The Collaboration Gap: Exploration and Benchmarking of Open-World Agentic Cooperation

Tim R. Davidson[†*] **Adam Fourney** [‡] **Saleema Amershi**[‡] **Robert West**[†]
Eric Horvitz[‡] **Ece Kamar**[‡]

[†]EPFL [‡]Microsoft Research

Abstract

The trajectory of AI development suggests that we will increasingly rely on agent-based systems powered by language models, composed of independently developed agents with different information, privileges, and tools. The success of these systems will depend critically on effective collaboration among these heterogeneous agents, even under partial observability. Despite intense interest, the literature lacks empirical studies evaluating agentic collaboration without relying on fixed communication protocols, limiting insights for open-world deployments. We propose an illustrative collaborative maze-solving benchmark that (i) isolates collaborative capabilities, (ii) modulates problem complexity, (iii) enables scalable automated grading, and (iv) imposes no output-format constraints, benchmarking unguided, natural communication. Using this benchmark, we evaluate 32 leading open- and closed-source models in solo, homogeneous, and heterogeneous pairings. Our results reveal a surprising *collaboration gap*: models that perform well solo often degrade substantially when required to collaborate. We identify mitigations that show remarkable influence. For example, a small nudge via a *relay inference* approach, where a stronger agent leads before handing off to a weaker one, closes much of the gap. Our findings argue for (1) collaboration-aware evaluation, (2) training strategies to enhance collaborative capabilities, and (3) deliberate interaction design to elicit agents' collaboration skills, principles relevant to AI–AI and human–AI settings where agents must establish common ground. [§]

1 Introduction

Although collaborative interaction is remarkably efficient for humans (Garrod & Pickering, 2004), it is unclear to what extent such collaborative competency holds for AI agents powered by language models (LMs). Collaborative competency among agents is a rising concern with the deployment of agentic systems. The quest for faster innovation and economic progress through AI promises a steep increase in systems composed of multiple, independently developed AI agents. Due to privacy and organizational constraints, agents are provided with varying capabilities, contextual information, privileges, tools, and allowable actions.

Current attempts at integrating multi-agent solutions strongly rely on predetermined interoperability protocols, e.g., MCP (Anthropic, 2024), A2A (Google, 2025), and ACP (Besen, 2025), or centrally orchestrated architectures (Guo et al., 2024). Such reliance poses two clear obstacles for open-world integration of AI agents: First, given the fast-paced rise of heterogeneous agentic actors and systems, defining a fixed communication protocol in advance for every possible real-world situation is not feasible. Second, even if developing an expressive, standardized protocol *was* feasible, there is no way to enforce such protocols on agents controlled by multiple parties. Consequently, successful open-world deployment of AI agents will rely on their ability to engage in flexible, on-the-fly communication.

[*]Research conducted during an internship at Microsoft Research.

[§]Code for this project is available at `https://github.com/trdavidson/collaboration_gap`.

The emergence of sophisticated capabilities from pre-training data is a defining feature of modern LMs (Brown et al., 2020; Wei et al., 2022). Yet, most benchmarks focus on static reasoning capabilities (Chang et al., 2024), instead of capabilities unique to agentic use cases such as dynamic, long-horizon, multi-agent collaboration in partially observed environments (Davidson et al., 2024). Without targeted evaluations, this leaves a critical question regarding the limits of current training recipes: How does solo performance on tasks relate to *ad hoc* collaborative competency of AI agents (Stone et al., 2010)?

So far, research on collaboration involving LM-powered agents has predominantly focused on human–AI collaboration, optimizing AI's role of a "helpful assistant" (Bai et al., 2022). Still, it is unclear how much progress in this area transfers to AI–AI collaboration. Since collaboration underpins many societal functions, understanding how AI agents work together is vital for their successful integration.

