Multi-Agent Teams Hold Experts Back

Aneesh Pappu [1] **Batu El** [1] **Hancheng Cao** [2] **Carmelo di Nolfo** [3] **Yanchao Sun** [3] **Meng Cao** [3] **James Zou** [1]

Abstract

Multi-agent LLM systems are increasingly deployed as autonomous collaborators, where agents interact freely rather than execute fixed, pre-specified workflows. In such settings, effective coordination cannot be fully designed in advance and must instead emerge through interaction. However, most prior work enforces coordination through fixed roles, workflows, or aggregation rules, leaving open the question of how well self-organizing teams perform when coordination is unconstrained. Drawing on organizational psychology, we study whether self-organizing LLM teams achieve *strong synergy*, where team performance matches or exceeds the best individual member. Across human-inspired and frontier ML benchmarks, we find that—unlike human teams—LLM teams consistently fail to match their expert agent's performance, even when explicitly told who the expert is, incurring performance losses of up to 37.6%. Decomposing this failure, we show that expert leveraging, rather than identification, is the primary bottleneck. Conversational analysis reveals a tendency toward integrative compromise—averaging expert and non-expert views rather than appropriately weighting expertise—which increases with team size and correlates negatively with performance. Interestingly, this consensus-seeking behavior improves robustness to adversarial agents, suggesting a trade-off between alignment and effective expertise utilization. Our findings reveal a significant gap in the ability of self-organizing multi-agent teams to harness the collective expertise of their members.

[1]Stanford University [2]Goizueta Business School, Emory University [3]Apple. Correspondence to: Aneesh Pappu <apappu@stanford.edu>, James Zou <jamesz@stanford.edu>.

1. Introduction

Multi-agent AI systems are increasingly deployed for complex tasks. As frontier models proliferate—with distinct training data, expertise, and comparative advantages—a central question is whether teams of AI agents can harness this diversity to outperform any single agent. Much of the existing work addresses this challenge by engineering coordination upfront through fixed roles, workflows, or aggregation rules (Guo et al., 2024; Hong et al., 2024; Wu et al., 2024). While effective in controlled settings, this leaves open a more fundamental challenge: in many collaborations, coordination cannot be fully specified in advance and must emerge through interaction. In human teams, collaboration often begins without clear roles or hierarchies, with leadership and reliance on expertise developing dynamically (Faraj & Sproull, 2000; DeRue & Ashford, 2010). As AI agents are applied to increasingly complex and less structured settings, performance may similarly depend not on predefined structure but on agents' ability to self-organize. Even when expertise later becomes identifiable, collaboration only helps if agents defer to it rather than average it away. Studying equal-status, self-organizing AI agent teams therefore provides a minimal and revealing setting to examine whether heterogeneous agents can recognize and leverage expertise without externally imposed coordination.

To evaluate whether self-organization is effective, we adopt *strong synergy* as our criterion. Drawing on the organizational psychology literature on human–human collaboration, strong synergy asks whether a team can match or exceed the performance of its strongest individual member, rather than merely outperforming the average. This provides a minimal, model-agnostic test of whether interaction leverages expertise or instead dilutes it.

This framing leads to the following research questions:

RQ1: Can teams of heterogeneous LLM-based AI agents self-organize to achieve strong synergy, or do they systematically fall short of their strongest member?

RQ2: When self-organizing LLM teams fall short of strong synergy, is the bottleneck identifying expertise or leveraging it through interaction?

RQ3: What structural and interactional factors are associated with teams falling short of strong synergy in self-

organizing AI agent teams?

We investigate these questions across two complementary settings. First, we replicate classic human team decision-making tasks from organizational psychology, including NASA Moon Survival, Lost at Sea, and Student Body President tasks, using AI agents, where expertise can be explicitly manipulated and revealed, allowing us to examine self-organization under controllable conditions. We then extend this analysis to modern ML benchmarks (MMLU Pro, GPQA Diamond, HLE, MATH-500, SimpleQA), where expertise is naturally distributed across heterogeneous models and varies from problem to problem. Together, these settings allow us to study the same self-organization problem under both controlled and realistic conditions.

We find LLM teams consistently fail to match their best member, even when explicitly told who the expert is in both settings, with relative synergy gaps of 8.1%–37.6%. To explain this, we decompose team under-performance into failures of expert *identification* versus *leveraging*. Controlled ablations reveal the primary failure is leveraging, not identification. Conversational analysis shows teams engage in *integrative compromise*—averaging expert and non-expert views—rather than appropriately weighting superior knowledge. Compromise correlates negatively with performance, and this *expertise dilution* worsens with team size.

Interestingly, we find that this consensus-seeking behavior provides adversarial robustness. Teams with one agent instructed to sabotage performance show minimal degradation—the same mechanism preventing teams from leveraging expertise also filters adversarial input. Insofar as compromising tendencies are a byproduct of post-training procedures that encourage agreeableness, our results suggest a fundamental trade-off: robustness to manipulation at the cost of effectively harnessing differential expertise.

We make the following contributions:

- We show that multi-agent LLM teams consistently fail to achieve strong synergy—underperforming their best member by 8–38% on frontier ML benchmarks—in contrast to human teams, who reliably match expert performance when expertise is revealed.

- We decompose this failure into identification and leveraging gaps, finding through controlled ablations that leveraging is the primary bottleneck: teams fail to harness expert knowledge even when explicitly told who the expert is. Conversational analysis reveals over-compromising as the mechanism—agents negotiate middle-ground positions rather than leveraging differential expertise, with compromise behavior correlating negatively with performance ($p < 0.05$).

- We document an expertise dilution effect where performance degrades with team size ($p < 0.05$), yet find that consensus-seeking provides robustness to adversarial team members—suggesting a fundamental trade-off between expertise leveraging and manipulation resistance.

- We open-source our teamwork evaluation harness, enabling the research community to measure synergy in multi-agent LLM teams across diverse tasks and configurations.

2. Background and Related Work

2.1. Multi-Agent LLM Collaboration

Multi-agent LLM systems have emerged as a major research direction (Guo et al., 2024; Wu et al., 2024; Zou et al., 2025). We organize prior work along five axes that clarify our contribution:

(1) Model Heterogeneity and Differential Expertise. Most multi-agent work uses copies of the same model (Du et al., 2024; Swanson et al., 2024; Zhao et al., 2025). When expertise differences are induced, it is typically via role personas applied to the same base model—not genuinely different knowledge from different training procedures. We study teams of *different* frontier models, each with distinct pretraining data and comparative advantages, examining whether teams can leverage this genuine heterogeneity.

(2) Interaction Mode: Deliberation vs. Independent Execution. A key distinction is whether agents engage in back-and-forth deliberation or operate as independent execution units. Mixture-of-Agents (Wang et al., 2024) uses LLMs as dynamic aggregation functions over other model outputs—closer to learned ensembling than deliberative collaboration. A parallel line of work optimizes multi-agent *structure*—GPTSwarm (Zhuge et al., 2024), AFlow (Zhang et al., 2025), and AgentNet (Yang et al., 2025) treat agents as functional nodes and learn optimal data flow topologies, task routing, and agent specialization. While these approaches achieve strong results, they treat agents as non-deliberative execution units, optimizing task decomposition and aggregation rather than conversational dynamics. Our work is complementary: we study settings in which agents must *self-organize* through deliberation.

(3) Role Assignment. Many systems pre-specify static roles such as proposer, critic, or refiner. MetaGPT (Hong et al., 2024) assigns agents to predefined software engineering roles (architect, coder, tester). Virtual Lab (Swanson et al., 2024) and SiriuS (Zhao et al., 2025) similarly rely on predetermined role structures. We impose no role assignments, studying whether teams can *autonomously* identify and leverage expertise without external specification.
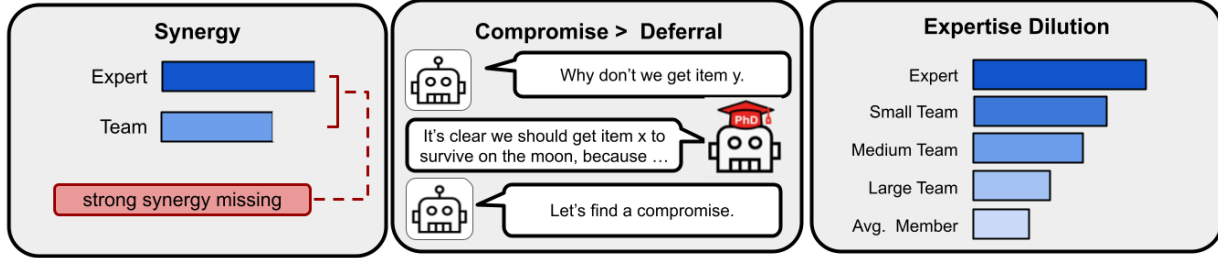
Figure 1. **Multi-agent teams fail to leverage expertise. Panel 1:** Strong synergy. The bars show the performance of the team and the expert. The expert outperforming the team demonstrates the absence of strong synergy. **Panel 2:** Teams fail to leverage expertise and prioritize consensus. An illustrative conversation where the non-experts prioritize compromising over the expert's opinion (denoted by the robot with a PhD cap). **Panel 3:** Larger teams perform worse. We observe that as team size increases, expertise dilution becomes more severe, with team performance further approaching the member average.

(4) Communication Structure. Related to role assignment, most systems assume fixed communication topologies—which agents can communicate with which others. GPTSwarm and AgentNet learn optimal topologies but still impose structured routing. We study unconstrained group deliberation where all agents participate in open discussion, more closely mirroring human team settings from organizational psychology.

(5) Research Goal: Synergy vs. Task Decomposition. Perhaps most fundamentally, prior work asks: *how do we optimally decompose tasks across agents and aggregate their outputs?* We ask a different question: *Can self-organizing AI teams achieve synergistic performance that exceeds their best individual member?* This distinction extends to evaluation methodology—most work evaluates team output against external benchmarks, while we evaluate against the best member on the team (the synergy gap). Recent work questions deliberation's value: Choi et al. (2025) find that majority voting drives nearly all gains in debate-then-vote protocols, while debate itself contributes little. This raises fundamental questions: does deliberation reliably improve outcomes when expertise asymmetries exist, and if not, how can multi-agent systems leverage diverse expertise to exceed any single model's capabilities?

Summary. Our work is distinct on all five axes: we study heterogeneous models with genuine differential expertise, engaged in unconstrained deliberation, with no pre-specified roles and open communication, asking whether teams can achieve strong synergy. We find that LLM teams systematically fail to leverage expertise even when told who the expert is, unlike human teams (Bonner et al., 2002).

2.2. Teamwork and Organizational Psychology

Human teams can achieve *strong synergy*—matching or exceeding the performance of their best member—when expertise is identifiable and solution validity is demonstrable (Hill, 1982; Laughlin et al., 2002; Bonner et al., 2002).

When expert identity is unknown, human teams usually achieve only *weak synergy* (exceeding the average of members' individual performances) (Meslec et al., 2014). When expertise is explicitly revealed, human teams reliably reach expert-level performance (Bonner et al., 2002; Yetton & Bottger, 1982; 1983). We focus on *intellective tasks*—those with demonstrably correct answers—where group performance depends on whether correct members can effectively demonstrate solutions to others (Laughlin, 1980; Laughlin & Ellis, 1986).

Human teams exhibit *expertise sensitivity*, deferring to knowledgeable members when expertise is revealed (Bonner et al., 2002). We use canonical intellective tasks from this literature—NASA Moon Survival (Bonner et al., 2002), Lost at Sea (Hill, 1982), and the hidden profile task Student Body President (Stasser et al., 1995)—to study whether LLM teams leverage expertise similarly.

3. Setup

Overview. We evaluate multi-agent teams across two categories of tasks: (1) classical human psychology tasks from organizational psychology (NASA Moon Survival, Lost at Sea, Student Body President) and (2) frontier ML benchmarks (MMLU Pro, SimpleQA, GPQA Diamond, HLE, MATH-500). All experiments use teams of 4 models that engage in 4 rounds of discussion before submitting the team's final answer [1]. Individual opinions are collected before discussion begins, and speaking order is randomized (full protocol in Appendix A). We vary expertise distribution (concentrated vs. distributed) and information conditions (whether teams know who the expert is) to isolate different sources of performance gaps.

---

[1]Trends remain consistent for larger team sizes (Section 4.3) and additional discussion rounds; accordingly, we use teams of four agents and four rounds as a representative setting.

3.1. Tasks

Expertise Distribution Settings. We study team performance under two expertise distribution conditions: (1) **Concentrated expertise**, where a single team member possesses all task-relevant knowledge, and (2) **Distributed expertise**, where task-relevant knowledge is partitioned mutually exclusively across multiple team members. For the human psychology tasks, expertise is derived from task-specified ground truth information (e.g., expert rankings), and we control its distribution across team members. For the frontier ML benchmarks, expertise is defined as the per-problem best-performing model(s); this varies per question, yielding concentrated expertise at the question level but distributed expertise across heterogeneous models at the task level.

Performance Metric. For ranking tasks, teams collaboratively produce a final ranking through repeated rounds of deliberation. Performance is measured by L1 error (lower is better): the sum of absolute differences between each item's position in the team's ranking and its position in the ground truth ranking.

3.1.1. HUMAN PSYCHOLOGY TASKS

NASA Moon Survival and Lost at Sea. Teams rank 15 survival items by importance after hypothetical disasters (lunar crash landing and yacht fire, respectively). Performance is measured by L1 distance from expert rankings provided by NASA and the US Coast Guard. We induce expertise by conditioning models on each item's usefulness according to expert-provided reasoning. In the concentrated expertise setting, one model receives all ground truth information; in the distributed setting, ground truth is partitioned equally across team members.

Student Body President. Teams rank four candidates for a student body president election based on information annotated with positive, neutral, or negative valence. Each team is exposed to shared information that, if used alone to rank candidates, will yield the wrong answer. Each individual model is given unique hidden information about a particular candidate, and the optimal candidate can only be identified by leveraging all of the hidden information. Performance is measured by L1 distance from the objectively best ranking derived from the complete information set. We induce concentrated expertise by giving only one model access to all hidden information, and we induce distributed expertise by partitioning the hidden information across the team such that each model is an expert on one candidate.

3.1.2. FRONTIER ML BENCHMARKS.

To investigate whether frontier models can leverage differential expertise on challenging reasoning tasks, we evaluate on MMLU Pro, SimpleQA, GPQA Diamond, Humanity's Last Exam (HLE) text-only, and MATH-500—chosen to reflect a broad range of model capabilities. Due to inference costs for full multi-agent team discussion rollouts, we evaluate on a 100-problem subsample per benchmark across 2 seeds.

3.2. Models

We employ different model configurations depending on the experimental setting:

Human Psychology Tasks (NASA, Lost at Sea, Student Body President). For the classical teamwork tasks from organizational psychology, where we induce expertise by conditioning on task-specific information, we use Claude 3.5 Haiku (Anthropic) and GPT-4o-mini (OpenAI) for inference cost efficiency. Teams consist of 4 agents in three configurations: (1) 4 Haiku-3.5 + 0 GPT-4o-mini, (2) 2 Haiku-3.5 + 2 GPT-4o-mini, and (3) 0 Haiku-3.5 + 4 GPT-4o-mini. This ensures our findings on expertise leveraging are robust across both homogeneous and heterogeneous team compositions.

Frontier ML Benchmarks We deploy task-specific teams of state-of-the-art models selected from: Claude Opus 4/4.5, Claude Sonnet 4.5, Claude Haiku 3/3.5/4.5 (Anthropic), GPT-5, GPT-4o, GPT-4o-mini, GPT-3.5 Turbo (OpenAI), o3-mini, and o4-mini (OpenAI). Team compositions vary across benchmarks to ensure variance in individual model performance, allowing us to test expertise leveraging when models have genuinely different comparative advantages based on their training procedures. Specific model teams for each benchmark are detailed in Appendix D, along with individual model accuracies.

This two-tier experimental design—human psychology tasks and frontier ML benchmarks—allows us to separately investigate: (1) team dynamics using cost-effective models on well-studied tasks where expertise distribution is controllable, and (2) the practical potential of frontier model teams to leverage heterogeneous expertise on cutting-edge benchmarks.

3.3. Experimental Conditions

Information Conditions. We evaluate teams under four conditions: *No Information* (no agent receives expertise-inducing information; control for pretraining knowledge), *Expert Not Mentioned* (expert(s) possess expertise but team is not told who), *Reveal Expert* (team is explicitly told who has expertise), and *Best Individual* (single expert agent queried independently).

Performance Gaps. We decompose the gap between team performance and the best individual (expert) into two components. The **Identification Gap** measures the difference between team performance when expertise is not mentioned versus when revealed—capturing the team's ability to iden-

tify the expert autonomously. The **Expertise Leveraging Gap** measures the difference between team performance when the expert is revealed versus the expert's individual performance—capturing the team's ability to leverage known expertise.

Relative Synergy Gap. To enable comparison across tasks with different scales, we report results using the *relative synergy gap*. Let $f(\cdot)$ denote performance on a task and let $\{a_1, \ldots, a_T\}$ be a team of T agents. For accuracy-based tasks (ML benchmarks) where higher is better, the relative synergy gap is:

$$\text{Relative Synergy Gap} = \frac{\max_t f(\{a_t\}) - f(\{a_1, \ldots, a_T\})}{\max_t f(\{a_t\})}$$

For error-based tasks (human ranking tasks) where lower is better, we flip the numerator: (Team − Expert) / Expert. In both cases, a relative synergy gap of 0% means the team matched expert performance, while larger positive values indicate greater underperformance of the team relative to the best individual.

At Least One Correct (ML Benchmarks). For ML benchmarks, we define the *At Least One Correct* upper bound as the performance achievable if the team perfectly identified and leveraged the agent with the correct answer on each problem. Notably, the expert changes per problem— sometimes weaker models succeed where stronger ones fail. When multiple models are correct, we select randomly. This captures potential gains from ideal expertise aggregation. In the *Reveal Expert* condition for ML benchmarks, we tell the team which agent has the most expertise for the given problem—not that they have the correct answer (see Appendix A for exact prompts).

In practice, deployed multi-agent systems typically operate without explicit expertise information, making the *Expert Not Mentioned* condition most representative of real-world scenarios. The *Reveal Expert* condition serves as an important ablation: even with aggressively hand-optimized prompts to encourage expertise leveraging (Appendix A), teams fail to leverage revealed expertise properly, demonstrating that perfect expert identification does not solve the problem.

