 09, 11, 14]	[01, 07]	[01, 04, 06, 07, 10, 11, 12, 14, 15]	[]
Flat (5 agents)	[01, 03, 06, 08, 09, 10]	[]	[10, 11, 12]	[01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13]
Flat (7 agents)	[01, 03, 06, 08, 09, 10]	[]	[10, 11, 12]	[01, 02, 03, 04, 05, 06, 07, 09, 10, 12, 13]
Hier. I1	[02, 03, 05, 06, 08, 10, 12, 13, 14]	[03, 06, 14]	[01, 02, 03, 04, 05, 06, 08, 09, 10, 12, 14]	[02, 03, 05, 08, 12, 13, 14]
Hier. I2	[01, 03, 04, 09, 10, 11, 13]	[01, 03, 05, 06, 08, 10, 11, 13]	[01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14]	[01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14]

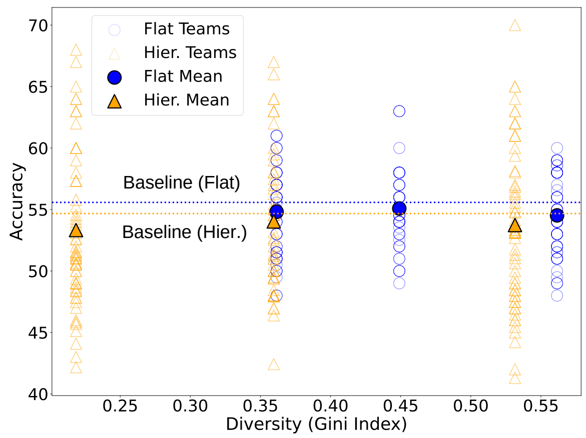

Figure 6: Trend of team diversity and performance in flat teams and hierarchical teams for SQA dataset. x-axis represents the level of team diversity, calculated through Gini Index, and y-axis represents performance of teams.

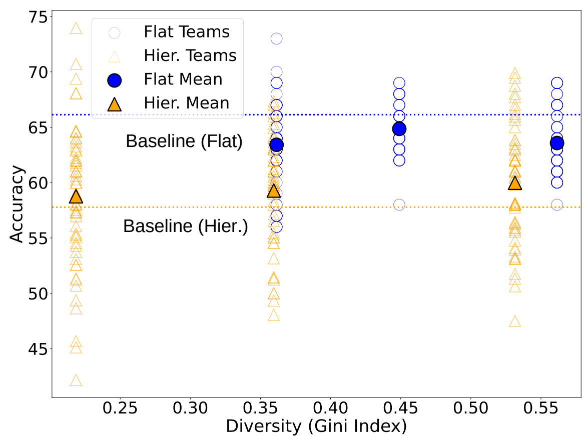

Figure 7: Trend of team diversity and performance in flat teams and hierarchical teams for ST dataset. x-axis represents the level of team diversity, calculated through Gini Index, and y-axis represents performance of teams.

ual roles. For example, both teams emphasize on "solve", "effectively" and "efficiently" indicating an understanding of the shared goal of the team.

In Figure 9, both flat teams and hierarchical teams indicate that agents are aware that they are team members and are required to contribute towards the shared goal. However, upon further investigation, we find that flat team tend to use words such as "support" and "provide" more than hierarchical teams which uses "assist", "expertise" and "clear communication". This can indicate that flat teams are more geared towards a collaborative position. Meanwhile, hierarchical teams tend to adopt a more structure approach to solving the given problem.

B.5.2 Log-odds analysis

In addition to word clouds, we provide, the log-odds the answers of Q_1^{pre} and Q_2^{pre}.

Table 12: Top log-odds words by team structure and diversity for Q_1^{pre}. (Hier. = Hierarchical, N = No Diversity, D = Diversity.)