AI's impressive performance in areas like creative writing, coding, and STEM subjects makes it tempting to also expect proficiency in collaboration, a capability so natural to humans it is easily taken for granted. However, effective collaborative interaction among humans has been shown to rely on multiple levels of analysis, drawing upon numerous subtle cues and assumptions (Clark, 1996; Garrod & Pickering, 2004). Research on grounding — or achieving mutual understanding — between AI systems and people, building on results from studies in psychology, further emphasizes the need for multiple perceptual and inferential skills to achieve fluid collaboration in the open world (Pejsa et al., 2014). Our limited understanding of AI agents' collaborative capabilities can be partially explained by their recency. More fundamentally, evaluating advanced AI capabilities is simply very challenging; benchmarks quickly become outdated (Maslej et al., 2025; Kwa et al., 2025), confounded by other factors (Henderson et al., 2018), or fail ecological plausibility (De Vries et al., 2020). To study autonomous AI–AI collaboration at scale, we need tasks with automated outcome metrics that allow us to modulate complexity, isolate collaborative capabilities, and put minimal constraints on model outputs.

The arrival of strong LMs enabled the creation of general-purpose agents using rich natural language. This shifted focus from learning low-level control problems through, e.g., reinforcement learning (RL) (Zhu et al., 2024), to orchestrating high-level reasoning and task execution among capable, pre-trained agents (Zhuge et al., 2025). Although such orchestrated solutions and carefully crafted agentic teams have shown great promise (Guo et al., 2024; Fourney et al., 2024; Hong et al., 2024; Chen et al., 2024; Tran et al., 2025), a recent study of popular multi-agent systems collaborating on math and coding tasks identified "*failures arising from ineffective communication, poor collaboration, [and] conflicting behaviors among agents*" as an important failure mode (Pan et al., 2025). These results highlight that collaborative capabilities can form a bottleneck even in tightly orchestrated systems.

Studies of emergent interactions in role-based societal simulations offer interesting insights into model behavior (Park et al., 2023), but lack control and measurable outcomes relevant to collaboration. Other recent works by Wu et al. (2025) and Zhou et al. (2025) optimize LMs' collaborative capabilities for human–AI collaboration through RL. While these studies use LMs to simulate the human actions, the core focus is not AI–AI collaboration. In general, most studies explore *homogeneous* interactions, where all agents are powered by the same underlying model, or a limited selection of *heterogeneous* models, likely limiting their generalization (Wynn et al., 2025). Davidson et al. (2024) use negotiations to evaluate interactions between heterogeneous agents with minimal output constraints. While negotiations have important collaborative elements, they often contain incentives to withhold information and deceive.

We introduce a scalable maze-solving benchmark designed to isolate dynamic collaborative capabilities. Using this benchmark across 32 leading open- and closed-source LMs, we identify a critical and counterintuitive phenomenon we term the "collaboration gap:" models that are highly capable solo performers exhibit a statistically reliable performance drop when collaborating with an identical copy of themselves. Our contributions are threefold: (1) We formally define and provide empirical evidence for the collaboration gap, showing it appears especially severe in distilled models. (2) We analyze the dynamics of heterogeneous collaboration, revealing that task performance is heavily influenced by which agent initiates

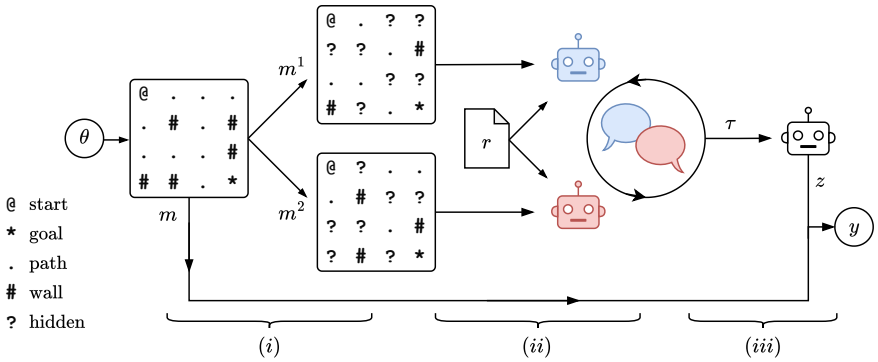


Figure 2.1: **Core System Diagram.** In sequence, we (i) first generate a random maze $m \sim \mathcal{M}(\theta)$ and split it into two copies, m^1, m^2, with roughly half the cells obfuscated using "?" symbols, (ii) each agent is then given a maze copy along with the rules, r, after which they engage in a dialogue until either the maximum number of turns is reached or the task is completed, finally (iii) the raw transcript, τ, is passed to a "grader" agent that extracts the agreed upon route, z, which is checked against the ground-truth maze to decide outcome y.

the dialogue. (3) Based on this insight, we propose "relay inference," a new collaborative strategy where a more capable model "seeds" the initial steps of a task for a weaker one, substantially boosting collaborative performance of weaker models. We interpret our results as an existence proof that effective collaboration represents a distinct axis of capability that current training strategies insufficiently capture.