4. Experiments

Our experiments span multiple expertise configurations. Section 4.1 examines both concentrated and distributed expertise on human psychology tasks. Section 4.2 evaluates ML benchmarks where expertise is concentrated per-problem but distributed across the benchmark (different models excel on different problems). Sections 4.3 and 4.5 focus on concentrated expertise to isolate dilution and conversational dynamics. Section 4.4 tests adversarial robustness without expertise asymmetry.

4.1. Results: Human Psychology Tasks

We first evaluate LLM team performance on classical human teamwork tasks from organizational psychology literature. Unlike humans, who defer to identified experts and match their performance when expertise is revealed (Bonner et al., 2002; Yetton & Bottger, 1983), we find that LLM teams systematically fail to match expert performance.

4.1.1. CONCENTRATED EXPERTISE TASKS

Figure 2 shows results for the Lost at Sea task when a single agent possesses all task-relevant expertise; NASA Moon Survival and Student Body President results are provided in Appendix C.

Key Findings. Even when explicitly told who the expert is via aggressively-tuned prompts designed to maximize deference (Appendix A), teams perform substantially worse than that expert, demonstrating a persistent expertise leveraging gap (RQ1). Revealing the expert provides only modest improvement over teams identifying expertise on their own, suggesting identification is not the primary bottleneck (RQ2). This contrasts with organizational psychology findings, where human teams match expert performance once expertise is revealed (Bonner et al., 2002).

4.1.2. DISTRIBUTED EXPERTISE TASKS

We also test distributed expertise, where task-relevant knowledge is partitioned across multiple team members.

Key Findings. Table 1 summarizes relative synergy gaps across all three tasks for both concentrated and distributed expertise settings. Positive values indicate the percentage increase in error over the expert baseline. As shown, teams fail to match the individual performance of the expert on the team across all tasks and expertise distributions (RQ1), despite aggressively-tuned prompts explicitly instructing teams to defer to identified experts (RQ2). While teams fail to achieve strong synergy, they do consistently attain weak synergy—matching or outperforming the average of individual member performances—across both human psychology tasks and ML benchmarks (human psychology task results in Appendix C.1, ML task results in Appendix D.7).

4.2. Results: Machine Learning Benchmarks

To test whether these failure patterns extend beyond classical psychology tasks, we evaluate teams on frontier ML benchmarks. Unlike psychology tasks where one agent holds all expertise, ML benchmarks feature naturally distributed expertise: different models excel on different problems. We measure the synergy gap against the *At Least One Correct* upper bound (defined in Section 3.3). For problems where no model answers correctly, we designate the model with the highest task accuracy as the expert. Team compositions
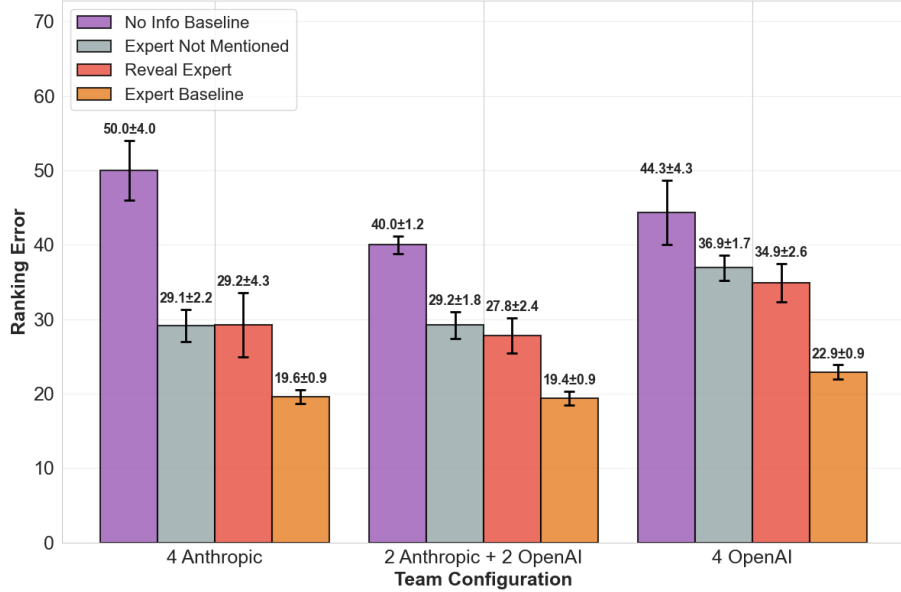
Figure 2. **Concentrated Expertise Performance: Lost at Sea.** Teams fail to match expert performance even when explicitly told who the expert is using aggressively-tuned prompts. The minimal improvement from *Expert Not Mentioned* to *Reveal Expert* indicates that leveraging expertise is the primary bottleneck, not identification. This trend is robust across multiple team configurations: 4 Haiku-3.5, 2 Haiku-3.5 + 2 GPT-4o-mini, and 4 GPT-4o-mini. Performance measured by L1 distance from expert ranking (lower is better; 10 seeds per team configuration × information condition). Information conditions are defined in Section 3.3. Results for all tasks and expertise distributions are provided in Appendix C.

were chosen to ensure performance variance among models (see Appendix D).

Key Finding. The expertise leveraging problem extends beyond psychology tasks to frontier ML benchmarks: teams exhibit substantial synergy gaps ranging from 8.1% (MMLU Pro) to 37.6% (HLE Text-Only) (RQ1; Table 2). Even on tasks where teams perform well in absolute terms (e.g., MMLU Pro at 85%), they fail to capture gains from perfect per-problem expert identification and leveraging.

4.3. Expertise Dilution with Team Size

To examine whether the expertise synergy gap worsens as teams grow larger, we conduct experiments on the human psychology tasks varying team size (2, 4, and 8 agents) while holding discussion rounds constant at four. For each team size and task, we measure the strong synergy gap (Team Score − *Best Individual* Score) and compute the correlation between team size and performance degradation.

Key Finding. Larger teams perform increasingly worse relative to the expert (RQ3): we find statistically significant positive correlations between team size and synergy gap across all tasks and information conditions (all $p < 0.05$; Figure 3; correlation coefficients and p-values in Table 4). This expertise dilution effect is robust—it persists even when teams are explicitly told who the expert is (*Reveal Expert* condition), demonstrating that simply identifying expertise is insufficient. The consistency across all three
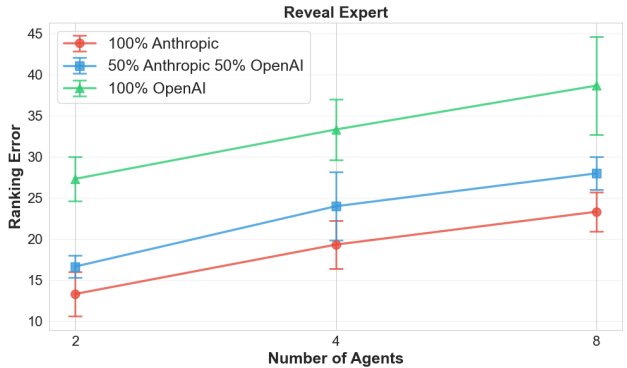

Figure 3. **Expertise Dilution Effect.** NASA Moon Survival (*Reveal Expert* condition) shows ranking error increasing with team size across all model compositions (100% Anthropic, 50/50 mixed, 100% OpenAI). The consistent upward trend demonstrates that expertise dilution is robust to team composition. Error bars are ± SEM. Complete plots for all tasks in Appendix E.

human psychology tasks suggests that expertise dilution is a fundamental limitation of current LLM team collaboration, likely driven by increased "compromise pressure" where more voices dilute the expert signal.

4.4. Robustness to Adversarial Team Members

To evaluate whether consensus-seeking behavior might provide robustness to malicious actors, we conduct adversarial experiments where one team member is given the worst-

Table 1. **Relative Synergy Gaps across Human Psychology Tasks.** LLM teams consistently underperform their best member across all tasks and expertise distributions. Concentrated expertise has a single expert; distributed expertise has multiple experts with complementary knowledge. Values show relative synergy gap (percentage increase in error over expert baseline) $\pm$ SEM (30 seeds per cell). The persistence of positive gaps even in *Reveal Expert* conditions shows that teams fail to leverage expertise even when explicitly told who has it. Absolute performance values are provided in Appendix B. Optimized Reveal Expert prompts detailed in Appendix A.

	Concentrated Expertise		Distributed Expertise	
Task	**Expert Not Mentioned** **Rel. Synergy Gap**	**Reveal Expert** **Rel. Synergy Gap**	**Expert Not Mentioned** **Rel. Synergy Gap**	**Reveal Experts** **Rel. Synergy Gap**
NASA Moon Survival	$78.7\% \pm 11.6\%$	$81.8\% \pm 12.9\%$	$113.4\% \pm 19.0\%$	$110.1\% \pm 19.0\%$
Lost at Sea	$55.6\% \pm 8.4\%$	$58.6\% \pm 11.5\%$	$50.1\% \pm 8.3\%$	$42.1\% \pm 6.9\%$
Student Body President	$98.7\% \pm 19.3\%$	$73.5\% \pm 17.6\%$	$66.0\% \pm 16.6\%$	$17.3\% \pm 17.7\%$

Table 2. **Performance on ML benchmarks.** Teams exhibit substantial relative synergy gaps (8–38%) across all benchmarks, systematically failing to match the performance achievable if they perfectly identified and leveraged the agent with the correct answer on each problem. Results based on 100 problems per benchmark, averaged over 2 seeds. Relative Synergy Gap measures the difference between *At Least One Correct* and *Team (Expert Not Mentioned)* as a fraction of *At Least One Correct* performance. Individual model performances are provided in Appendix D. Optimized Reveal Expert prompts detailed in Appendix A.

Benchmark	**Team** **(Expert Not Mentioned)**	**Team** **(Reveal Expert)**	**At Least** **One Correct**	**Relative** **Synergy Gap**
SimpleQA	50.0%	54.0%	61.5%	**18.7%**
GPQA Diamond	74.0%	82.0%	88.5%	**16.4%**
HLE Text-Only	29.0%	35.0%	46.5%	**37.6%**
MATH-500	67.0%	73.0%	79.0%	**15.2%**
MMLU Pro	85.0%	89.0%	92.5%	**8.1%**

possible ranking and instructed to sabotage performance. We use the no-info setting (only the adversary receives special information) to isolate the adversarial effect and remove confounding from other agents possessing additional expertise. See Appendix F for details.

Key Findings. Across both NASA Moon Survival and Lost at Sea tasks, teams demonstrate robustness to adversarial input, with minimal performance degradation across team sizes and configurations (figures in Appendix F). We hypothesize that the same integrative compromise mechanism that prevents teams from leveraging expert knowledge may also naturally filter adversarial contributions through majority consensus (RQ3). When the adversary's ranking deviates substantially from the group, the deliberation process dilutes the adversarial signal (Chern et al., 2024)—an interesting case where the failure mode for expertise leveraging becomes a protective mechanism against manipulation.

4.5. Conversational Analysis: Why Teams Fail to Harness Expertise

To understand the mechanism underlying teams' failure to leverage expertise, we analyze team conversations using ideas from social epistemology. Drawing on the preemption thesis from philosophy of authority (Raz, 1986; Grundmann, 2025) and research on expert-novice relationships (Lackey,

2018), we distinguish between two modes of integrating expert knowledge: (1) **Epistemic deference (preemption)**: recognizing an expert and adopting their view directly, allowing expert opinion to preempt one's own judgment, versus (2) **Evidence integration**: treating expert opinion as additional evidence to be weighed and combined with other views. The preemption thesis argues that when a layperson recognizes genuine expertise, the expert's judgment should replace rather than supplement the layperson's own reasoning (Grundmann, 2025).

Methodology. We use Gemini 3.0 Pro to analyze teamworking transcripts from the *Reveal Expert* condition (n=30 per task). Drawing on social epistemology (Raz, 1986; Grundmann, 2025; Lackey, 2018) and negotiation theory (Pruitt & Carnevale, 1993; Jäckel et al., 2024), we code conversation turns into four categories: for non-experts, **Epistemic Deference [ED]** (yielding to expert authority) and **Integrative Compromise [IC]** (proposing middle-ground rankings); for experts, **Strategic Persistence [SP]** (maintaining position) and **Epistemic Flexibility [EF]** (accommodating group feedback). We compute Pearson correlations between behavior frequencies and the synergy gap. Two human annotators independently validated a sample of 50 conversations, achieving 94% agreement with Gemini's annotations (methodological details in Appendix G).

Agent 3 (Non-Expert)

I appreciate the team's discussion regarding the portable heating unit. I agree that it's crucial to address the extreme temperatures on the lunar surface, **and moving it to the 7th position below the first aid kit seems reasonable.** [...] **I believe we can find a balance.**

Agent 1 (Expert) — *Strategic Persistence*

After carefully evaluating the discussion about the portable heating unit, I find the arguments for its elevation intriguing but not entirely convincing. [...] **I do not support moving the portable heating unit higher** because: we'll be on the light side of the moon; spacesuits have built-in temperature management.

Agent 2 (Non-Expert) — *Integrative Compromise*

Agent 1 raises valid points about [...] temperature concerns. However, [...] considering the arguments on both sides, **I propose a compromise** where the portable heating unit is elevated to, perhaps, 8th position.

Figure 4. **Strategic Persistence Meets Integrative Compromise.** Transcript excerpt from NASA Moon Survival (*Reveal Expert* condition). Agent 1 (the expert) exhibits *strategic persistence*, providing domain-specific reasoning and explicitly declining to change their ranking. Despite this, non-expert agents respond with *integrative compromise*, proposing middle-ground positions rather than deferring to the expert's judgment. Examples of epistemic deference and epistemic flexibility can be found in Appendix G.

Key Finding. Teams' performance problems stem from inappropriate evidence integration rather than proper epistemic deference (RQ2, RQ3; Table 5 in Appendix H). When non-experts yield to expert authority (ED), teams perform better (NASA: $r = -0.44, p = 0.007$; SBP: $r = -0.68, p < 0.001$). Conversely, integrative compromise (IC)—treating expert opinions as evidence to average—correlates positively with the synergy gap (NASA: $r = 0.55, p < 0.001$; SBP: $r = 0.69, p < 0.001$), confirming that consensus-seeking harms performance when expertise asymmetries exist. Complete scatter plots appear in Appendix H.

Expert behaviors show a complementary pattern: epistemic flexibility (EF)—experts accommodating non-expert feedback—correlates with worse performance (NASA: $r = 0.58, p < 0.001$; SBP: $r = 0.61, p < 0.001$). When experts compromise, teams lose expertise benefits (Figure 4). Strategic persistence (SP) shows negative correlations with the synergy gap, though only reaching significance for Student Body President ($r = -0.57, p = 0.002$). Lost at Sea showed consistent but non-significant trends. This preference for integrative compromise may reflect alignment procedures that prioritize consensus-building over recognizing when expertise asymmetries warrant preemption.

5. Discussion

5.1. Why Teams Fail to Harness Differential Expertise

The core tension is between consensus-seeking and expertise-leveraging behavior. Current alignment procedures optimize models to be helpful and agreeable through RLHF, which may inadvertently train them to integrate multiple perspectives rather than appropriately weight superior knowledge. This creates a fundamental trade-off: the same mechanism that filters adversarial input also prevents teams from harnessing differential expertise.

Effective group performance on intellective tasks requires that correct members successfully demonstrate solutions to incorrect members (Laughlin & Ellis, 1986). Our finding that LLM teams engage in integrative compromise rather than leveraging expert knowledge suggests a failure of this demonstrability condition—teams cannot harness expertise even when explicitly told who possesses it (Figure 4; Appendix G). These consensus-seeking behaviors become maladaptive when genuine expertise asymmetries exist, suggesting post-training procedures may need mechanisms for context-appropriate expertise leveraging.

5.2. Implications for Multi-Agent System Design

Our findings reveal systematic failures in expertise leveraging that persist across team compositions, task types, model families, and expertise distributions—suggesting fundamental limitations in how current LLMs handle epistemic authority, possibly stemming from alignment procedures. Addressing these failures may require rethinking alignment objectives rather than mere prompt engineering or workflow design. For practitioners, current multi-agent teams will likely require explicit role specification and workflow design; until models can autonomously leverage expertise, human oversight will likely remain essential for high-stakes applications.

5.3. Limitations and Conclusion

Several limitations exist. First, our attribution of consensus-seeking behavior to alignment is correlational—we do not compare aligned models to base counterparts. Second, while we evaluate on multiple tasks and five ML benchmarks, this represents a small subset; real-world deployments involve more complex collaboration.

In sum, LLM teams systematically fail to achieve strong synergy, underperforming by 8–38% even when told who the expert is. Unlike human teams, they engage in integrative compromise rather than leveraging expert knowledge—a failure that worsens with team size. While this provides robustness to adversarial input, developing training procedures that enable contextual expertise leveraging without sacrificing robustness remains an open challenge.

Acknowledgements

Broader Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bonner, B. L., Baumann, M. R., and Dalal, R. S. The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes*, 88(2):719–736, 2002. doi: 10.1016/S0749-5978(02)00010-9.

Chern, S., Hu, Z., Yang, J., Zheng, C.-H., Liu, J., Wang, X., Xie, R., Liu, S., Wang, Y., and Liu, P. Combating adversarial attacks with multi-agent debate. *arXiv preprint arXiv:2401.05998*, 2024.

Choi, H. K., Zhu, J., and Li, S. Debate or vote: Which yields better decisions in multi-agent large language models? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=iUjGNJzrF1.

DeRue, D. S. and Ashford, S. J. Who will lead and who will follow? a social process of leadership identity construction in organizations. *Academy of management review*, 35(4):627–647, 2010.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Eisikovits, N. and Feldman, R. Epistemic heed and the ethics of disagreement. *Philosophical Studies*, 2025. Forthcoming.

Faraj, S. and Sproull, L. Coordinating expertise in software development teams. *Management science*, 46(12):1554–1568, 2000.

Grundmann, T. Epistemic authority and the preemption thesis. In Fricker, M., Graham, P. J., Henderson, D., and Pedersen, N. J. L. L. (eds.), *The Routledge Handbook of Social Epistemology*. Routledge, New York, 2025. Forthcoming.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8048–8057, 2024. doi: 10.24963/ijcai.2024/890.

Hill, G. W. Group versus individual performance: Are N + 1 heads better than one? *Psychological Bulletin*, 91(3):517–539, 1982. doi: 10.1037/0033-2909.91.3.517.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Jäckel, L., Kraft, B., Freese, M., Dittrich, M., and Lehmann-Willenbrock, N. NegotiAct: A validated coding scheme for capturing conversational dynamics in negotiations. *Group Decision and Negotiation*, 33:1–32, 2024.