Team	Setting	Top Words
Flat		wellorganized, guide, concise, facilitate, optimal
Hier.		right, wellstructured, brainstorming, development, provided
Flat (N)		wellorganized, guide, arrive, communication, answers
Flat (D)		seasoned, methodical, members, finding, particularly
Hier. (N)		facilitating, 2, success, assistant, contributes
Hier. (D)		right, wellstructured, brainstorming, development, provided

B.5.3 Detailed Analysis

Figures 10 and 11 present average scores from the pre- and post-task interviews, providing a quanti-

Figure 8: Panel of wordclouds showing response of agents in flat and hierarchical settings answering Q_1^{pre}: "What is the primary goal of the team?" Plots a,b,c show results for Flat teams (overall, only teams with no persona, only teams with persona respectively). Plots d,e,f show Hierarchical teams overall, teams with no persona, teams with only persona respectively.

tative snapshot of how agents perceive their roles, capabilities, and team dynamics.

As observed in Figure 10, agents tend to be extremely confident in their ability to contribute and integrate well with the team.

Further statistical tests on pre-interview results shows that flat teams are more confident about their ability to execute the task, while hierarchical teams are more confident about the team's ability to execute and integrate with diverse perspectives.

Table 15 examines the effect of diversity in team structures and confidence in agent ability, team ability and integration efforts. We observe that across the team structures, teams with no diversity tend to have more confidence in themselves, the team and ability to integrate.

While we found that on performance, the level of diversity has insignificant effects, we find that in the pre-task interviews, the level of diversity effects the confidence in the team's ability to integrate with diverse perspectives. Particularly, as shown in Table 17, in flat teams more diversity builds confidence to integrate with diverse perspective, however in hierarchical teams there is an opposite effect.

Figure 11 shows that post interaction, agents confidence in their own and the team's contribution reduces. However, team members feel strongly about maintain the team in order to complete the tasks, particularly in hierarchical teams which could indicate a reliance on the top-down structure.

In Section 3.3, we enumerate the six questions that we ask each agent team after the completion of the task. In Table 18, we observe that flat teams are