2 Measuring agents' collaborative capabilities

2.1 Collaborative maze solving

Following seminal studies on human collaboration (Garrod & Anderson, 1987; Garrod & Doherty, 1994), we employ "maze solving" as our target task. Mazes contain desirable properties that make them especially suited to study collaboration. First, they represent "fixed" constructs that nevertheless can be described in a variety of ways; none more correct than another. For example, the same maze may be described using row-column notation or using visual descriptions, creating the need to "align" representations between parties. Secondly, while there are many ways to describe a solution, its correctness can be readily verified. Thirdly, a maze can be solved both "solo" and as a team. This allows us to disentangle agents' "raw" maze-solving capabilities from those needed for collaboration. Finally, mazes can be made arbitrarily complex by simply scaling up their sizes and wall density.

In our study, mazes consist of N×N grids with a start and goal state, separated by path and wall cells. The dimensions of a maze, distance between start and goal state, and wall density are parameterized by θ. Following Figure 2.1 (i), we first sample a maze instance, $m_i \sim \mathcal{M}(\theta)$. Next, we introduce a "twist," designed to incentivize agents to collaborate: Instead of simply providing each agent a copy of the complete maze, we distribute the information by randomly obfuscating half the cells of each copy resulting in m_i^1 and m_i^2. Combined, the two copies recover the complete map ($m_i = m_i^1 \cup m_i^2$). To fairly compare solo to collaborative performance on the decomposed task, we can present a single agent with both of the distributed maps to measure its capability to handle distributed information.

To generate trajectories, we provide two agents (parameterized by LMs) each an incomplete maze copy as a visual, text-based representation of N×N symbols, along with the following rules, r: (1) "*both agents must agree on a move before it is executed*," and (2) "*only one move can be executed at a time*." The first rule prevents a single agent from unilaterally jeopardizing the rollout, while the second ensures a minimal number of interactions. Except for a predetermined completion phrase, we do not enforce or provide *any* other guidance on the communication protocol the agents must use or output format they must follow. Given the visual format of the maze copies, this leaves considerable room for interpretation. The