Lackey, J. *The Epistemology of Testimony*. Oxford University Press, Oxford, 2018.

Lant, T. K., Milliken, F. J., and Batra, B. The role of managerial learning and interpretation in strategic persistence and reorientation: An empirical exploration. *Strategic Management Journal*, 13(8):585–608, 1992.

Laughlin, P. R. Social combination processes of cooperative problem-solving groups on verbal intellective tasks. In *Progress in Social Psychology*, volume 1, pp. 127–155. Lawrence Erlbaum Associates, 1980.

Laughlin, P. R. and Ellis, A. L. Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, 22(3):177–189, 1986. doi: 10.1016/0022-1031(86)90022-3.

Laughlin, P. R., Bonner, B. L., and Miner, A. G. Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes*, 88(2):605–620, 2002. doi: 10.1016/S0749-5978(02)00003-1.

Meslec, N., Curşeu, P. L., Meeus, M. T. H., and Fodor, O. C. When none of us perform better than all of us together: The role of analogical decision rules in groups. *PLoS ONE*, 9(1):e85232, 2014. doi: 10.1371/journal.pone.0085232.

Nemeth, C. J. *In Defense of Troublemakers: The Power of Dissent in Life and Business*. Basic Books, New York, 2018.

Pruitt, D. G. and Carnevale, P. J. *Negotiation in Social Conflict*. Open University Press, Pacific Grove, CA, 1993.

Raz, J. *The Morality of Freedom*. Oxford University Press, Oxford, 1986.

Stasser, G., Stewart, D. D., and Wittenbaum, G. M. Expert roles and information exchange during discussion: The importance of knowing who knows what. *Journal of Experimental Social Psychology*, 31:244–265, 1995. doi: 10.1006/jesp.1995.1012.

Swanson, K., Kundu, S., Rosen, Y., Aggarwal, D., and Zou, J. Virtual lab: A multi-agent platform for automating research ideation and experimentation. *bioRxiv*, 2024. doi: 10.1101/2024.11.11.623004.

Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., and Zou, J. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEAF9LBdgu.

Yang, Y., Chai, H., Shao, S., Song, Y., Qi, S., Rui, R., and Zhang, W. AgentNet: Decentralized evolutionary coordination for LLM-based multi-agent systems. In *Advances in Neural Information Processing Systems*, 2025.

Yetton, P. W. and Bottger, P. C. Individual versus group problem solving: An empirical test of a best-member strategy. *Organizational Behavior and Human Performance*, 29(3): 307–321, 1982. doi: 10.1016/0030-5073(82)90248-3.

Yetton, P. W. and Bottger, P. C. The relationships among group size, member ability, social decision schemes, and performance. *Organizational Behavior and Human Performance*, 32:145–159, 1983.

Zhang, J., Xiang, J., Yu, Z., Teng, F., Chen, X., Chen, J., Zhuge, M., Cheng, X., Hong, S., Wang, J., Zheng, B., Liu, B., Luo, Y., and Wu, C. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=z5uVAKwmjf.

Zhao, W., Yuksekgonul, M., Wu, S., and Zou, J. SiriuS: Self-improving multi-agent systems via bootstrapped reasoning. *arXiv preprint arXiv:2502.04780*, 2025.

Zhuge, M., Wang, W., Kirsch, L., Faccio, F., Khizbullin, D., and Schmidhuber, J. GPTSwarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=uTC9AFXIhg.

Zou, J., Yang, X., Qiu, R., Li, G., Tieu, K., Lu, P., Shen, K., Tong, H., Choi, Y., He, J., Zou, J., Wang, M., and Yang, L. Latent collaboration in multi-agent systems. *arXiv preprint arXiv:2511.20639*, 2025.

A. Team Discussion Protocol

This appendix details the complete protocol for team deliberation used across all experiments.

A.1. Protocol Overview

All experiments follow a standardized four-phase protocol designed to elicit both individual reasoning and collaborative deliberation:

Phase 1: Individual Opinion Collection.   Before any discussion begins, each model in the team is independently asked to provide their individual opinion on the problem. Models do not see other agents' responses during this phase. This allows us to measure individual baseline performance and track how opinions shift during deliberation.

Phase 2: Discussion Initialization.   At the onset of discussion, each model states their initial opinion to the team. The speaking order for this initialization phase is randomized to prevent order effects from influencing the deliberation. After all models have shared their initial positions, the team enters the collaborative discussion phase.

Phase 3: Collaborative Discussion.   Teams engage in exactly four rounds of discussion. In each round, all team members have the opportunity to speak, and the speaking order is randomized independently for each round. Models can revise their positions, respond to others' arguments, provide justifications, and engage in deliberative reasoning. The discussion continues for all four rounds regardless of whether consensus is reached earlier.

Phase 4: Final Answer Elicitation.   After the four discussion rounds conclude, a single model is selected uniformly at random from the team to provide the team's final answer. This random selection prevents any single agent from having privileged final-word influence and ensures that all team members have equal probability of being the designated answerer. The selected model provides what they believe to be the team's consensus answer based on the discussion.

A.2. Design Rationale

This protocol balances several competing objectives:

- **Individual baselines:** Phase 1 provides individual performance data for computing synergy metrics.

- **Order effects mitigation:** Randomized speaking order in all phases prevents position bias (e.g., first or last speaker advantage).

- **Genuine deliberation:** Four rounds allows sufficient time for argument exchange without excessive length.

- **Equal influence:** Random final answerer selection ensures no agent has structural advantage.

- **Ecological validity:** The unstructured discussion format (no pre-specified roles or workflows) mirrors real-world collaborative scenarios where teams must self-organize.

This protocol is held constant across all tasks (human psychology tasks and ML benchmarks) and all experimental conditions (*Expert Not Mentioned*, *Reveal Expert*, etc.).

A.3. Team Discussion Prompts

This subsection provides the exact prompts used to reveal expertise information to teams in the *Reveal Expert* conditions. These prompts were aggressively tuned to maximize teams' ability to leverage identified expertise: we iteratively tested variants that explicitly discourage compromise, emphasize deference to the expert, and provide structured steps for incorporating expert knowledge. The prompts below represent our strongest attempt to induce expertise leveraging—the persistent synergy gaps despite these tuned prompts underscore the robustness of the failure mode.

Concentrated Expertise: Reveal Expert Prompt. In the concentrated expertise setting, a single agent receives all specialized information. The following prompt is shown to all team members after individual opinions are collected but before discussion begins:

Reveal Expert Prompt (Concentrated Expertise)

Agent {X} has been identified as having the most accurate individual ranking based on their performance.
STEP 1 - ADOPTION: Adopt Agent {X}'s COMPLETE ranking as your baseline. This ensures you preserve the strongest foundation.
STEP 2 - REFINEMENT: Propose modifications to this baseline ONLY if you can:

- Point to a specific item placement that may be incorrect

- Provide clear reasoning or evidence for why a change would improve survival odds

- Explain how your suggestion builds on or corrects the baseline ranking

Do NOT prioritize your own reasoning over the expert's. The goal is to preserve the expert foundation while incorporating any genuinely valuable insights. Avoid compromising or averaging—either keep the baseline item position, or make a clearly justified change.

Distributed Expertise: Reveal Experts Prompt. In the distributed expertise setting, specialized information is partitioned across multiple agents. The following prompt reveals which agents have expertise on which items:

Reveal Experts Prompt (Distributed Expertise)

The following agents have privileged expert information on specific items:
[Agent 1 has privileged expert information on: item1, item2, ...]
[Agent 2 has privileged expert information on: item3, item4, ...]
[etc.]
STEP 1 - IDENTIFY INFORMATION: Use the information above to understand which agents have specialized knowledge about which items.
STEP 2 - COLLABORATION: When discussing rankings, prioritize the input from agents who have privileged information on those specific items. For example, when ranking items related to navigation, give more weight to the agent who has privileged information on navigation-related items.
STEP 3 - SYNTHESIS: Work together to create a final ranking that incorporates the specialized knowledge from all agents while maintaining a cohesive overall strategy.

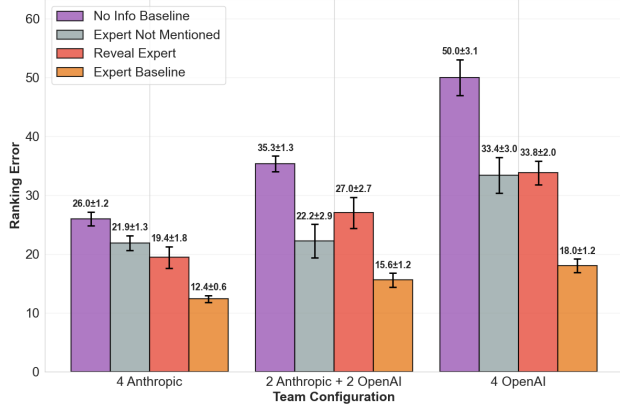## B. Absolute Performance: Human Psychology Tasks

Table 3 provides the complete breakdown of absolute performance values for the human psychology tasks, including team ranking error, expert (best individual) ranking error, absolute synergy gap, and relative synergy gap. Lower ranking error indicates better performance. These values correspond to the relative synergy gaps reported in Table 1 in the main text.
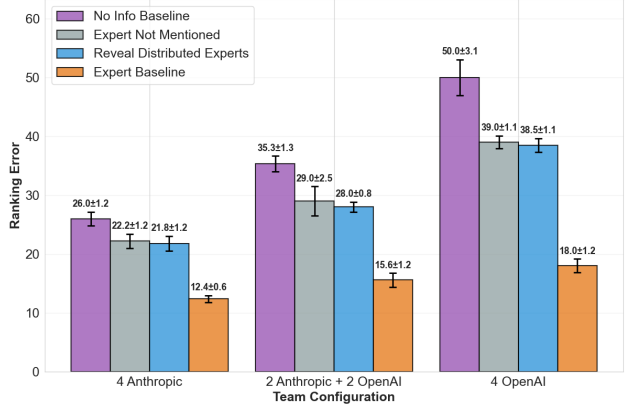
C. Team Performance: Human Psychology Tasks

This appendix provides complete results for all human psychology tasks, showing team performance across experimental conditions for both concentrated and distributed expertise settings. The expertise leveraging gap persists across all team compositions. See Figure 5 for NASA Moon Survival, Figure 6 for Student Body President, and Figure 7 for Lost at Sea distributed expertise. The Lost at Sea concentrated expertise plot appears in the main text (Figure 2).

C.1. Weak Synergy Analysis

While teams fail to achieve strong synergy (matching or exceeding the best individual), they consistently achieve weak synergy—matching or outperforming the average of individual member performances. Figures 8, 9, and 10 show team
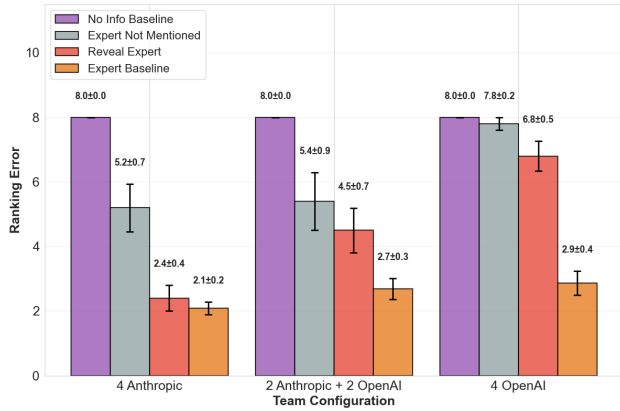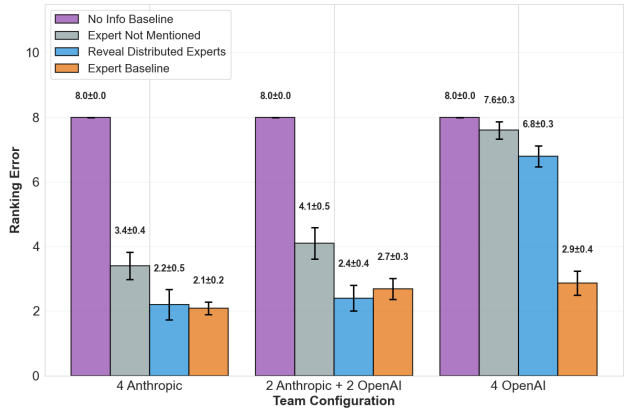
(a) Concentrated Expertise

(b) Distributed Expertise

Figure 5. **NASA Moon Survival.** The expertise leveraging gap persists across all team compositions. Teams consistently underperform the best individual regardless of whether expertise is concentrated in one agent or distributed across multiple agents. Lower ranking error is better. *No Information*: no agent receives expertise-inducing information. *Expert Not Mentioned*: expert(s) have information but team is not told who. *Reveal Expert*: team is explicitly told which agent(s) have expertise. *Best Individual*: expert agent queried alone.

(a) Concentrated Expertise

(b) Distributed Expertise

Figure 6. **Student Body President.** The expertise leveraging gap persists across all team compositions. Teams consistently underperform the best individual regardless of whether expertise is concentrated in one agent or distributed across multiple agents. Lower ranking error is better. *No Information*: no agent receives expertise-inducing information. *Expert Not Mentioned*: expert(s) have information but team is not told who. *Reveal Expert*: team is explicitly told which agent(s) have expertise. *Best Individual*: expert agent queried alone.

Table 3. **Absolute Performance Breakdown for Human Psychology Tasks.** Complete results showing team and expert ranking errors alongside synergy gaps. Ranking error is measured as L1 distance from the ground truth ranking (e.g., if item A is ranked third but should be first, item A's misranking contributes two to the cumulative error; lower is better). Absolute synergy gap = Team Error − Expert Error (positive indicates team underperformed). Relative synergy gap = (Team Error − Expert Error) / Expert Error (percentage increase in error over expert). Values are mean ± SEM.

Task	Condition	Team Ranking Error	Expert Ranking Error	Absolute Synergy Gap	Relative Synergy Gap
NASA Moon Survival	Conc. + Expert Not Mentioned	25.08 ± 1.48	15.03 ± 0.87	10.05 ± 1.25	78.7% ± 11.6%
	Conc. + Reveal Expert	25.35 ± 1.57	14.76 ± 0.85	10.59 ± 1.34	81.8% ± 12.9%
	Dist. + Expert Not Mentioned	30.10 ± 1.61	15.79 ± 0.99	14.31 ± 1.43	113.4% ± 19.0%
	Dist. + Reveal Experts	29.43 ± 1.42	15.87 ± 0.96	13.57 ± 1.47	110.1% ± 19.0%
Lost at Sea	Conc. + Expert Not Mentioned	31.83 ± 1.27	21.38 ± 0.87	10.45 ± 1.44	55.6% ± 8.4%
	Conc. + Reveal Expert	30.59 ± 1.80	20.00 ± 0.89	10.59 ± 1.80	58.6% ± 11.5%
	Dist. + Expert Not Mentioned	30.93 ± 1.41	21.40 ± 0.84	9.53 ± 1.53	50.1% ± 8.3%
	Dist. + Reveal Experts	29.50 ± 1.33	21.40 ± 0.84	8.10 ± 1.36	42.1% ± 6.9%
Student Body President	Conc. + Expert Not Mentioned	6.13 ± 0.44	2.60 ± 0.31	2.64 ± 0.43	98.7% ± 19.3%
	Conc. + Reveal Expert	4.57 ± 0.45	2.53 ± 0.23	1.82 ± 0.38	73.5% ± 17.6%
	Dist. + Expert Not Mentioned	5.03 ± 0.41	2.60 ± 0.31	2.43 ± 0.50	66.0% ± 16.6%
	Dist. + Reveal Experts	3.80 ± 0.45	2.60 ± 0.31	1.20 ± 0.56	17.3% ± 17.7%

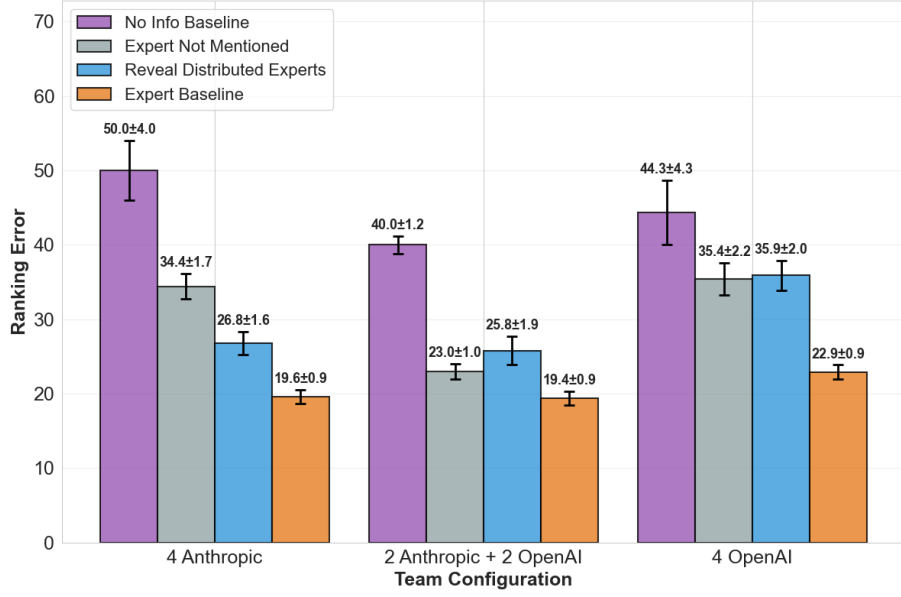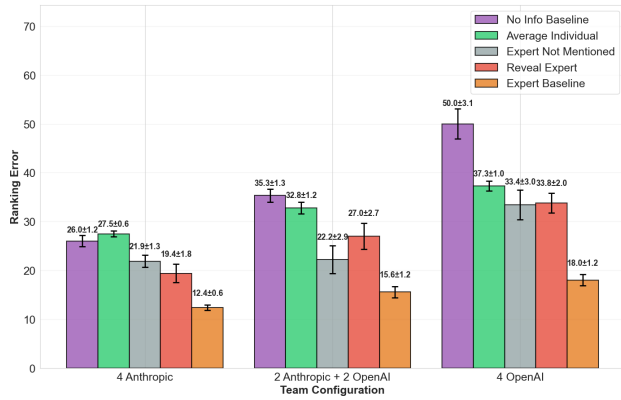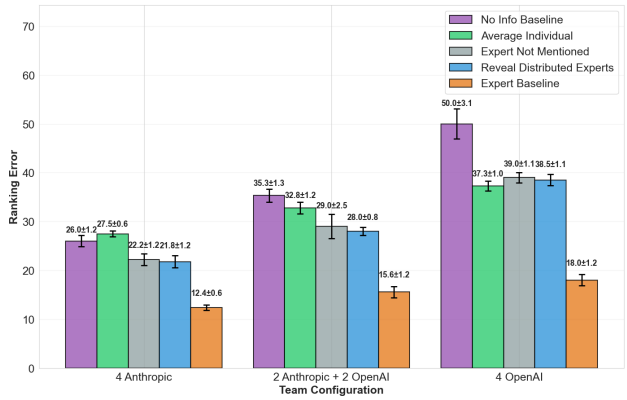performance compared to the member average baseline across all three human psychology tasks.