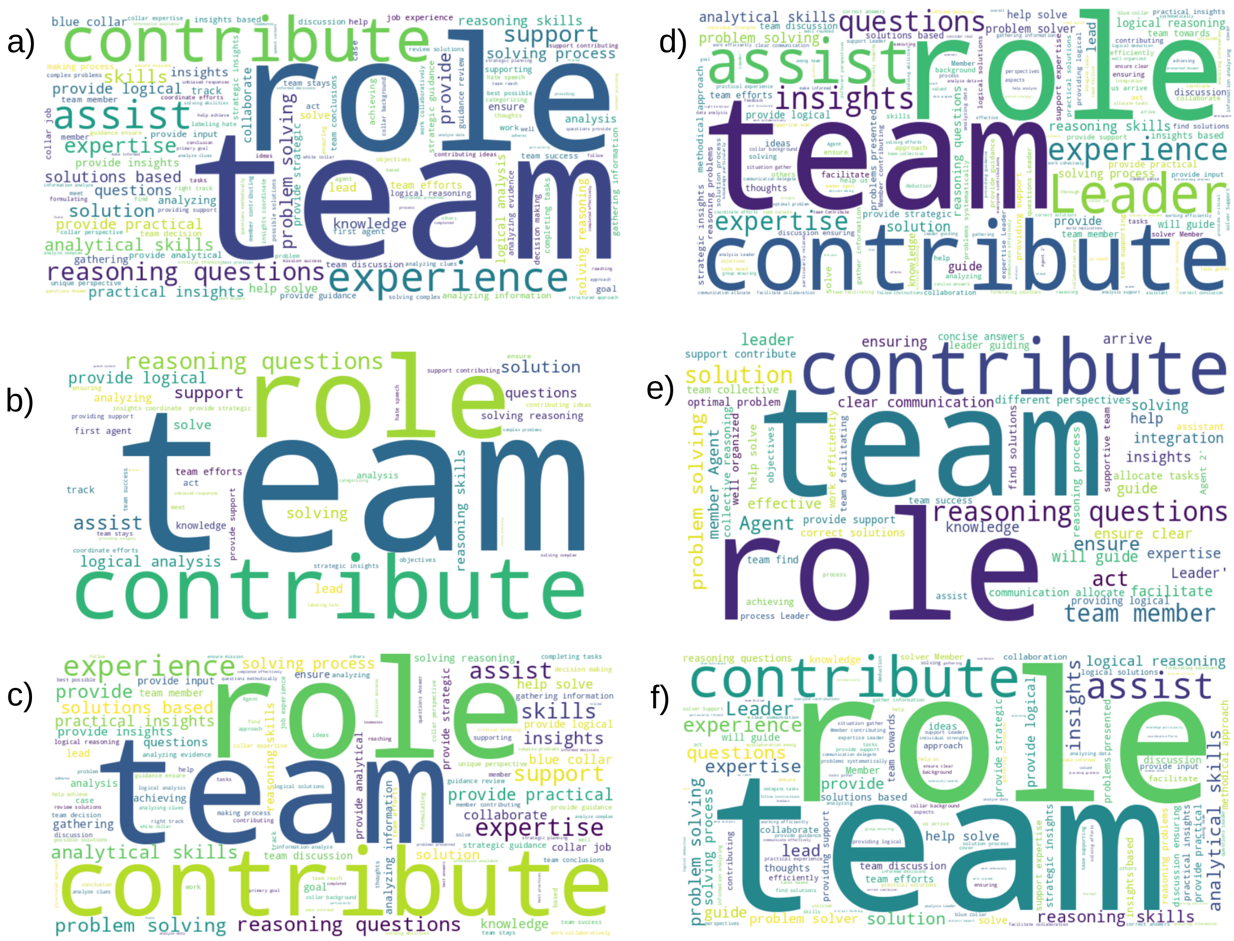

Figure 9: Panel of wordclouds showing response of agents in flat and hierarchical settings answering Q_2^{pre}: "What is your role in the team?". Plots a,b,c show results for Flat teams answering Q_2^{pre} (overall, only teams with no persona, only teams with persona respectively). Plots d,e,f show Hierarchial teams overall, teams with no persona, teams with only persona respectively answering Q2.

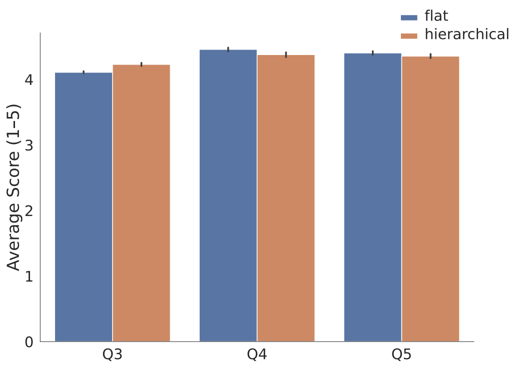

Figure 10: Plot shows average score for Q_3^{pre}, Q_3^{pre}, Q_5^{pre} in pre-Interview setting. Blue bars present flat teams and Orange bars represent hierarchical teams.

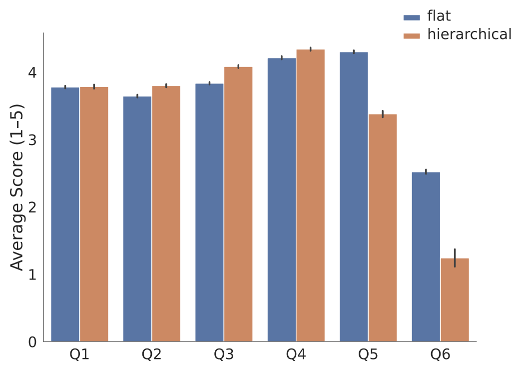

Figure 11: Plot shows average score all post-Interview questions. Blue bars present flat teams and Orange bars represent hierarchical teams.