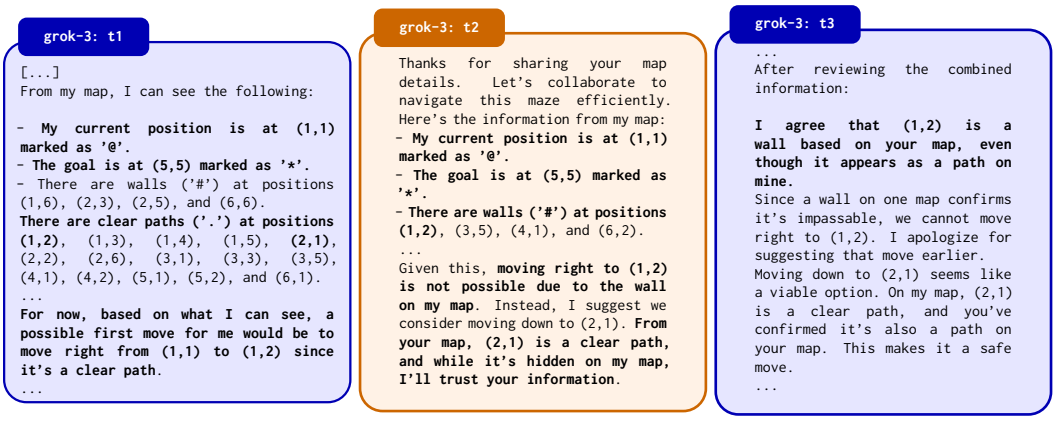


Figure 2.2: **Coordinate Grounding Failure.** In the above exchange between two grok-3 agents, the row-col coordinates $(1,2)$ and $(2,1)$ are a path/hidden for agent 1, and hidden/wall for agent 2. The first agent uses (row, col) coordinates to describe its map, but does not explicitly specify this (It also wrongly communicates *both* $(1,2)$ and $(2,1)$ as paths). Because the start and goal coordinates are symmetric, i.e., $(1,1)$ and $(5,5)$, this is not immediately clear to the second agent, which opts to use (col, row) coordinates.

agents then engage in a multi-turn dialogue attempting to solve the task. At each turn, agents have access only to their copy of the maze, the rules, and the dialogue thus far. This process continues until either the maximum number of turns is reached or the predefined completion phrase is uttered, producing final transcript τ_i (Figure 2.1 (ii)). Because no output structure is enforced, there is no straightforward way to deterministically extract the proposed solution. Instead, we use a third agent as a "grader," tasked with extracting the agreed upon moves at each turn from the raw transcript, $a_{\text{grader}} : \tau_i \to z_i$. Moves can follow different "maze schemes." For example, they can consist of directions like "up" and "down," or coordinates like "$(0,1) \to (0,2)$." They can also use different origin locations, e.g., top-left = 1, or bottom-left = 0. To ensure the most favorable interpretation we thus transform z_i under a large set of potential schemes and evaluate each against the original maze m_i to obtain the final outcome y (Figure 2.1 (iii)). See App. B for prompts used.

In contrast to human collaboration studies to date, the ability to use auto-grading allows us to massively scale our experiments. We perform extensive sensitivity analyses to measure the stability of repeated and different grader models, finding limited variance across average outcomes and virtually no evidence of model-specific biases (see Appendix E).

2.2 Communication challenges

While straightforward, our maze-solving task offers a rich evaluation environment to test diverse collaborative challenges. Perhaps the most important among those is the concept of "grounding," the process by which participants attempt to establish mutual understanding (Clark & Brennan, 1991). The goal is to ensure that exchanged information is understood by all parties the same way. The first challenge is to build a shared mental model of the maze itself. In order to plan a route, agents must exchange information about their partial views to *ground* their understanding of the full maze, e.g., they must establish a shared understanding of the start and goal position.

Absent a fixed communication protocol, agents must further agree on how to *refer* to locations and actions (Garrod & Anderson, 1987; Harnad, 1990). For instance, two agents are in trouble if one uses (row, col) coordinates and the other interprets these as (col, row) (see Figure 2.2). Similarly, the condition that both agents need to agree on each move necessitates *action grounding*, i.e., to ensure that the intended move is clear to the other agent.

As the trajectory progresses, the agents have to resolve different types of conflicts. For example, they might encounter *perceptual conflicts* where they disagree on the contents of a specific cell or their current location. This requires the ability to resolve the inconsistency and update their shared understanding of the maze. The unstructured nature of the task

also introduces a procedural challenge: who gets to propose the next move? Should one agent take the lead, or should they alternate? This can be especially useful in the case of different capabilities: can agents recognize which is more capable and dynamically defer? To that end, *Theory of Mind* (Premack & Woodruff, 1978) — the capacity to reason about the other agent's mental state — can be beneficial both to effectively convince another agent to take an action and to resolve potential conflicts, as we observe in Section 4.2.

3 Experimental setup

Maze details. Our main experiments are conducted using 6×6 mazes with a wall density of 30% and an average solution path of 7 to 9 steps. These parameters ensure most models can solve mazes solo, providing a meaningful collaborative baseline; ablations are in App. C.

Solo baseline. We establish a baseline for solo agents' "raw" maze-solving capabilities under two conditions: a single, fully visible map, and a "distributed map" where information is split across two maps as described in Section 2.1. The gap between these conditions reflects the challenge of using distributed information. In each setting an agent is permitted a "critic" step, in which it can review its proposed solution. We collect 100 samples per setting.