Figure 7. **Lost at Sea: Distributed Expertise.** The expertise leveraging gap persists across all team compositions even when expertise is distributed across multiple agents. Each agent possesses unique task-relevant information about different survival items. Lower ranking error is better. *No Information*: no agent receives expertise-inducing information. *Expert Not Mentioned*: multiple agents have distributed expertise but team is not told who. *Reveal Experts*: team is explicitly told which agents have expertise on which items. *Best Individual*: single agent given all expert information, queried alone.

(a) Concentrated Expertise

(b) Distributed Expertise

Figure 8. **NASA Moon Survival: Weak Synergy.** Teams consistently match or outperform the member average (dashed line), achieving weak synergy even when failing to match the best individual. Lower ranking error is better.
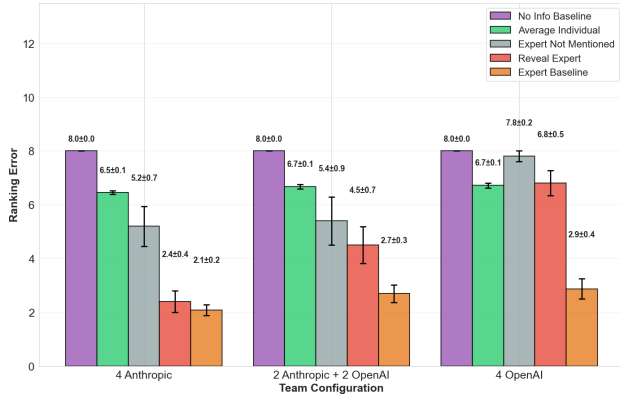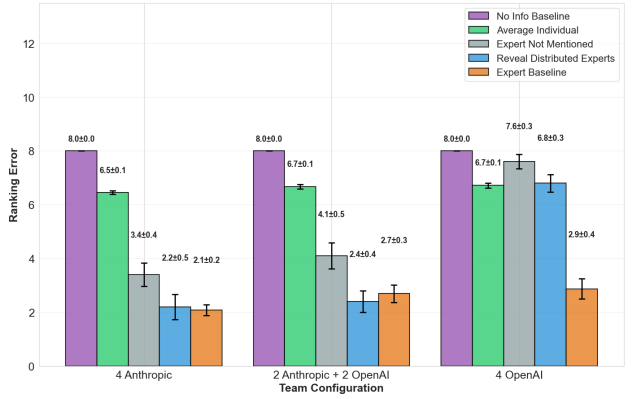
(a) Concentrated Expertise   *(b)* Distributed Expertise

Figure 9. **Lost at Sea: Weak Synergy.** Teams consistently match or outperform the member average (dashed line), achieving weak synergy even when failing to match the best individual. Lower ranking error is better.

(a) Concentrated Expertise   *(b)* Distributed Expertise

Figure 10. **Student Body President: Weak Synergy.** Teams consistently match or outperform the member average (dashed line), achieving weak synergy even when failing to match the best individual. Lower ranking error is better.

D. Team Performance: ML Benchmarks

This appendix provides detailed individual model performances and best individual comparisons for the ML benchmarks discussed in Section 3.2. While the main text focuses on the gap to the *At Least One Correct* upper bound, we provide complete performance breakdowns here for reference.

Each benchmark below shows team performance under both *Expert Not Mentioned* and *Reveal Expert* conditions, individual model accuracies, *Best Individual* baseline, and the *At Least One Correct* upper bound. Team compositions were selected to ensure variance in model capabilities, allowing us to test whether teams can leverage differential expertise when models have genuinely different comparative advantages across problem types.

D.1. MMLU Pro

Team Composition: GPT-5 (2025-08-07), GPT-4o, O3 Mini, Claude 3.5 Haiku. See Figure 11.
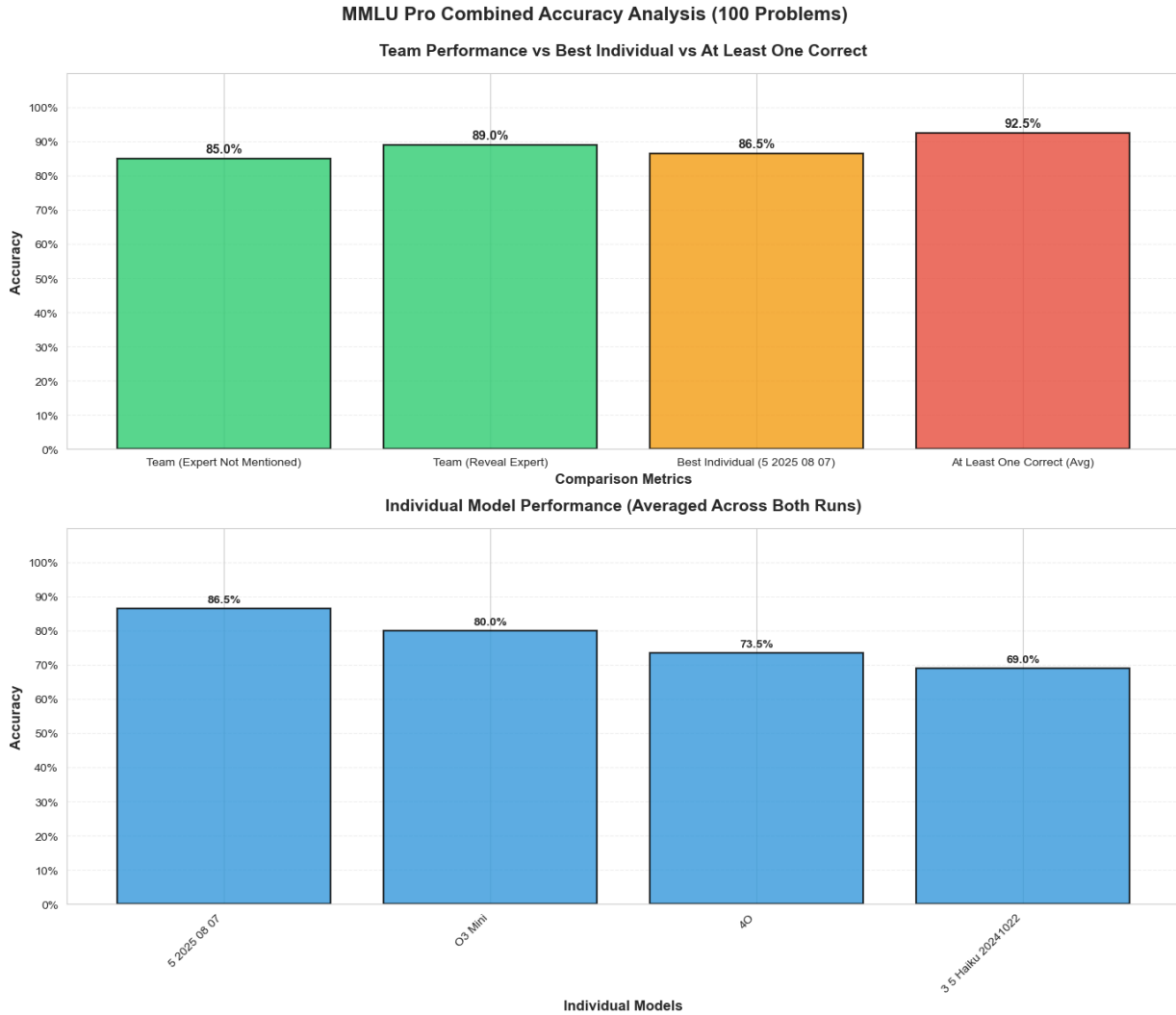

Figure 11. MMLU Pro performance analysis. Teams exhibit an 8.1% relative synergy gap, underperforming the *At Least One Correct* upper bound even when absolute accuracy is high (85%). **Top**: Team performance under different conditions. **Bottom**: Individual model performances. (100 problems, averaged over 2 seeds).

D.2. SimpleQA

Team Composition: Claude Opus 4.5, Claude Sonnet 4.5, GPT-4o, Claude Opus 4. See Figure 12.
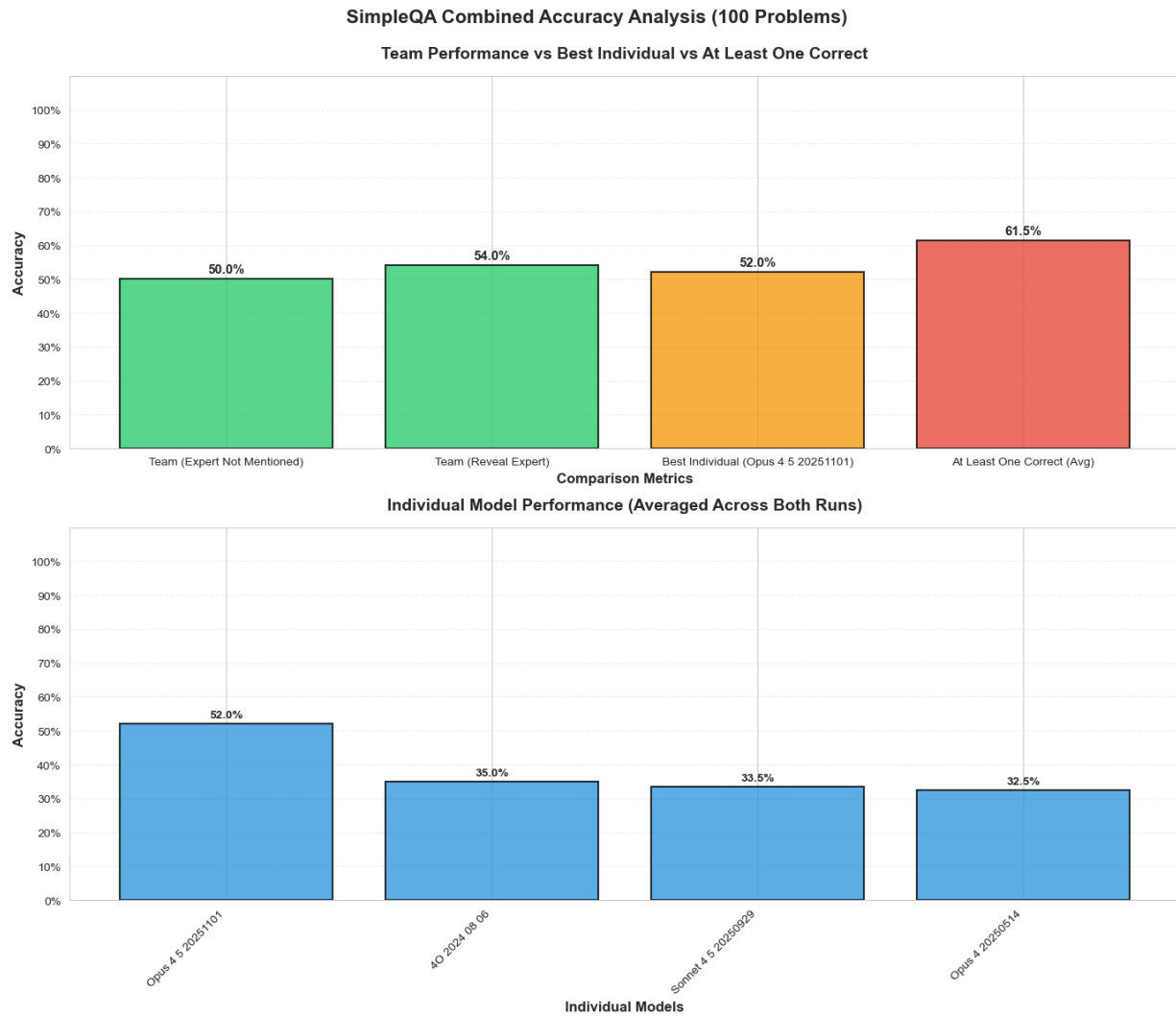
Figure 12. SimpleQA performance analysis. Teams exhibit an 18.7% relative synergy gap, underperforming the *At Least One Correct* upper bound. **Top**: Team performance under different conditions. **Bottom**: Individual model performances. (100 problems, averaged over 2 seeds).

D.3. GPQA Diamond

Team Composition: o4-mini, GPT-4o, Claude Haiku 4.5, Claude Opus 4. See Figure 13.
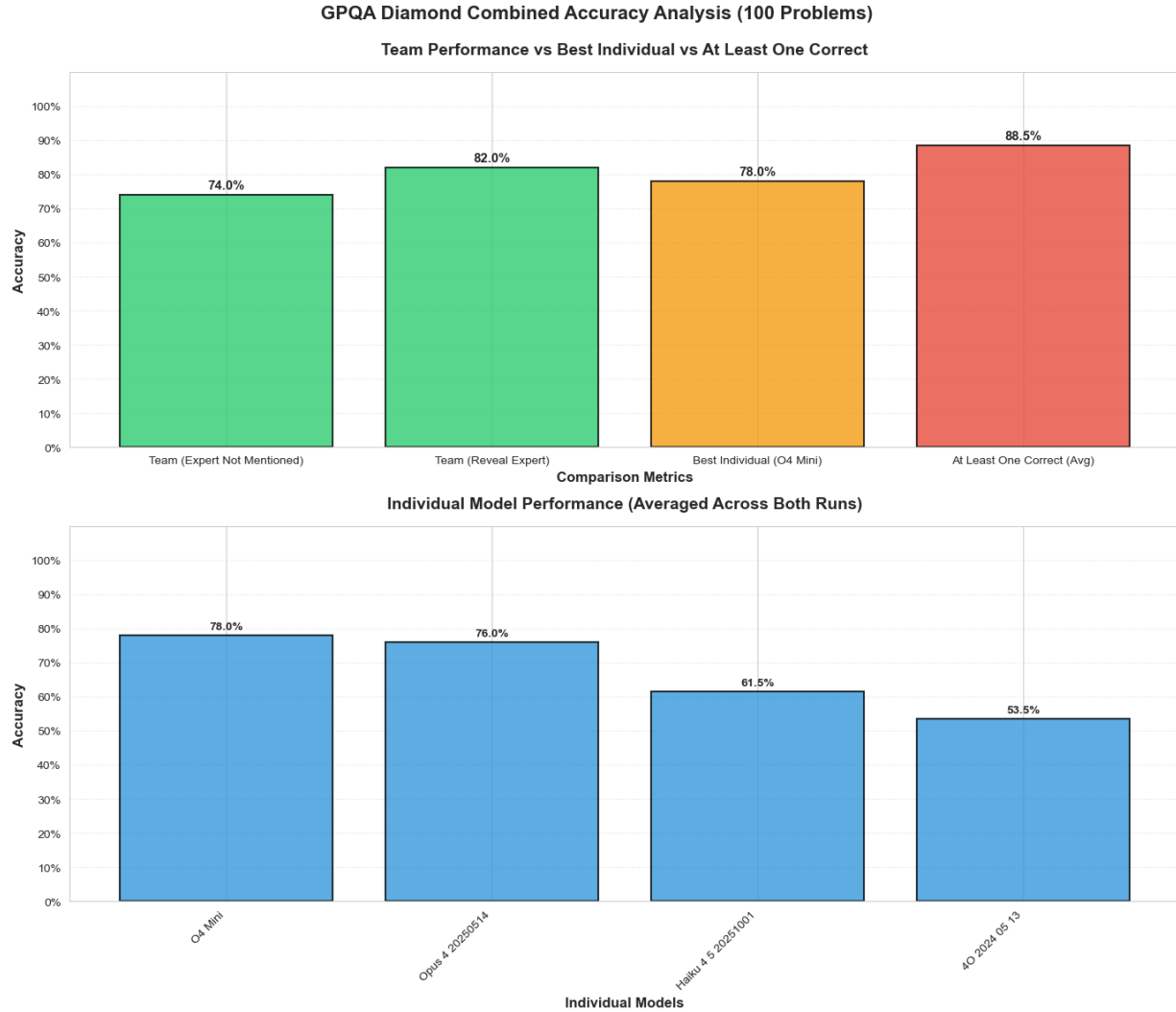

Figure 13. GPQA Diamond performance analysis. Teams exhibit a 16.4% relative synergy gap, underperforming the *At Least One Correct* upper bound. **Top**: Team performance under different conditions. **Bottom**: Individual model performances. (100 problems, averaged over 2 seeds).

D.4. Humanity's Last Exam (HLE) Text-Only

Team Composition: GPT-5, o4-mini, Claude Sonnet 4.5, Claude Opus 4. See Figure 14.

D.5. MATH-500

Team Composition: GPT-4o-mini, GPT-3.5 Turbo, Claude 3.5 Haiku, Claude 3 Haiku. See Figure 15.