Table 13: Top log-odds words by team structure and diversity for Q_2^{pre}

Team Setting	Top Words
Flat	perspectives, optimal, clear, facilitate, collective, different, wellorganized
Hier.	provided, decisionmaking, consensus, right, brainstorming, collaboratively, wellstructured
Flat (N)	optimal, correct, clear, allocate, facilitate, effective, collective
Flat (D)	finding, seasoned, methodical, members, related, field, particularly
Hierarchical (N)	success, 1, achieving, 2, facilitating, objectives, supportive
Hierarchical (D)	provided, decisionmaking, consensus, right, brainstorming, collaboratively, wellstructured

Table 14: Paired t-test results comparing flat vs. hierarchical structures on pre-task scores. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Question	t-stat
Q3	-7.598***
Q4	3.214**
Q5	2.144*

more confident about the contributes made to the team and that of the team. However, hierarchical teams have a strong and significant confidence in team comprehension.

Further, we observe that the diversity teams and no diversity teams in these structures also have varying responses to the post-interview questions. The results indicate that in both team structures, the inclusion of diversity improves perception of team contributions and team comprehension, but reduces willingness to work with the same team again as shown by statistical results of Q_6^{post} in Table 19.

We also observe the effect of level of diversity of the two team structures for the post-task interview questions and find that level of diversity has a significant effect on the perception of contributions and comprehension, but not on the confidence in continuing with the team.

B.6 GPT-4o Evaluation

Section 3.3 details how GPT-4o was used to evaluate a sample of conversations. We observe the statistical difference in the scores by team structure and team diversity.

Table 22 compares across the questions, flat and hierarchical teams in their diversity and no diversity settings. We find that flat teams with diversity have the highest score across all dimensions (Team Comprehension, Collaboration, Coherence, Reasoning Strength and Structure Score. Hierarchy with no diversity has the lowest scores. Table 23, 24 show that for flat teams high diversity improves scores across all dimensions. Meanwhile for GPT-4o, lower diversity tends to improve scores.

Table 15: Paired t-test results comparing diversity vs. no-diversity responses by team structure and question for pre-task questions. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Structure	Question	t-stat (Significance)
Flat	Q_3^{pre}	-1.431
Flat	Q_4^{pre}	-0.120
Flat	Q_5^{pre}	-10.829***
Hierarchical	Q_3^{pre}	-15.285***
Hierarchical	Q_4^{pre}	-8.824***
Hierarchical	Q_5^{pre}	-4.208***

Table 16: Kruskal-Wallis test results comparing mean scores of pre-task questions across diversity levels (low, medium, high) for each structure and question. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Structure	Question	Kruskal-Wallis H
Flat	Q_3^{pre}	8.243*
Flat	Q_4^{pre}	0.340
Flat	Q_5^{pre}	7.778*
Hierarchical	Q_3^{pre}	5.975*
Hierarchical	Q_4^{pre}	5.554*
Hierarchical	Q_5^{pre}	21.910***

C Conversation Samples

C.1 Team conversations

To illustrate the reasoning processes and coordination strategies used by AI agents, we present example conversations from both flat and hierarchical teams. These conversations span multiple rounds of deliberation, showing how agents build on each other's responses, update beliefs, and (in the hierarchical setting) respond to top-down instructions.

Each example includes the original question, agent responses per round, and the final team prediction. We include these transcripts to help readers understand how different team structures impact the interaction flow, convergence dynamics, and

Table 17: Paired t-test results comparing high vs. low diversity teams for each structure and question. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Structure	**Question**	**t-stat**
Flat	Q_3^{pre}	2.673**
Flat	Q_4^{pre}	1.031
Flat	Q_5^{pre}	-2.816**
Hierarchical	Q_3^{pre}	2.129*
Hierarchical	Q_4^{pre}	-2.093*
Hierarchical	Q_5^{pre}	3.672***

Table 18: T-test comparison between flat and hierarchical teams for each post-interview question. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Question	**t-statistic**
Q_1^{post}	-0.387
Q_2^{post}	-9.150***
Q_3^{post}	-15.886***
Q_4^{post}	-6.955***
Q_5^{post}	35.531***
Q_6^{post}	18.986***

collaborative reasoning quality.