Homogeneous collaboration. In a natural extension of the distributed solo setting, e.g., distributed information and iterative discussion, an agent now has to collaborate with an independent copy of itself, effectively isolating the collaboration capability. As each agent copy only has access to half the map, they need to engage the other agent to fill in the missing pieces. We perform 100 rollouts per agent, with a maximum of 50 turns.

Heterogeneous collaboration. This setting follows the same mechanics as the homogeneous one, but focuses on collaborations between agents of different model families and/or different strengths. This allows us to examine possible "ordering" effects on collaborative outcomes. For each pairing, we perform at least 50 rollouts with a maximum of 50 turns.

Relay inference. We further explore the opportunities of efficient heterogeneous agent deployment using a "relay inference" approach. Given two available agents, A and B, where A is stronger (costlier) than B, assume we have to deploy an agent to collaborate with another party's agent, and that we cannot control the choice of agent used by the other party. In the event the other party chooses the weaker B to save costs, we would like to know if a strong model can be used to (i) "prime/ground" a rollout before switching to a weaker model, and (ii) "recover" after starting with a weaker model. To that end, we first generate a set of rollouts using AB and BB. We then "freeze" the first $K \in \{2, 4, 6, 8\}$ turns, after which respectively the weaker or stronger model is swapped in to complete the rollout. We perform at least 100 rollouts per model pairing and relay point K.

Models used. For the solo and homogeneous collaborative experiments we include most commercially available frontier models (full list in App. B). To avoid combinatorial explosions in heterogeneous settings, we focus on selected models of the most popular providers: Anthropic, Cohere, Google, OpenAI, and xAI. For each, we investigate "in-family" pairings and selected "cross-family" pairings. We use gpt-4.1 as the grading model (see App. E).

Outcome metrics. We distinguish between a binary success rate, i.e., the provided solution is a correct path from start to goal state, and a "weighted outcome," which is defined as the ratio between (a) the optimal path length from start to goal and (b) the distance from the last valid position on a path to the goal state, $\frac{a-b}{a}$. Note that this ratio can be negative. To provide a continuous evaluation signal, we report weighted outcomes in the results section.

4 Results

The results section will focus on weighted outcome results in the various settings. Additional dialogue samples, ablations, and results are available in the Appendix.

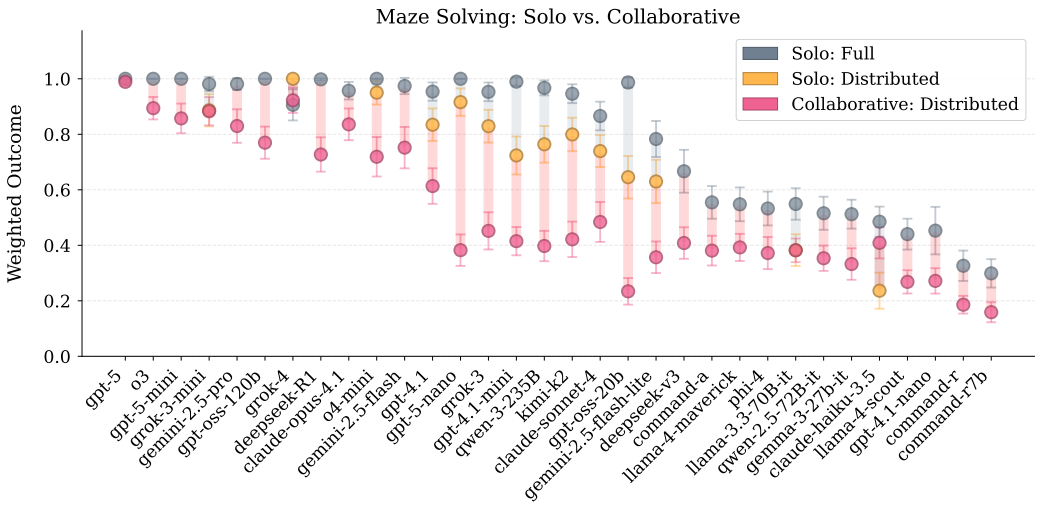


Figure 4.1: **The Collaboration Gap.** Mean weighted outcomes of 6×6 maze-solving experiments in solo and homogeneous collaboration mode with 95% CI. We set a "maze solving" capability baseline by presenting a model with the full maze (gray marks), showing weighted outcomes surpass 0.5 for most models. Next, we present models the same mazes but "distribute" the information over two map copies (yellow marks), reporting a performance degradation only if values pass a 95% CI between the solo settings. Finally, we evaluate models' collaborative capabilities by pairing them with an independent copy of themselves, providing each a distributed half of the maze (red marks). Most models display a significant performance "gap" between solving mazes solo and collaboratively.