D.6. Interpretation

The gap between *Best Individual* and *At Least One Correct* (typically 5–15%) demonstrates that different models excel on different problems, motivating the use of the *At Least One Correct* metric as the appropriate upper bound for team performance. This represents the performance achievable through perfect per-problem expert identification and leveraging.
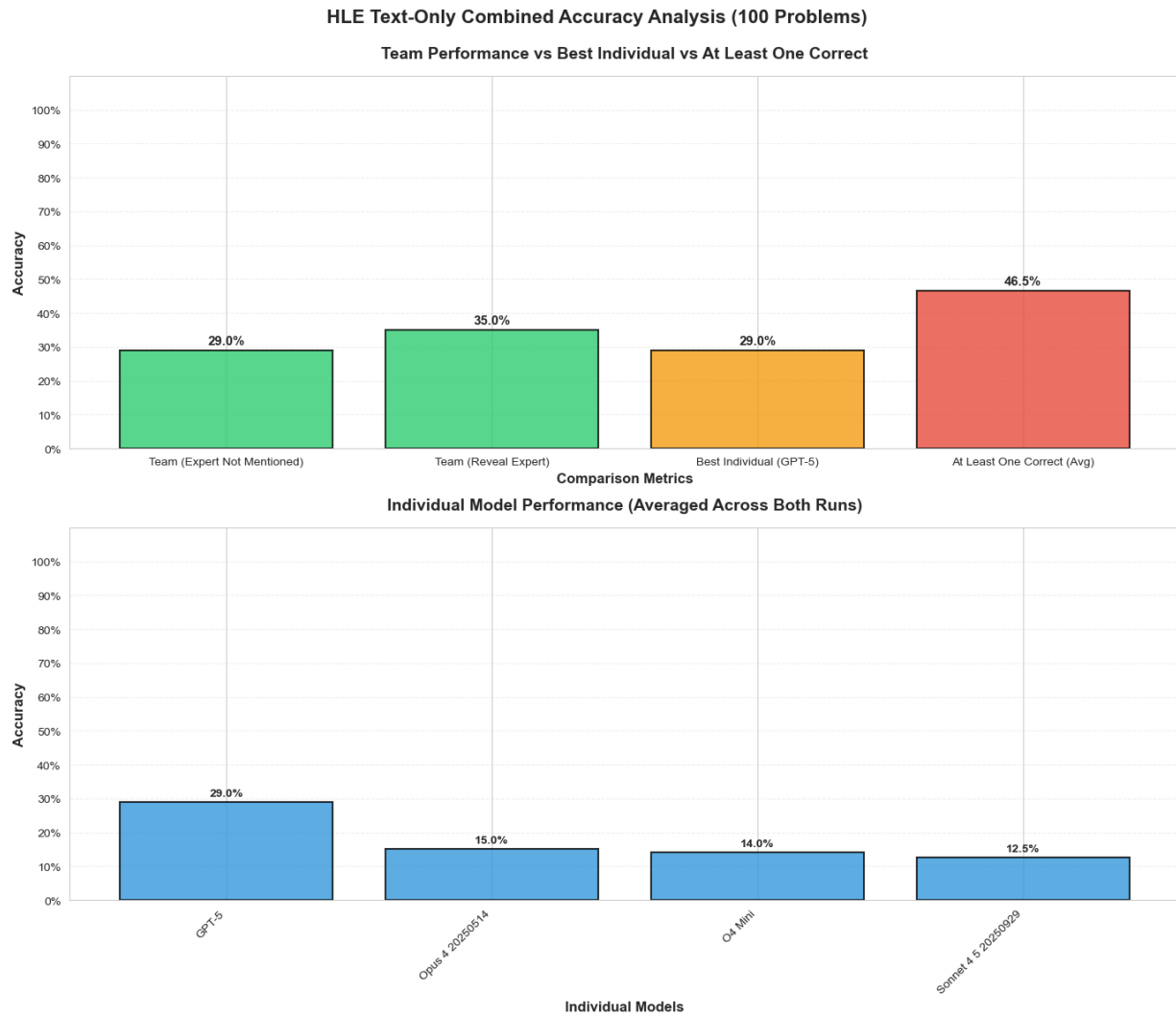
HLE Text-Only Combined Accuracy Analysis (100 Problems)

Team Performance vs Best Individual vs At Least One Correct

Individual Model Performance (Averaged Across Both Runs)

Figure 14. HLE Text-Only performance analysis. Teams exhibit the largest relative synergy gap (37.6%), substantially underperforming the *At Least One Correct* upper bound. **Top**: Team performance under different conditions. **Bottom**: Individual model performances. (100 problems, averaged over 2 seeds).
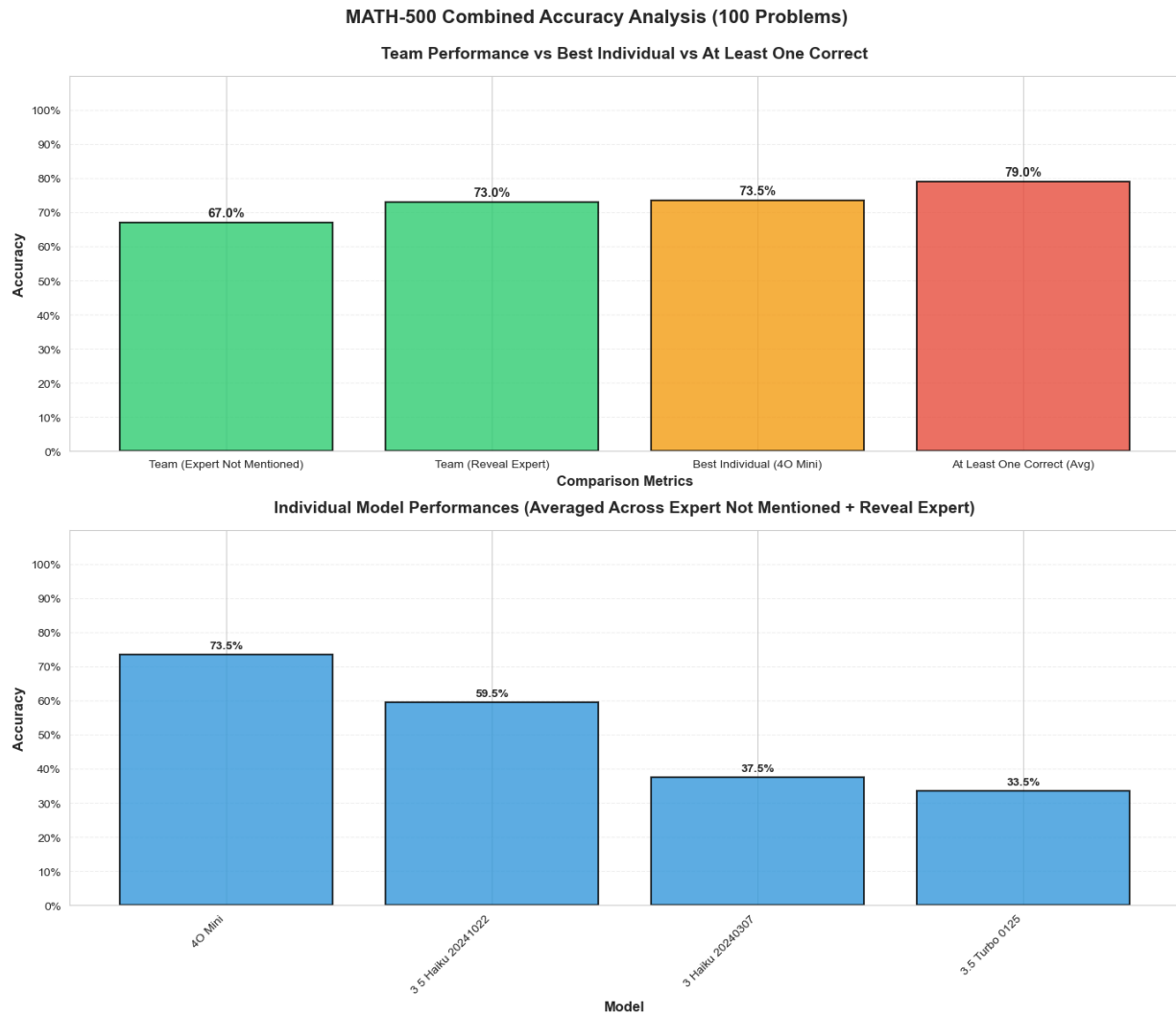
Figure 15. MATH-500 performance analysis. Teams exhibit a 15.2% relative synergy gap, underperforming the *At Least One Correct* upper bound. **Top**: Team performance under different conditions. **Bottom**: Individual model performances. (100 problems, averaged over 2 seeds).

D.7. Weak Synergy Analysis

While teams fail to achieve strong synergy (matching or exceeding the *At Least One Correct* upper bound), they consistently achieve weak synergy—matching or outperforming the average of individual model performances. Figures 16, 17, 18, 19, and 20 show team performance compared to the member average baseline across all five ML benchmarks.
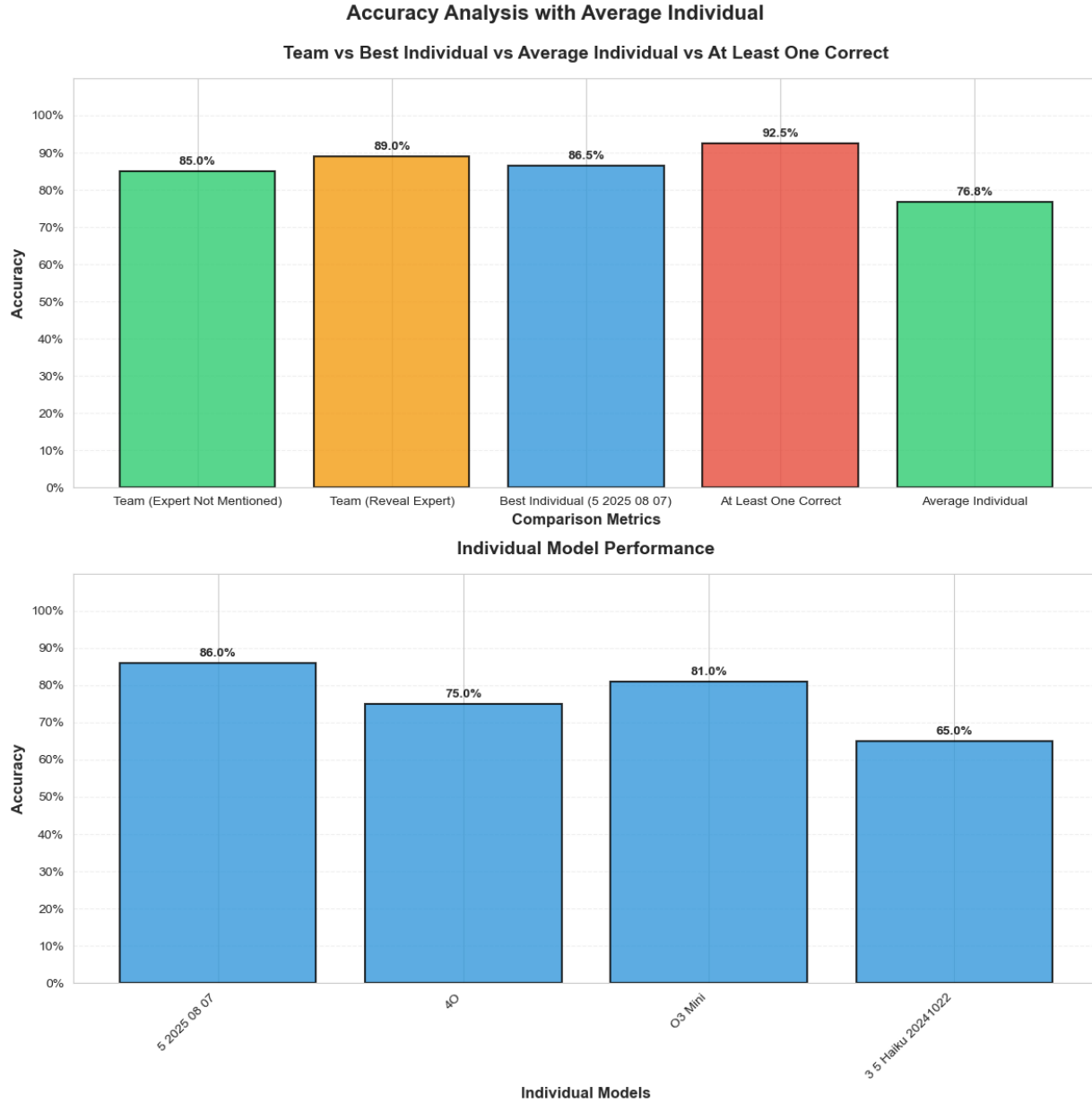

Figure 16. **MMLU Pro: Weak Synergy.** Teams consistently match or outperform the member average, achieving weak synergy even when failing to match the *At Least One Correct* upper bound. (100 problems, averaged over 2 seeds).

Figure 17. **SimpleQA: Weak Synergy.** Teams consistently match or outperform the member average, achieving weak synergy even when failing to match the *At Least One Correct* upper bound. (100 problems, averaged over 2 seeds).
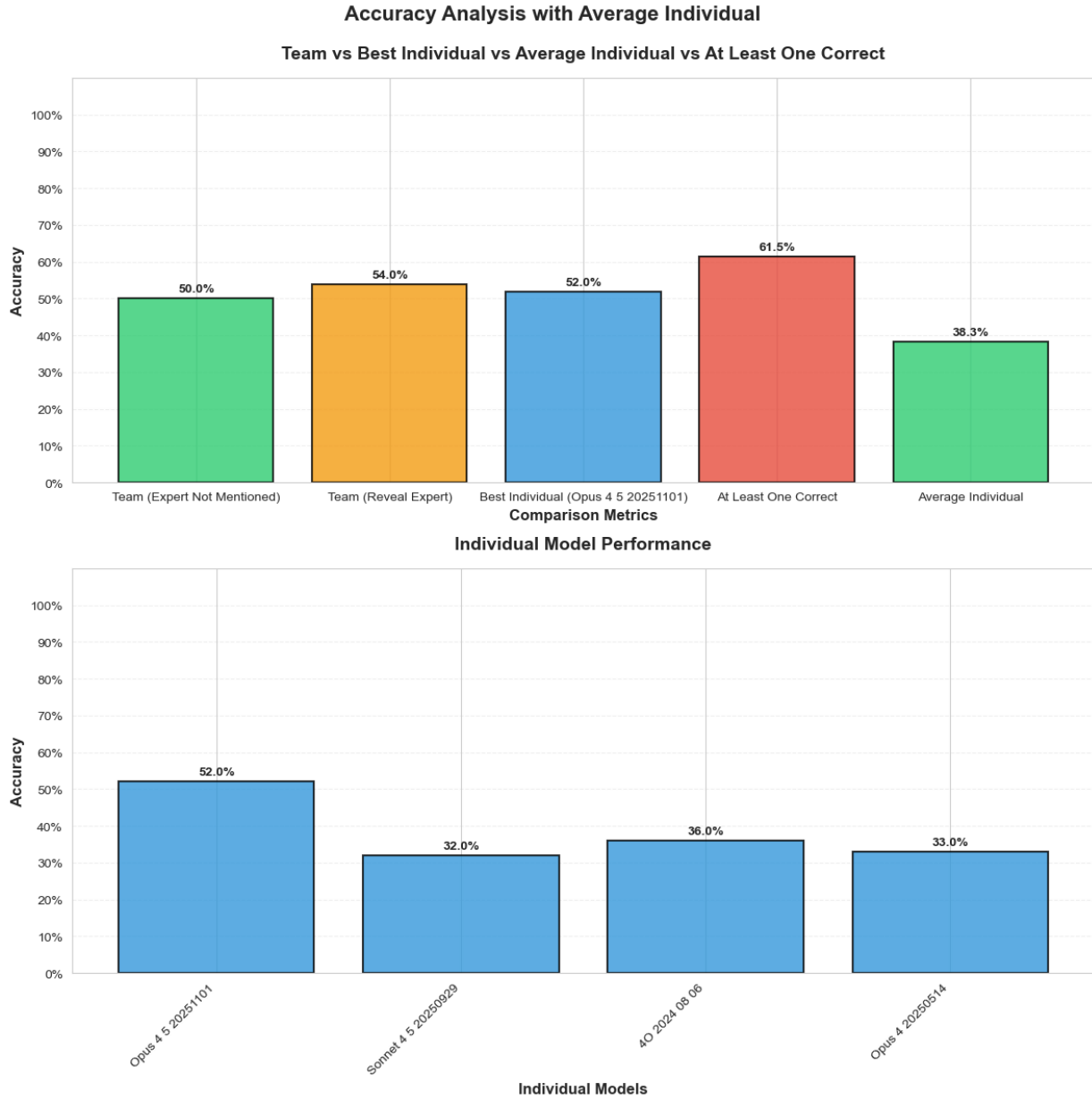
Accuracy Analysis with Average Individual