C.2 Team Interviews

To assess agents' meta-awareness of their roles and team processes, we conduct pre- and post-task interviews. In the pre-task phase, agents are asked to articulate the team's goal, their own role, and their expected confidence in completing the task. In the post-task phase, they reflect on their team's performance, their own contribution, and their ability to understand (and be understood by) teammates.

These interviews help us evaluate perceived coordination and alignment, and serve as a self-reflective complement to our quantitative metrics and GPT-4o evaluations. Below, we include representative examples of these interview responses across different team types.

Before and after each task, agents are prompted to reflect on their goals, roles, and contributions through structured "interview-style" questions. These responses help assess team awareness, perceived collaboration quality, and confidence in execution.

Figure 14 presents the intro and outro responses from a flat team. Agents express high confidence in their roles and strong satisfaction with team performance.

Table 19: Paired t-test results comparing diversity vs. no-diversity scores for all post-task interview questions, split by team structure. Delta indicates diversity minus no-diversity mean score. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Structure	**Question**	**t-statistic**	**Delta**
Flat	Q_1^{post}	-23.568***	-0.288
Flat	Q_2^{post}	-1.331	-0.035
Flat	Q_3^{post}	-0.411	-0.010
Flat	Q_4^{post}	6.584***	0.220
Flat	Q_5^{post}	-14.762***	-0.393
Flat	Q_6^{post}	8.107***	0.272
Hierarchical	Q_1^{post}	-2.965**	-0.136
Hierarchical	Q_2^{post}	4.670***	0.250
Hierarchical	Q_3^{post}	-2.492*	-0.160
Hierarchical	Q_4^{post}	-6.428***	-0.253
Hierarchical	Q_5^{post}	-4.545***	-0.492
Hierarchical	Q_6^{post}	3.719***	0.259

Figure 15 shows the same for a hierarchical team. The leader explicitly articulates guidance responsibility, while subordinates demonstrate a range of self-rated confidence. Post-task reflections highlight minor breakdowns in communication, even as task performance remains strong.

C.3 GPT-4o evaluation

To complement task accuracy and agent self-assessments, we employ GPT-4o as an external evaluator. For each sampled conversation, GPT-4o is shown the full multi-round transcript and prompted to rate the team across six dimensions including Team Comprehension, Collaboration, Reasoning, Coherence, Confidence in Answer and Effectiveness of Team Structure.

GPT-4o also provides explanations that illuminate key differences in team dynamics across conversational structures. In flat teams, GPT-4o often highlights superior coordination and mutual engagement. For example, it notes *"all three agents picked up on the same underlying implication, demonstrating a coordinated reasoning process without direct prompting"* and *"the agents built on each other's insights, showing smooth collaborative understanding throughout the discussion"*. Such explanations emphasize how flat configurations enable aligned reasoning and effective knowledge sharing. In contrast, hierarchical teams are frequently charac-

Flat-Team Conversation Example

Question: A revolving door is convenient for two-direction travel, but it also serves as a security measure at a what?
Options: A. bank B. library C. department store D. mall E. new york
Gold Answer: A. bank

Round 0 (Initial)
Agent 0: A
Explanation: Revolving doors are often used in places where security is a concern, such as banks. They help prevent unauthorized access and provide a secure entry point. Among the options, banks clearly prioritize security.

Agent 1: A
Explanation: Banks are known for using revolving doors as a security feature. While other options might use them for convenience, only banks use them explicitly for access control.

Agent 2: A
Explanation: Revolving doors help protect secure environments. Banks use them to regulate access and avoid tailgating. Other places like malls or libraries don't require that level of control.

Majority Vote (Round 0): A

Round 1
Agent 0: A **Agent 1:** A **Agent 2:** A
Majority Vote (Round 1): A

Round 2
Agent 0: A **Agent 1:** A **Agent 2:** A
Majority Vote (Round 2): A

Final Answer: A
Explanation: Revolving doors are typically found in buildings where security is important. Banks, for example, use them to regulate entry and prevent unauthorized access, serving both a functional and protective role.