4.1 Solo baselines versus homogeneous collaboration

Weighted outcomes for solo experiments are shown in Figure 4.1 (gray and yellow markers). Most evaluated models are capable of at least partially solving mazes under full visibility, with half the models reaching near perfect solve-rates (gray). Performance drops significantly for about a third of the models when changing to distributed information (yellow). Overall, most models are capable of (partially) solving 6×6 mazes. Virtually all studied models experience a statistically reliable performance drop when moving from a solo to a collaborative setting (red markers), a degradation unlikely to be explained by context or turn limits (see Appendix D.5). Larger models and "thinking" models tend to perform better than smaller, "regular" models, though exceptions exist. For example, grok-3 and kimi-k2 both underperform grok-3-mini.[2] Distilled models appear disproportionately affected, e.g., note the large drops from gpt-5 to gpt-5-nano and gemini-2.5-pro to gemini-2.5-flash-lite, suggesting that our current distillation strategies may lose more than just information from the tail-ends of a base model's knowledge distribution (Du et al., 2023).

o3: turn 1

Here's my view of the maze **(row 0/col 0 = top-left)**:
Row 0: (0,0)=@, (0,1)=?, (0,2)=., (0,3)=?, (0,4)=?, (0,5)=?
Row 1: (1,0)=?, (1,1)=?, (1,2)=., (1,3)=#, (1,4)=#, (1,5)=?
Row 2: (2,0)=., (2,1)=?, (2,2)=., (2,3)=?, (2,4)=., (2,5)=?
Row 3: (3,0)=?, (3,1)=?, (3,2)=., (3,3)=., (3,4)=?, (3,5)=.
Row 4: (4,0)=#, (4,1)=?, (4,2)=., (4,3)=?, (4,4)=., (4,5)=?
Row 5: (5,0)=., (5,1)=?, (5,2)=?, (5,3)=#, (5,4)=?, (5,5)=* (goal)

1. Could you tell me the contents of any "?" cells you can see, especially (0,1) and (1,0) so I know a safe first move?
2. Where is your current position on your map? We'll use coordinate format (row,col).

gpt-4.1-mini: turn 1

Hello! I see that we are **starting at the position marked '@'** on my map, and **the goal is marked '*'**. There are some **hidden cells marked with '?'** that we need to figure out together.

To begin, I suggest we **share what we see around our starting position. On my map, from '@', I can move right to '.', or down to '?'. What do you see around your starting position?**

Figure 4.2: **Primer Messages.** The difference in the first "primer" message to start a dialogue sent by a strong (o3) and weaker model (gpt-4.1-mini). Note the effort to unambiguously "**ground**" language on multiple levels by o3, and the lack of clear grounding by gpt-4.1-mini.

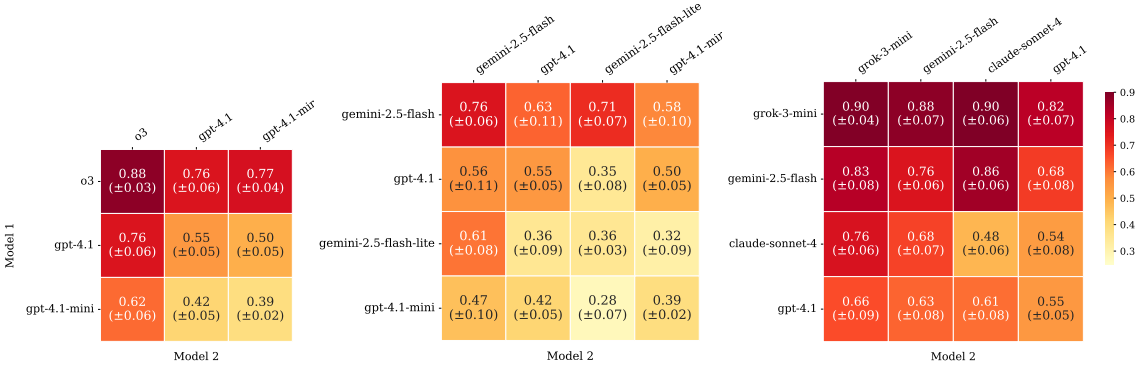
What explains these differences? As established above, both o3 and gpt-4.1-mini are capable of solving mazes solo, yet have drastically different outcomes in collaborative mode. Qualitative analysis of their first messages is illuminating: In Figure 4.2, we display the