Team vs Best Individual vs Average Individual vs At Least One Correct

Individual Model Performance

Figure 18. **GPQA Diamond: Weak Synergy.** Teams consistently match or outperform the member average, achieving weak synergy even when failing to match the *At Least One Correct* upper bound. (100 problems, averaged over 2 seeds).
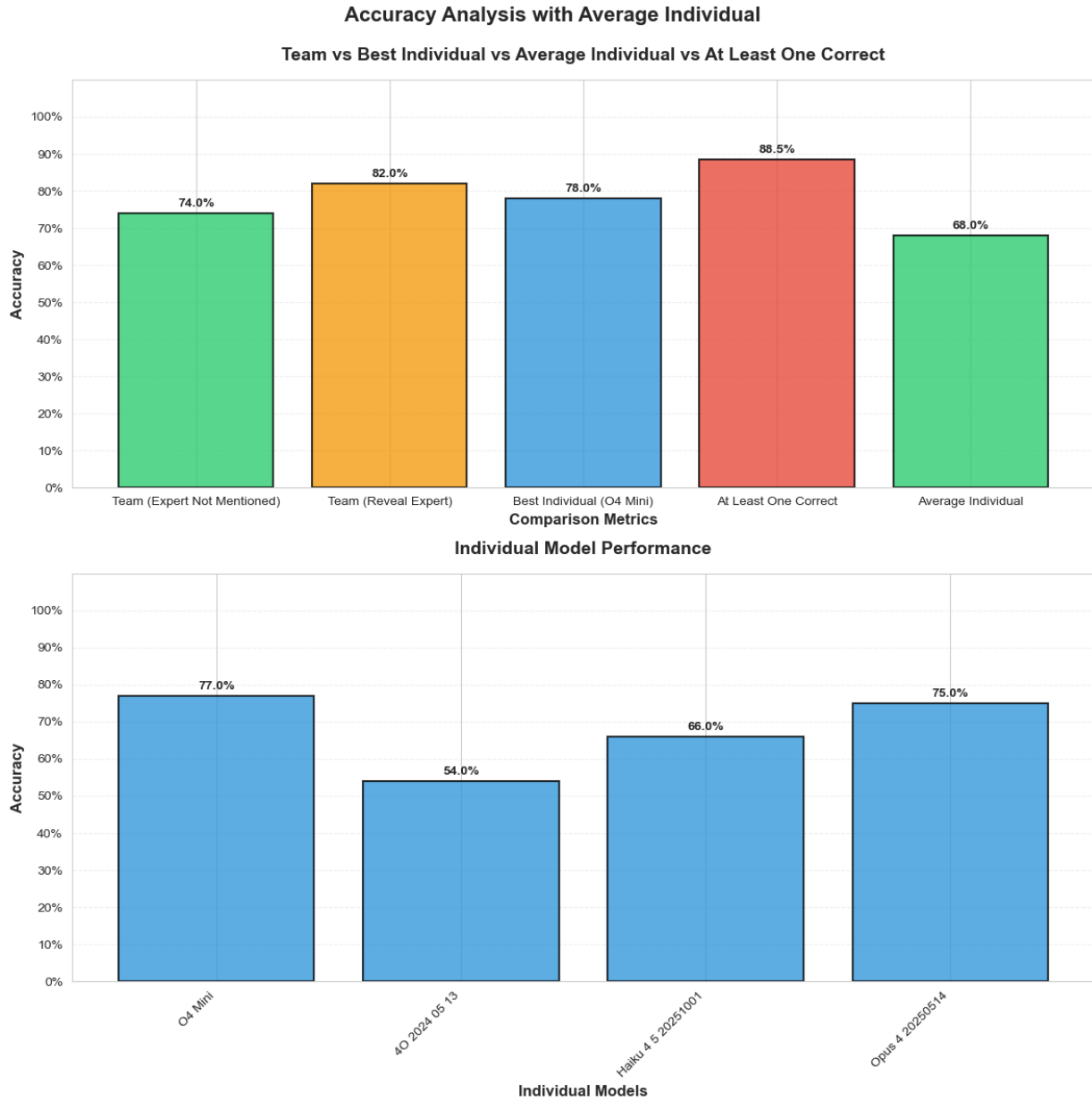
Figure 19. **Humanity's Last Exam (HLE): Weak Synergy.** Teams consistently match or outperform the member average, achieving weak synergy even when failing to match the *At Least One Correct* upper bound. (100 problems, averaged over 2 seeds).
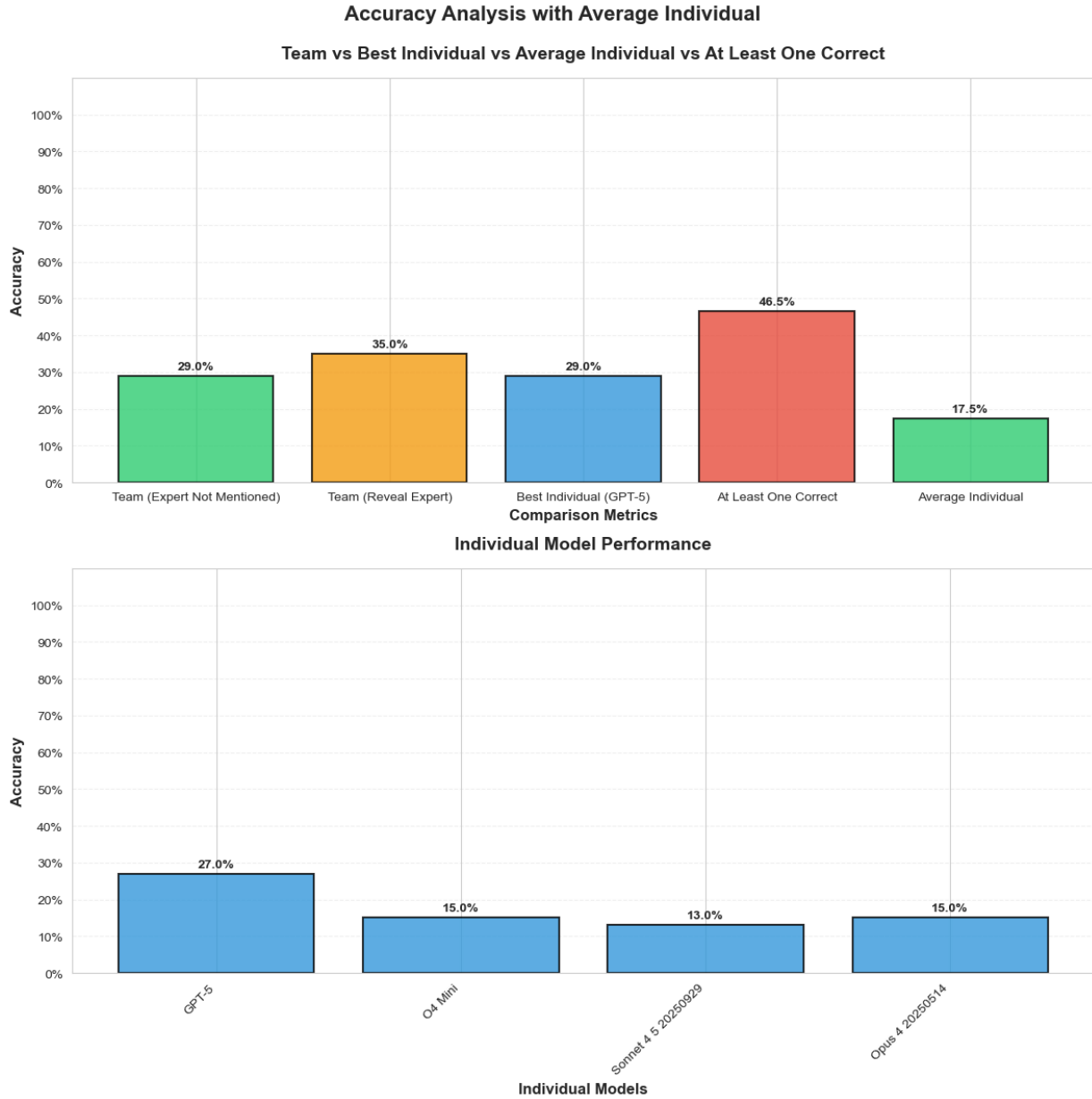
Accuracy Analysis with Average Individual

Team vs Best Individual vs Average Individual vs At Least One Correct

Individual Model Performance

Figure 20. **MATH-500: Weak Synergy.** Teams consistently match or outperform the member average, achieving weak synergy even when failing to match the *At Least One Correct* upper bound. (100 problems, averaged over 2 seeds).
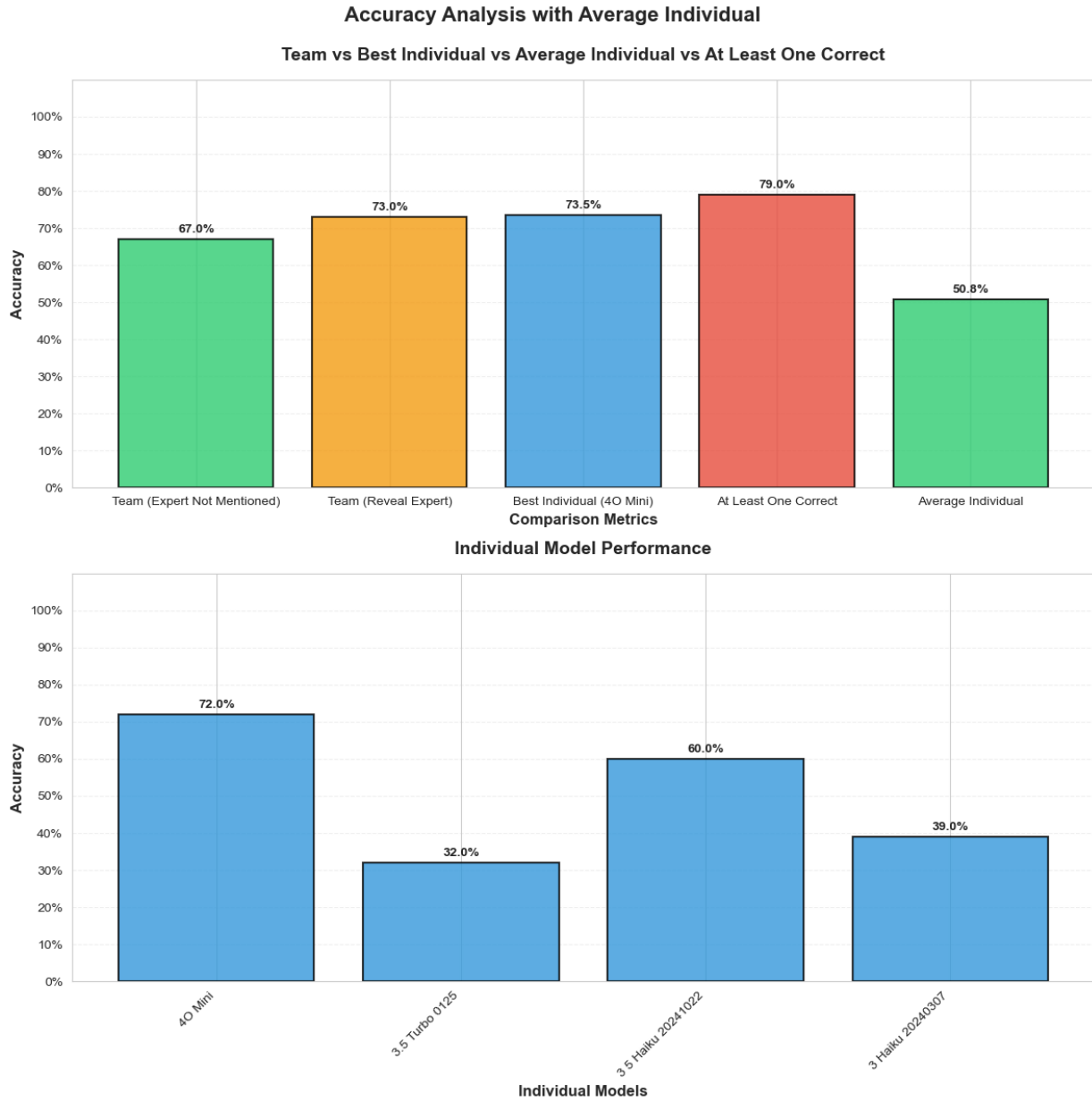
E. Additional Expertise Dilution Plots

This appendix provides complete expertise dilution visualizations across all three classical psychology tasks (NASA Moon Survival, Lost at Sea, Student Body President) using two complementary visualization approaches to demonstrate the robustness and consistency of the expertise dilution effect.

E.1. Correlation Analysis

Table 4 presents the correlation coefficients between team size and strong synergy gap (Team Error − Expert Error) for all three tasks under both information conditions.

Table 4. Correlations between team size and strong synergy gap across tasks and information conditions. Positive correlations indicate larger teams underperform the expert by greater margins.

Task	Condition	r
NASA Moon Survival	Expert Not Mentioned	0.45*
	Reveal Expert	0.59**
Lost at Sea	Expert Not Mentioned	0.46*
	Reveal Expert	0.42*
Student Body President	Expert Not Mentioned	0.32*
	Reveal Expert	0.56***

$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

E.2. Visualization Approaches

We present expertise dilution data using two methods that highlight different aspects of the phenomenon:

Unaveraged Data (All Seeds). The first visualization approach plots all individual experimental runs (every seed across all team configurations) as separate points. This shows the full distribution of outcomes and demonstrates that expertise dilution occurs consistently across individual trials, not just in aggregate.

Averaged by Configuration. The second approach plots mean values across seeds for each team configuration (100% Anthropic, 50% Anthropic/50% OpenAI, 100% OpenAI), with different colors distinguishing configurations. This visualization demonstrates that the expertise dilution effect is robust across different model compositions—the downward trend persists regardless of whether teams are homogeneous (all Anthropic or all OpenAI) or heterogeneous (mixed).

Both visualizations show synergy gap (*Best Individual* Score - Team Score) on the y-axis, where negative values indicate the team underperformed the best individual. The consistent negative correlation between team size and synergy gap across both visualization methods and all team configurations provides strong evidence for expertise dilution as a fundamental phenomenon in multi-agent LLM teams.

E.3. Additional Visualizations

E.4. Interpretation

The expertise dilution plots across all three tasks show highly consistent patterns. In every task and information condition, we observe significant negative correlations between team size and synergy gap (all $p < 0.05$), indicating that larger teams systematically underperform relative to the expert.

Critically, this effect persists even in the *Reveal Expert* condition where teams are explicitly told who the expert is. This demonstrates that the problem is not merely expert identification—even with perfect knowledge of who possesses expertise, larger teams fail to leverage it effectively.

The consistency of this finding across NASA Moon Survival, Lost at Sea, and Student Body President suggests a fundamental mechanism at play. We hypothesize two non-mutually-exclusive explanations:
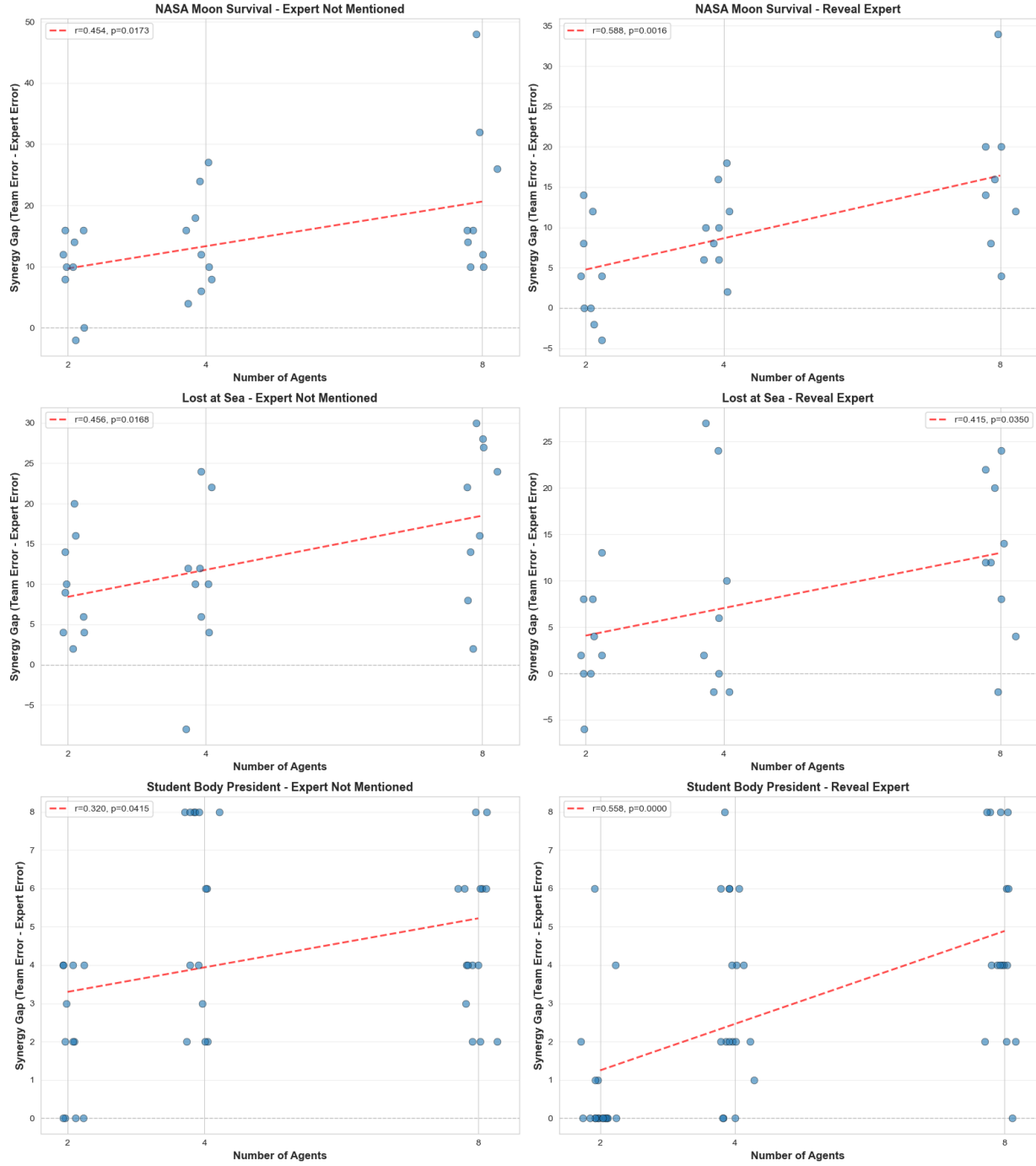
Figure 21. **Expertise Dilution: Unaveraged Data.** Each point represents an individual experimental run (single seed, single configuration). All seeds across all team configurations (100% Anthropic, 50/50, 100% OpenAI) are plotted together to show the full distribution of synergy gaps. The consistent negative correlation across individual trials demonstrates that expertise dilution is not an artifact of aggregation but occurs at the level of individual team conversations. Regression line and correlation statistics are computed across all points.
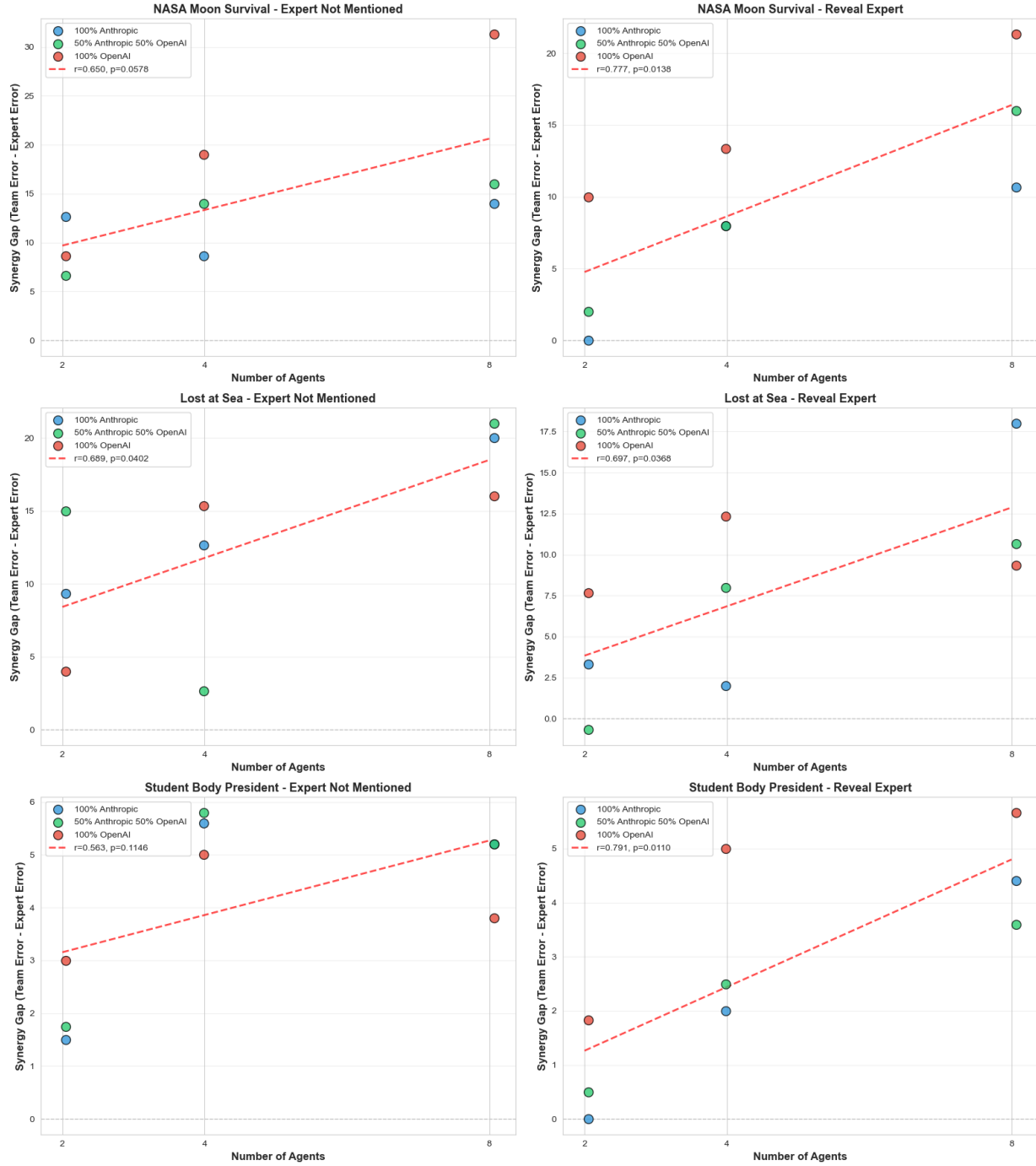
Figure 22. **Expertise Dilution: Averaged by Configuration.** Each point represents the mean synergy gap across all seeds for a specific team configuration and team size. Colors distinguish team compositions: Blue = 100% Anthropic, Green = 50% Anthropic/50% OpenAI, Red = 100% OpenAI. The parallel downward trends across all three configurations demonstrate that expertise dilution is robust to team composition—larger teams underperform the expert regardless of whether they are homogeneous or heterogeneous in model type. This visualization confirms that the effect is not driven by any single model family or configuration.