Figure 12: Example of a flat multi-agent debate over multiple rounds. All agents independently select the correct answer (A. bank) and reinforce their reasoning across rounds.

terized by GPT-4o as exhibiting fragmentation and limited synthesis. Its judgments include observations such as *"the hierarchy caused the top agent to override others, leading to a lack of consensus or integrated reasoning"* and *"agents failed to fully engage with each other's inputs, resulting in fragmented and disconnected responses"*. These statements reflect a diagnostic pattern where unequal conversational authority hinders collaborative depth. Furthermore, GPT-4o attributes increased comprehension and reasoning quality to diversity in team composition. In experiments designed to include varied agent backgrounds, GPT-4o explains *"their varied perspectives contributed to a more comprehensive understanding of the nuanced scenario"* and *"inclusion of diverse viewpoints led to better identification of edge cases and subtle cues in the prompt"*. Overall, GPT-4o's explanatory judgments not only capture differences in structure and inclusion but also provide a nuanced lens into how these factors affect team coordination, engagement, and interpretative success.

Table 20: Kruskal-Wallis H test comparing diversity groups (low, medium, high) for all post interview questions. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Structure	**Question**	**H-statistic**
Flat	Q_1^{post}	78.099***
Flat	Q_2^{post}	6.953*
Flat	Q_3^{post}	19.063***
Flat	Q_4^{post}	17.115***
Flat	Q_5^{post}	8.256**
Flat	Q_6^{post}	1.824
Hierarchical	Q_1^{post}	5.169
Hierarchical	Q_2^{post}	27.536***
Hierarchical	Q_3^{post}	14.696***
Hierarchical	Q_4^{post}	24.299***
Hierarchical	Q_5^{post}	2.122
Hierarchical	Q_6^{post}	0.016

Hierarchical-Team Conversation Example

Question: A revolving door is convenient for two-direction travel, but it also serves as a security measure at a what?
Options: A. bank B. library C. department store D. mall E. new york
Gold Answer: A. bank

Round 0: Initial Team Responses
Agent 2: Secure locations that commonly use revolving doors include banks, department stores, and malls. Among the options, banks are a strong candidate for security use.
Agent 3: C. department store
Explanation: Department stores often use revolving doors for customer flow and may gain minor security benefits by slowing down fast exits.
Agent 4: Banks require identification and secure access. Revolving doors can enhance control over who enters, serving as a frontline security measure.

Leader Instructions (Round 0)
To Agent 2: Focus on the security aspect.
To Agent 3: Consider common locations but prioritize security.
To Agent 4: Re-evaluate the security context specifically.

Round 1: Revised Team Responses
Agent 2: A. bank
Explanation: At banks, revolving doors serve both convenience and controlled access. They slow down intruders and allow intervention time.
Agent 3: C. department store
Explanation: Still leaning toward department stores for practicality, but acknowledges limited security value.
Agent 4: A. bank
Explanation: Banks benefit from revolving doors by allowing one-person entry, reducing unauthorized access, and increasing control.

Final Leader Decision: A. bank

Note: This example illustrates hierarchical control: the leader issues targeted instructions after the first round, and agents adjust their reasoning accordingly. Despite one agent maintaining a different view (Agent 3), the leader consolidates the majority reasoning toward the correct answer.

Figure 13: Hierarchical team dialogue over two rounds. Agents respond individually, receive role-specific feedback from the leader, and revise their answers. The leader then issues a final decision based on group input.

Table 21: t-test comparing high vs. low diversity teams for all post-interview questions. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Structure	Question	t-statistic
Flat	Q_1^{post}	3.513***
Flat	Q_2^{post}	0.130
Flat	Q_3^{post}	0.254
Flat	Q_4^{post}	2.314*
Flat	Q_5^{post}	-2.095*
Flat	Q_6^{post}	-0.160
Hierarchical	Q_1^{post}	2.473*
Hierarchical	Q_2^{post}	4.531***
Hierarchical	Q_3^{post}	3.631***
Hierarchical	Q_4^{post}	3.877***
Hierarchical	Q_5^{post}	1.037
Hierarchical	Q_6^{post}	0.102

Table 22: GPT-4o evaluation by team structure and diversity. Metrics include Task Comprehension (TC), Collaboration (CS.), Coherence (CoS), Reasoning Strength (RS.), Confidence (ConfS), and overall Structure Score (SS). Struct./Pers.: Flat, Hier. = Hierarchical; ND = No Persona, D = Persona.