[2]Crucially, grok-3-mini is reportedly *not* a distilled model, but purpose-built for its size (xAI, 2025).

first message sent by o3 (left) and the first message sent by gpt-4.1-mini (right) to another copy of themselves. The stronger o3 immediately seeks to align its maze representation by providing a fully determined schema, request for missing information, and immediate next steps. It also makes sure to "ground" the starting position. In contrast, gpt-4.1-mini only attempts to ground the meaning of different symbols, without proposing a communication schema, or clearly defining its starting position.

App. D.2.1 formalizes this grounding-outcome link. We first annotate a large, balanced sample of rollouts for various grounding proxies (e.g., aligning on start and goal state) and strategy type (global vs. local planning). We then compute a grounding score and fit both an additive and interactive mediation model, predicting weighted outcomes while controlling for several confounding factors (e.g., model family and optimal solution length). Results of clustered bootstrapping on agent pairs are displayed in App. Table 2, showing a statistically significant relation between grounding proxies and weighted outcomes.

4.2 Heterogeneous collaboration



(a) OpenAI (b) OpenAI and Google (c) Anthropic, Google, OpenAI, xAI

Figure 4.3: **Heterogeneous Collaboration.** Each cell represents mean weighted outcomes with 95% CI below. We observe significant ordering and cross-family effects.

Results for selected heterogeneous collaborations are shown in Figure 4.3, with additional results available in Appendix D.3. The diagonals represent homogeneous collaboration.

4.2.1 Same model family: working with models of different strength

In panel (a) we observe that pairing a strong model like o3 with a weaker model fails to replicate the performance of the stronger model. This is surprising, as the stronger model retains the power to veto bad moves, even if it cannot unilaterally solve the maze. Secondly, we note that pairings can display pronounced ordering effects. For example, gpt-4.1-mini paired with o3 performs markedly better when o3 starts the collaboration (0.77 ± 0.04 vs. 0.62 ± 0.06). A sample exchange started by gpt-4.1-mini with o3 moving second is shown in Figure 4.4. In stark contrast to the grounded approach o3 displayed in Figure 4.2, o3 partially "imitates" the communication format introduced by the weaker gpt-4.1-mini. While o3 still proposes a coordinate system to use, it no longer attempts to ground the complete map, instead opting for a greedy "local" strategy. Similar imitation tendencies can be seen the other way around, where gpt-4.1-mini adopts conventions introduced by o3 (see Figure F.1).

gpt-4.1-mini: turn 1

Hello! I see **my current position is at the top-left corner marked '@'**, and **the goal '*' is somewhere in the bottom-right corner**, but many cells are hidden. Let's start by moving right to reveal more of the maze. Shall we move right?

o3: turn 2

Agreed—**moving right from the start puts us at (row 0, col 1)**. On my map that cell is clear, and the **next one to the right (row 0, col 2)** is also an open path '.'. The top row is clear until a wall at column 5, so we could keep heading right, or drop **down from here (row 1, col 1)** which is also open on my side. I propose we move right again to (0, 2). Sound good?

Figure 4.4: **Style Imitation.** Example of a stronger model imitating the style of a weaker one.