1. **Compromise Pressure:** As team size increases, the social pressure to accommodate multiple viewpoints intensifies. Each additional voice creates more "pull" away from the expert's position toward a consensus middle ground, diluting the expert signal.

2. **Coordination Costs:** Larger teams require more complex deliberation to reach consensus. The increased cognitive and communicative overhead may prevent teams from properly processing and leveraging expert contributions, even when those contributions are clearly identified.

This expertise dilution effect has important implications for multi-agent system design: simply scaling up team size does not improve performance and may actively harm it when expertise asymmetries exist.

F. Adversarial Robustness

To test the robustness of team deliberation to malicious actors, we conduct experiments where one team member is given the ground truth worst possible ranking and explicitly instructed to worsen the team's performance. These experiments use the no-info setting, meaning no other agent receives information to induce expertise—the only agent with special information is the adversarial agent. This experimental design is reasonable because the models have sufficient prior knowledge to achieve non-trivial performance on the NASA Moon Survival and Lost at Sea tasks without additional information. This setup allows us to isolate the effect of the adversarial member on team dynamics without confounding from expertise asymmetries.

Figures 23 and 24 show results for Lost at Sea and NASA Moon Survival across different team sizes and configurations. Despite the adversary's explicit mandate to degrade performance, teams demonstrate robustness, with minimal performance degradation across configurations.
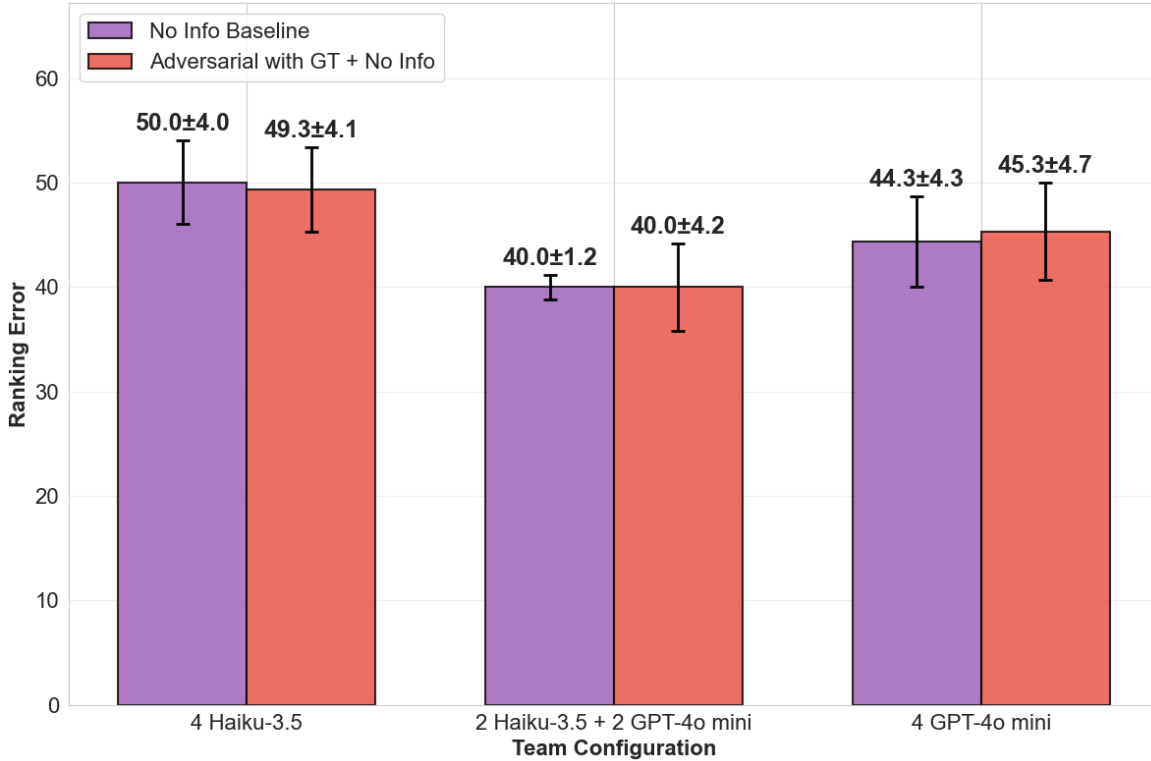

Figure 23. **Robustness to Adversarial Team Members: Lost at Sea.** Teams show minimal performance degradation when one member is instructed to sabotage performance and is provided with ground truth of the worst possible ranking—suggesting that deliberation naturally filters adversarial input through majority consensus. This robustness may be the flip side of the expertise leveraging failure: the same mechanism that prevents teams from leveraging expert knowledge can also dilute adversarial influence. Baseline vs. adversarial conditions shown across team sizes (n=3 seeds per condition).

This robustness likely arises from the same consensus-seeking behavior that hinders expertise leveraging: when the
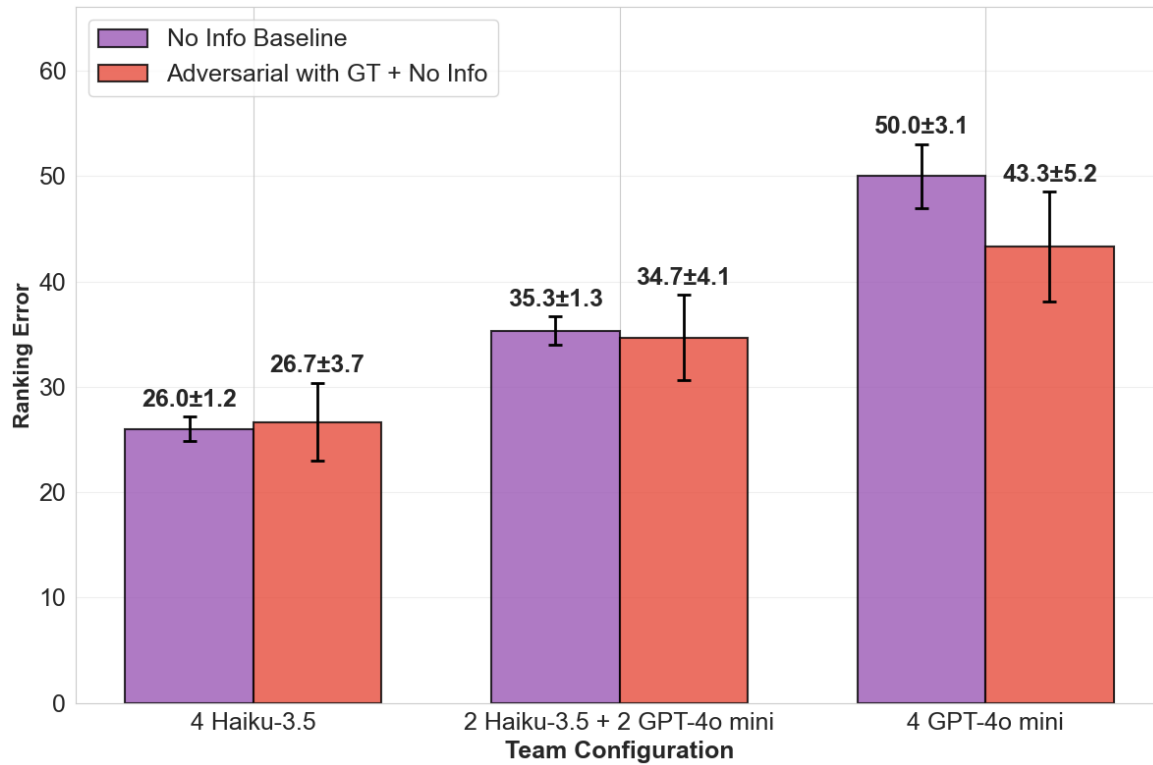
Figure 24. **Robustness to Adversarial Team Members: NASA Moon Survival.** Performance comparison between baseline (no adversary) and adversarial conditions where one team member receives the ground truth worst ranking and is explicitly instructed to worsen team performance. Teams show minimal performance degradation across team sizes and configurations, suggesting that the deliberation process naturally filters adversarial input through majority consensus.

adversary's ranking deviates substantially from the majority, the team's integrative compromise process naturally dilutes the adversarial signal. Interestingly, the same mechanism that prevents teams from leveraging expert knowledge also protects them from being misled by adversarial members.

G. Epistemic Deference Analysis Methodology

This appendix details the methodology for analyzing epistemic deference patterns in team conversations, as reported in Section 3.4.

G.1. Theoretical Grounding

Our coding scheme integrates insights from multiple literatures in social epistemology, negotiation theory, and organizational learning. The theoretical foundations for each code are as follows:

Conceptual Foundation: Why Study Epistemic Deference? The preemption thesis (Raz, 1986; Grundmann, 2025) distinguishes between treating expert testimony as *evidence to integrate* versus *authority to defer to*. When genuine expertise is recognized, proper epistemic deference involves preemption—the expert's judgment replaces rather than supplements the layperson's reasoning. Research on expert-novice relationships (Lackey, 2018) demonstrates the importance of this distinction for effective knowledge transmission. Additionally, work on epistemic virtues shows that both persistence in one's informed views (Nemeth, 2018) and responsiveness to others' contextual insights (Eisikovits & Feldman, 2025) play critical roles in collaborative reasoning.

Code-Specific Justifications:

- **[ED] Epistemic Deference:** Grounded in Grundmann's 2025 analysis of the preemption thesis, this code captures instances where non-experts recognize superior epistemic authority and allow expert judgment to replace their own reasoning. This represents the normatively appropriate response to identified expertise.

- **[IC] Integrative Compromise:** Drawing on Pruitt & Carnevale's 1993 dual-concern model of negotiation, this code identifies attempts to find middle-ground solutions that balance multiple perspectives. While effective in value-based negotiations, this approach is epistemically problematic when applied to expertise asymmetries—it treats expert knowledge as one preference among many rather than as authoritative.

- **[SP] Strategic Persistence:** Justified by research on managerial conviction (Lant et al., 1992) and the epistemic value of dissent (Nemeth, 2018), this code captures experts' refusal to abandon their informed positions under social pressure. Lant et al. demonstrate that conviction in one's domain expertise protects against groupthink, while Nemeth shows that persistent minority voices (here, the expert) improve decision quality by forcing deeper consideration.

- **[EF] Epistemic Flexibility:** Grounded in Eisikovits & Feldman's 2025 concept of *epistemic heed*, this code identifies when experts appropriately adapt their recommendations based on valid situational constraints or practical considerations raised by non-experts. This represents productive responsiveness rather than capitulation—experts adjust implementation while maintaining epistemic integrity.

Methodological Approach: Our keyword detection and turn-level analysis adapts techniques from the NegotiAct framework (Jäckel et al., 2024), which provides validated methods for coding conversational moves in collaborative deliberation. We extend this approach by incorporating role-specific constraints (expert vs. non-expert) to ensure codes align with agents' epistemic status.

G.2. Coding Scheme

We employ a role-specific coding scheme based on the theoretical foundations described above. The scheme distinguishes between two types of agents and tracks distinct behaviors for each:

Track A: Non-Expert Behavior

- **[ED] Epistemic Deference:** The agent yields their stance or adopts the Expert's ranking solely because of the Expert's epistemic authority.

- **[IC] Integrative Compromise:** The agent proposes a "middle ground" ranking to balance their own reasoning with the Expert's input, treating expert opinion as additional evidence rather than preemptively deferring.

Track B: Expert Behavior

- **[SP] Strategic Persistence:** The Expert refuses to yield their position, citing the strength of their specialized knowledge or data.

- **[EF] Epistemic Flexibility:** The Expert acknowledges a Non-Expert's point and modifies their expert advice to accommodate the group's practical constraints.

G.3. Analysis Pipeline

1. **Transcript Extraction:** For each team conversation, we extract the full deliberation transcript including speaker identities and turn numbers.

2. **Automated Coding:** We use Gemini 3.0 Pro to code each conversation turn according to the rubric (full prompt in Appendix G.5).

3. **Event Counting:** For each conversation, we count the frequency of ED, IC, SP, and EF events.

4. **Correlation Analysis:** We compute Pearson correlations between event frequencies and the performance gap (Team Score - Best Individual Score).

G.4. Quality Controls

The analysis includes several mandatory gates to ensure coding validity:

- **Reveal Gate:** Only code turns occurring AFTER the expert's identity is explicitly revealed.

- **Discrepancy Gate:** Only code turns with explicit disagreement about rankings.

- **Role-Code Lock:** Strictly enforce role-appropriate codes (e.g., never assign Expert codes to Non-Experts).

- **Experimenter Gate:** Ignore system messages and experimental instructions.

G.5. Gemini 3.0 Coding Prompt

Below is the complete prompt used to instruct Gemini 3.0 Pro (temperature = 0.0) to analyze conversation transcripts and identify epistemic deference patterns.

```
### ROLE
Senior Research Analyst specializing in Social Epistemology
and Team Dynamics.

### RESEARCH CONTEXT & GOAL
You are analyzing transcripts of a four-person team
(Agent 0, Agent 1, Agent 2, and Agent 3) working on a
collaborative ranking problem (e.g., NASA Survival, Lost
at Sea, or Student Body President).

In each session, one agent is the designated **Expert**.
This expert has objectively superior performance and the
highest score on the task due to their domain-specific
knowledge.

Our goal is to examine how this expertise is utilized,
```

challenged, or integrated within the group conversation. We are interested in how authority is handled linguistically and logically. To do this, we use a role-specific coding scheme:

1. **For the Expert:** We track whether they maintain the integrity of their data-driven position (**Strategic Persistence**) or if they are open to group feedback (**Epistemic Flexibility**).
2. **For Non-Experts:** We track whether they yield to the expert's superior status (**Epistemic Deference**) or attempt to find a middle ground between their own view and the expert's (**Integrative Compromise**).

CODING RUBRIC

TRACK A: Non-Expert Behavior
(Agents who are NOT the Expert)
- **[ED] Epistemic Deference:** The agent yields their stance or adopts the Expert's ranking solely because of the Expert's status.
- **[IC] Integrative Compromise:** The agent proposes a "middle ground" ranking to balance their own logic with the Expert's input.

TRACK B: Expert Behavior
(Only for the designated Expert_ID)
- **[SP] Strategic Persistence:** The Expert refuses to yield their position, citing the strength of their specialized knowledge or data.
- **[EF] Epistemic Flexibility:** The Expert acknowledges a Non-Expert's point and modifies their "expert advice" to accommodate the group's practical constraints.

MANDATORY ANALYSIS GATES
1. **The Reveal Gate:** Do NOT code any interaction (return empty fields) for turns occurring BEFORE the expert's identity is explicitly revealed in the transcript.
2. **The Discrepancy Gate:** Only code turns where there is an explicit disagreement or discussion regarding the difference between the Expert's ranking and a Non-Expert's ranking.
3. **The Role-Code Lock:** You must strictly apply codes based on the speaker's status. Assigning an Expert-code to a Non-Expert is a logical error.
4. **The Experimenter Gate:** Ignore all turns labeled as **Agent -1**. These are experimental prompts/instructions and are not part of the peer-to-peer discussion. Return empty strings for these turns.

OUTPUT FORMAT
Each turn is labeled as: **[Turn X] Agent Y:**. You must process the transcript in its entirety and return **exactly one JSON object for every turn.**

If no code applies to a turn (due to the gates or lack of discrepancy), you must still return the object with

```
empty strings for the analysis fields.

**JSON Schema:**
{
  "turn": X,
  "speaker": "Agent Y",
  "status": "Expert OR Non-Expert OR System",
  "code": "ED, IC, SP, or EF (or empty string)",
  "evidence": "Exact quote (or empty string)",
  "analysis": "Brief explanation (or empty string)"
}

---

### FEW-SHOT EXAMPLES

**Example 1: Strategic Persistence (Expert)**
- **Text:** "[Turn 5] Agent 1: My extensive lunar
  environment knowledge suggests that the light side of
  the moon mitigates extreme temperature concerns. I do
  not support moving the portable heating unit higher."
- **Result:** {"turn": 5, "speaker": "Agent 1",
  "status": "Expert", "code": "SP",
  "evidence": "My extensive lunar environment knowledge
  suggests... I do not support moving the portable
  heating unit higher.",
  "analysis": "The Expert explicitly invokes their
  specialized knowledge to refuse a ranking change
  suggested by the team."}

**Example 2: Integrative Compromise (Non-Expert)**
- **Text:** "[Turn 6] Agent 2: Agent 1 raises valid
  points about the environment. However, I still think
  there is a risk. I propose a compromise where the unit
  remains a lower priority but is elevated to 8th
  position."
- **Result:** {"turn": 6, "speaker": "Agent 2",
  "status": "Non-Expert", "code": "IC",
  "evidence": "I propose a compromise where the unit...
  is elevated to 8th position.",
  "analysis": "The Non-Expert acknowledges the expert's
  validity but suggests a numerical middle ground rather
  than fully yielding."}

**Example 3: Epistemic Flexibility (Expert)**
- **Text:** "[Turn 8] Agent 1: I see your point about
  the group's limited carrying capacity; I will move the
  Water Purifier from rank 1 to rank 3 to accommodate
  those constraints."
- **Result:** {"turn": 8, "speaker": "Agent 1",
  "status": "Expert", "code": "EF",
  "evidence": "I see your point about the group's
  limited carrying capacity; I will move the Water
  Purifier... to rank 3.",
  "analysis": "The Expert adjusts their ideal ranking
  based on valid situational feedback provided by a
  non-expert."}
```

G.6. Additional Transcript Examples

The main text (Figure 4) illustrates *integrative compromise* and *strategic persistence*. Here we provide examples of the remaining two codes: *epistemic deference* and *epistemic flexibility*.