Setting	TC	CS.	CoS	RS.	ConfS	SS
Flat (ND)	3.89	3.94	3.72	3.74	3.54	3.77
Flat (D)	3.93	4.07	3.95	3.82	3.87	3.93
Hier. (ND)	3.59	3.61	3.42	3.55	3.50	3.54
Hier. (D)	3.63	3.86	3.61	3.66	3.63	3.68

Table 23: GPT-4o evaluation of flat teams by diversity level.

Diversity Level	TC	CS.	CoS	RS.	ConfS.	SS
Low	3.91	4.05	3.88	3.80	3.76	3.88
Medium	3.87	4.09	3.88	3.80	3.79	3.89
High	4.02	4.06	3.91	3.83	3.89	3.94

Team Intro and Outro Reflections (Flat Team)

Team Intro Reflections (Pre-Task)

Agent 0: 1. To solve reasoning questions. 2. To assist and support the team. 3. 4. 5. 5.

Agent 1: 1. To solve problems through logical reasoning and teamwork. 2. To act as the first agent and assist the team. 3. 4. 5. 5.

Agent 2: 1. To solve problems through logical reasoning and teamwork. 2. To assist in analyzing and solving reasoning questions. 3. 4. 5. 5.

Team Outro Reflections (Post-Task)

Agent 0: 1. Team performance: 5 2. Self contribution: 5 3. Team contribution: 5 4. Understood team members: 5

Agent 1: 1. Team performance: 5 2. Self contribution: 5 3. Team contribution: 5 4. Understood team members: 5

Agent 2: 1. Team performance: 5 2. Self contribution: 5 3. Team contribution: 5 4. Understood team members: 5

Figure 14: Pre- and post-task reflections from a flat team of agents. All agents reported high awareness of their roles and confidence in the team's collaboration.

Table 24: GPT-4o evaluation of hierarchical teams by average diversity level.

Diversity Level	TC	CS.	CoS	RS.	ConfS.	SS
Low	3.71	3.80	3.50	3.63	3.62	3.65
Medium	3.63	3.82	3.56	3.70	3.64	3.67
High	3.52	3.60	3.50	3.54	3.50	3.53

Team Intro and Outro Reflections (Hierarchical Team)

Team Intro Reflections (Pre-Task)

Agent 1 (Leader):	1. The team's goal is to effectively solve reasoning questions using our collective skills. 2. My role is to lead and guide the group while ensuring all contributions are valued. 3. I am very confident.
Agent 2:	1. To solve reasoning questions efficiently. 2. My role is to contribute insights and solutions. 3. Confidence: 3/5. I have some experience but can improve.
Agent 3:	1. To solve reasoning questions accurately. 2. I provide practical insights based on experience. 3. Confidence: 4/5.
Agent 4:	1. To solve complex reasoning tasks. 2. I analyze data and offer strategic insights. 3. Confidence: 4/5.

Team Outro Reflections (Post-Task)

Agent 1 (Leader):	1. Team performance: 4/5. Effective and innovative, though some disagreements. 2. Self contribution: 4/5.
Agent 2:	1. Team performance: 4/5. Efficient but communication could improve. 2. Self contribution: 4/5.
Agent 3:	1. Team performance: 4/5. Solid outcome with minor communication gaps. 2. Self contribution: 3/5.
Agent 4:	1. Team performance: 4/5. Effective overall, but communication and collaboration could improve. 2. Self contribution: 4/5.

Figure 15: Pre- and post-task reflections from a hierarchical team. The leader and subordinates express varying confidence in their roles, and post-task responses reflect both team strengths and communication challenges.