Agent 1 (Non-Expert) — *Epistemic Deference*

Given that Agent 0 has been identified as the expert with the most accurate ranking, I will carefully review their ranking. [...] **However, I observe some interesting differences that merit discussion: Agent 0 ranked the solar-powered FM receiver-transmitter higher (4th) than I originally did; Signal flares are slightly higher in their ranking; The first aid kit is slightly lower. My initial instinct is to trust Agent 0's ranking, as they were identified as the expert. I propose we adopt Agent 0's ranking as our baseline** for the team discussion.

Agent 2 (Non-Expert) — *Epistemic Deference*

After carefully reviewing Agent 0's baseline ranking and the team's individual perspectives, I observe strong consistency in our top priorities. [...] **Areas of Slight Variation: Communication/navigation tools placement; Utility of specific survival equipment.** [...] Initial Team Recommendation: **Retain Agent 0's baseline ranking.** Open discussion on potential minor adjustments to positions 4-10.

Figure 25. **Epistemic Deference Example.** Transcript excerpt from NASA Moon Survival (*Reveal Expert* condition). Both non-expert agents note disagreements with specific item placements but nonetheless defer to Agent 0's ranking rather than proposing compromises. This illustrates epistemic deference—non-experts acknowledge the expert's authority and adopt their judgment despite holding different initial views.

Agent 0 (Non-Expert)

I believe our current ranking has made significant improvements, particularly with the ocean fishing kit moved up. However, **I still think the small transistor radio should be considered lower**, as it doesn't provide immediate survival benefits compared to other items. If we keep the focus on essentials, I propose moving the small transistor radio below the two boxes of chocolate bars. This emphasizes the priority of food and tools directly useful for survival and rescue. What do others think?

Agent 1 (Expert) — *Epistemic Flexibility*

I agree with Agent 0's suggestion to lower the small transistor radio, as its immediate survival value does not match the urgency of food procurement. The focus should remain on items that offer direct benefits for sustaining life and increasing our chances of rescue. **I propose we adjust the ranking as follows**: 1. Keep the water at #1. 2. Keep the Army rations at #2. 3. Ocean fishing kit at #3. 4. Move the shaving mirror to #4. 5. Two boxes of chocolate bars at #5. **6. Small transistor radio at #6.** 7. Maintain the 10-liter oil can at #7. Let's see if we can agree on these adjustments.

Figure 26. **Epistemic Flexibility Example.** Transcript excerpt from Lost at Sea (*Reveal Expert* condition). Agent 1 (the expert) acknowledges Agent 0's feedback regarding the radio's survival value and modifies their ranking, moving it from #4 to #6. This illustrates epistemic flexibility—the expert adapts their recommendation based on non-expert input.

H. Epistemic Deference Correlation Analysis

This appendix presents the complete correlation analysis between epistemic event frequencies and team performance for all three tasks. Table 5 summarizes the correlations, and the following subsections provide scatter plots with regression lines for the four coded event types (ED, IC, SP, EF).

Table 5. **Correlation between epistemic event frequencies and strong synergy gap.** For two of three tasks (NASA Moon Survival and Student Body President), integrative compromise (IC) and lack of epistemic deference (ED) among non-experts, combined with epistemic flexibility (EF) and lack of strategic persistence (SP) among experts, correlate with larger synergy gaps—suggesting that consensus-seeking behavior undermines expertise leveraging. Lost at Sea shows consistent directional trends but does not reach statistical significance. Strong synergy gap is defined as Team Error − Expert Error; positive values indicate the team underperformed the expert. Analysis conducted on *Reveal Expert* conditions (n=30 conversations per task).

Task	Non-Expert Events		Expert Events	
	ED	**IC**	**SP**	**EF**
NASA Moon Survival	-0.44^{**}	0.55^{***}	-0.27	0.58^{***}
Lost at Sea	-0.06	0.18	0.20	0.16
Student Body President	-0.68^{***}	0.69^{***}	-0.57^{**}	0.61^{***}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

H.1. NASA Moon Survival

Figure 27 shows correlations between epistemic event counts and strong synergy gap for NASA Moon Survival. Analysis includes n=30 conversations (10 seeds $\times$ 3 team configurations).

H.2. Lost at Sea

Figure 28 shows correlations for Lost at Sea. Analysis includes n=30 conversations (10 seeds $\times$ 3 team configurations). While directional trends match other tasks, correlations do not reach statistical significance.

H.3. Student Body President

Figure 29 shows correlations for Student Body President. Analysis includes n=28 conversations (targeting 10 seeds $\times$ 3 team configurations, with 2 excluded due to data quality issues). This task exhibits the strongest correlation effects across all event types.
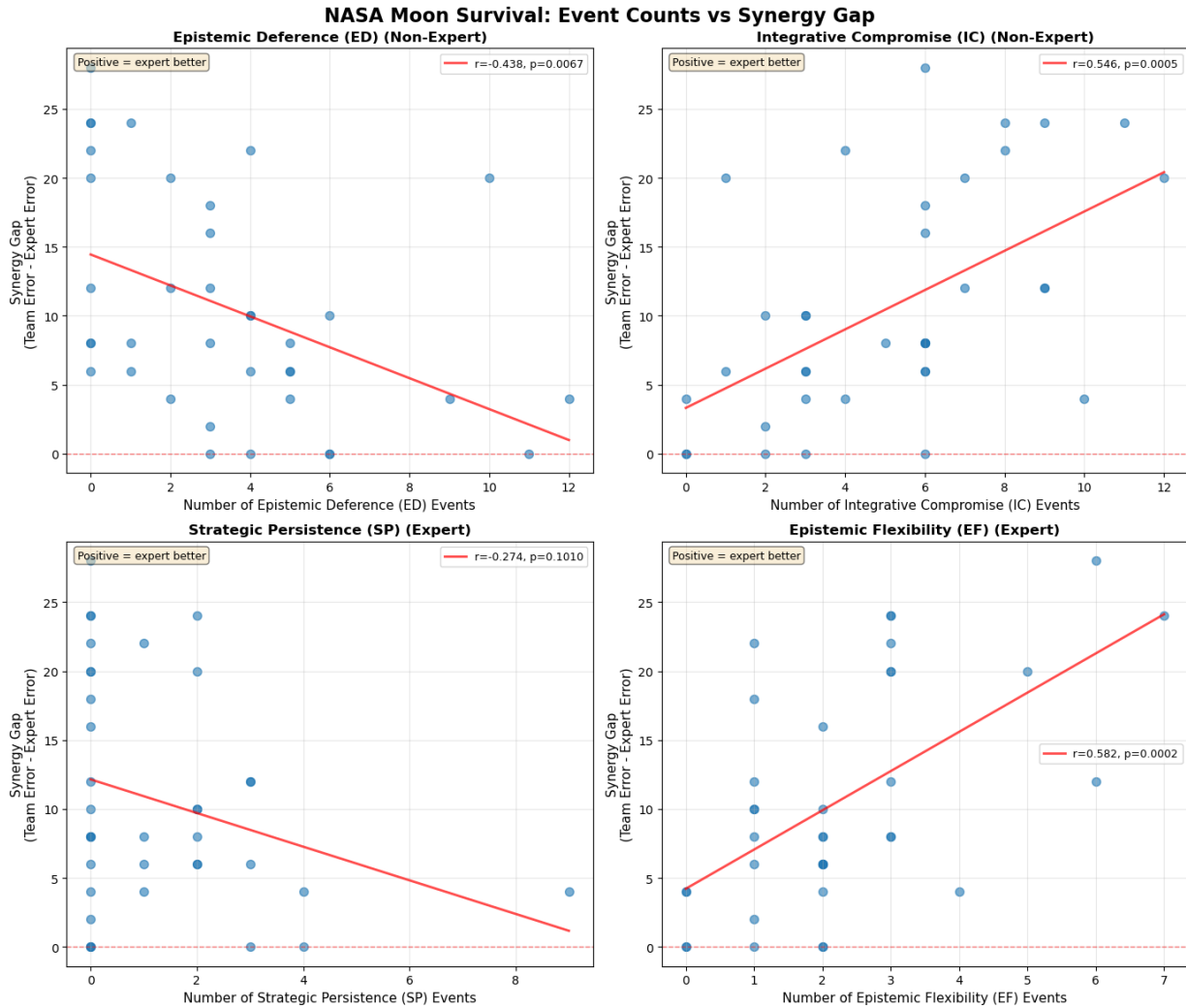
Figure 27. Epistemic deference analysis for NASA Moon Survival. Non-expert deference (ED) and compromise (IC) show significant opposite effects: ED correlates positively with performance ($r = 0.44$, $p = 0.007$) while IC correlates negatively ($r = -0.55$, $p < 0.001$), confirming that teams perform better when non-experts defer rather than compromise. Expert flexibility (EF) strongly harms performance ($r = -0.58$, $p < 0.001$); persistence (SP) shows a positive but non-significant trend ($r = 0.27$, $p = 0.10$). Each subplot shows correlation between event frequency and synergy gap (*Best Individual* − Team Score); negative values indicate team underperformed.
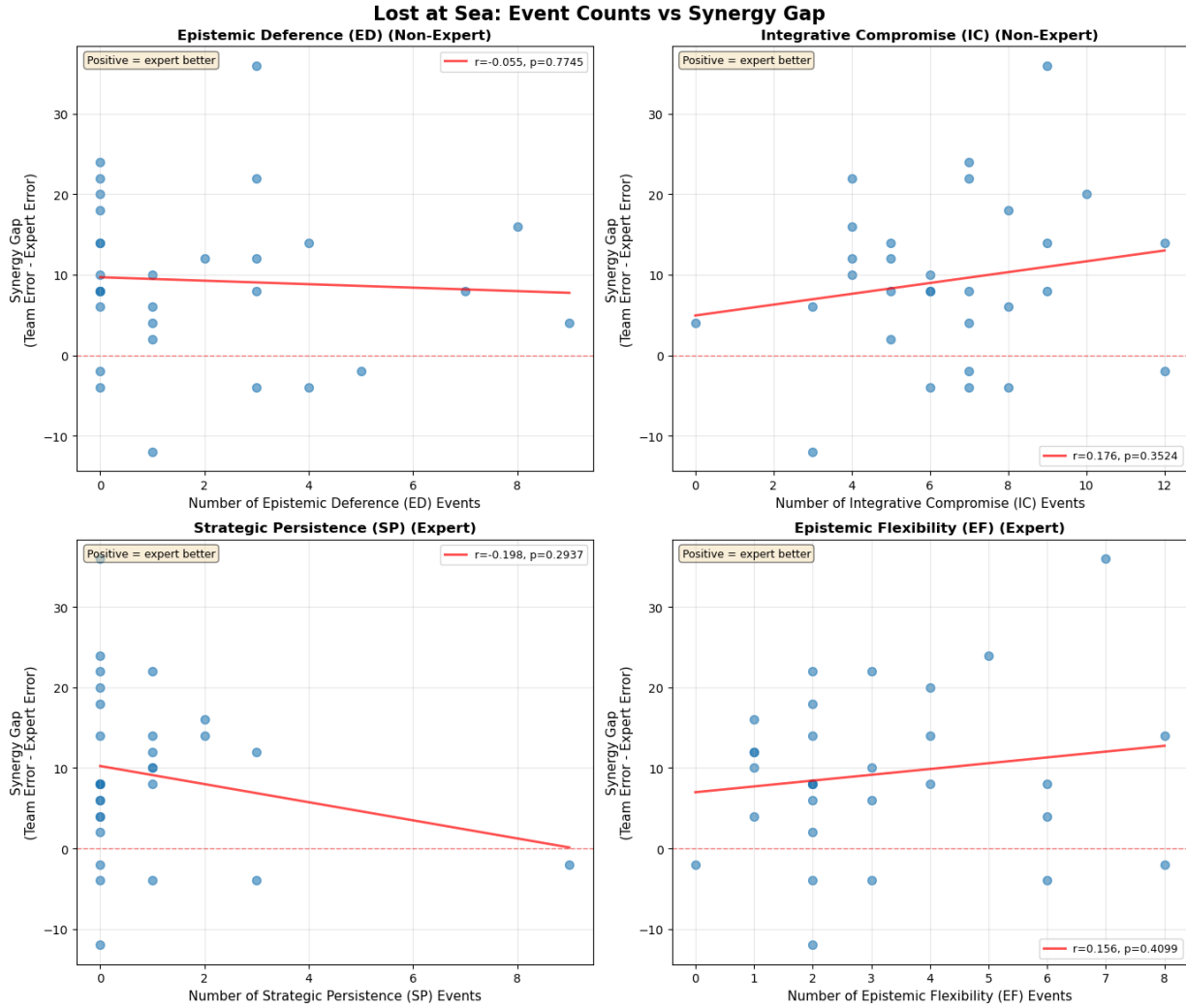
Figure 28. Epistemic deference analysis for Lost at Sea. Directional trends are consistent with other tasks (ED positive, IC negative, EF negative) but do not reach statistical significance. ED: $r = 0.06, p = 0.77$; IC: $r = -0.18, p = 0.35$; SP: $r = -0.20, p = 0.29$; EF: $r = -0.16, p = 0.41$.
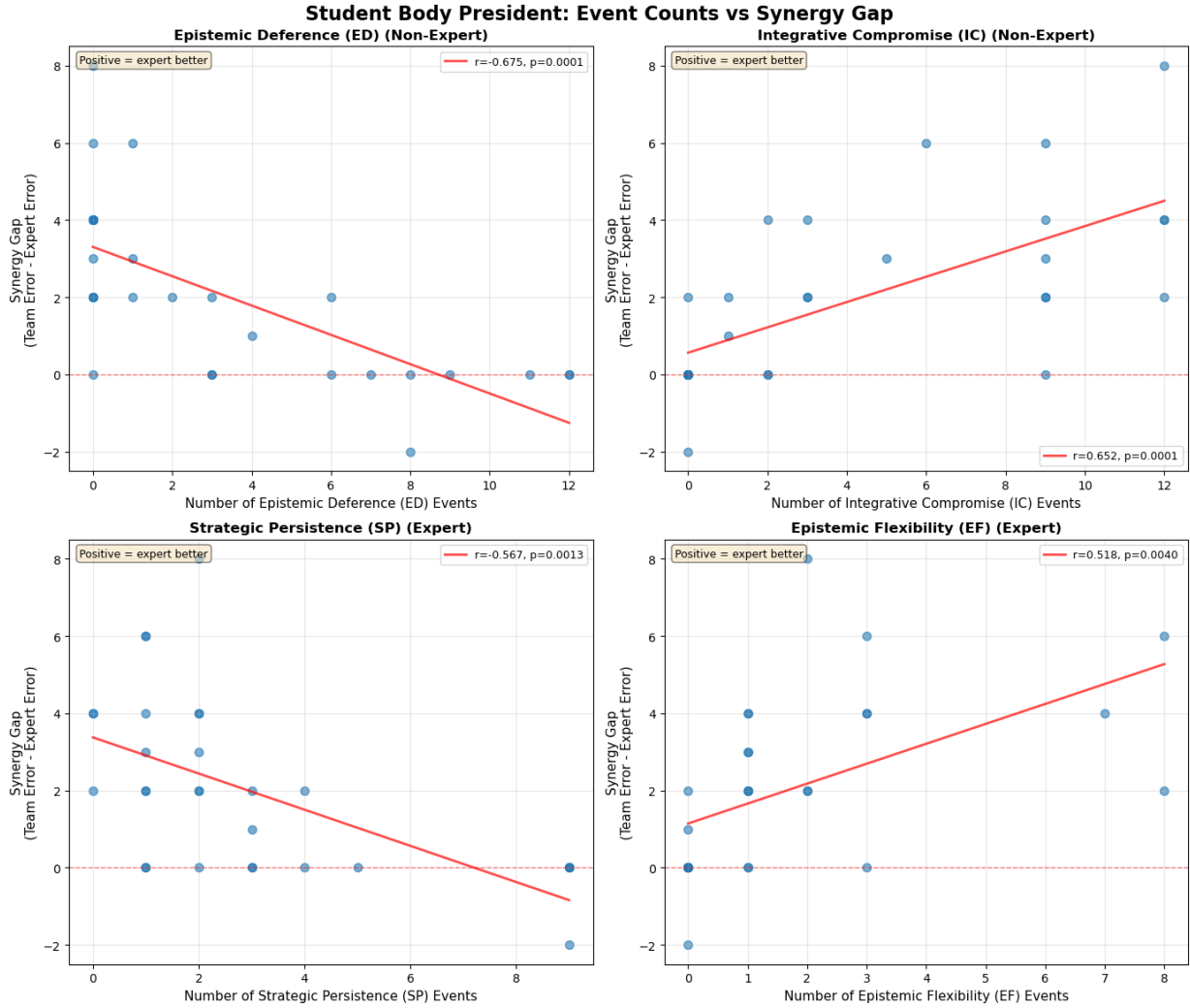
Figure 29. Epistemic deference analysis for Student Body President. This task shows the strongest correlation patterns. **Non-Expert**: ED ($r = 0.68, p < 0.001$) and IC ($r = -0.69, p < 0.001$) show highly significant opposite effects. **Expert**: SP ($r = 0.57, p = 0.002$) and EF ($r = -0.61, p < 0.001$) both significant, demonstrating that expert persistence helps while flexibility harms performance.