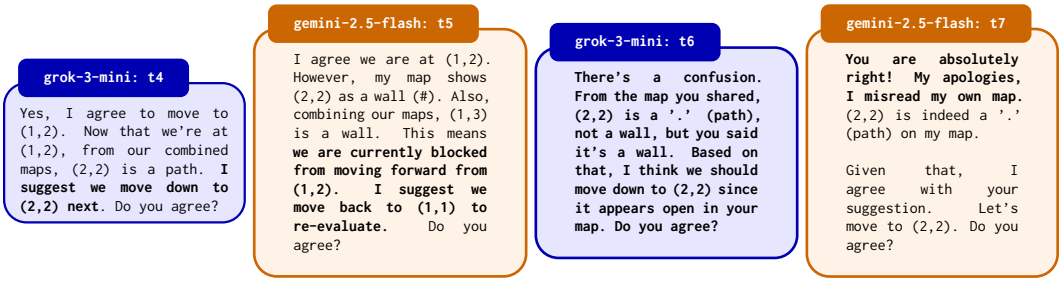


Figure 4.5: **Perceptual Grounding.** Example of two agents resolving a disagreement.

4.2.2 *Different model family: working with models of varying strength*

When crossing model families (b), we notice the same general trends as in (a): the stronger model generally serves as an upper bound on performance and ordering affects outcomes. We additionally observe that the weaker model's performance does not necessarily represent a lower bound, e.g., gemini-2.5-flash-lite and gpt-4.1-mini each perform better with copies of themselves than when combined. We also find evidence that models can display a certain affinity for models of their own family, e.g., note how gemini-2.5-flash-lite does not improve its performance when paired with the stronger gpt-4.1, but performs well when paired with gemini-2.5-flash (see also App. D.2.2).

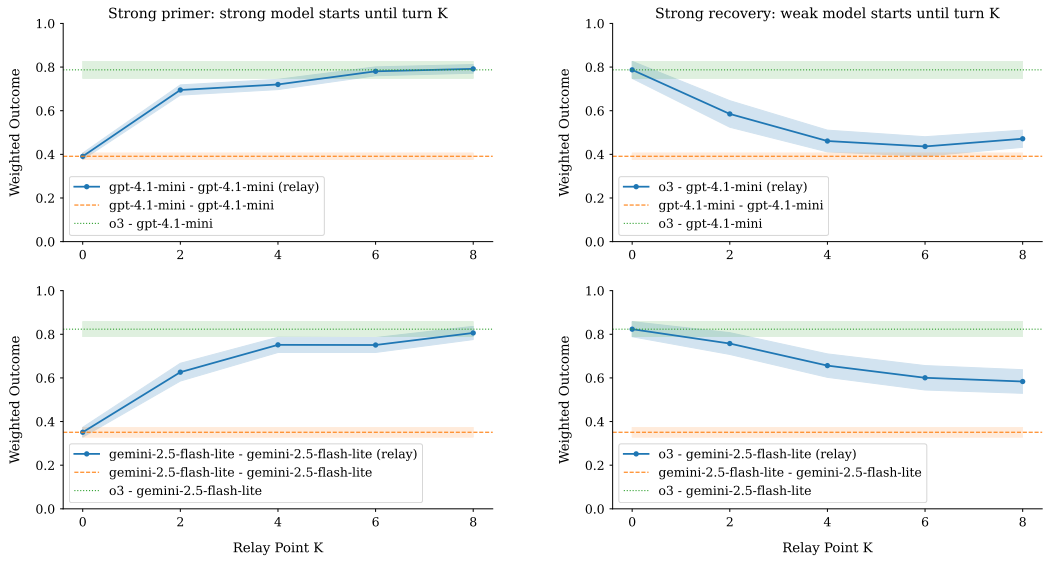
In panel (c) we show pairings of flagship models meant for everyday tasks from leading model builders. Across different model families, we find the same ordering effects. Although within the confidence interval, claude-sonnet-4 is the sole exception to this trend: its collaborative performance with gemini-2.5-flash and gpt-4.1 surpasses both of the models' homogeneous performance. Grok-3-mini shows itself an especially capable collaborator, managing to keep performance close to its high homogeneous baseline. Notably, it actively challenges incorrect moves rather than deferring to weaker partners. For instance, see the exchange between grok-3-mini and gemini-2.5-flash in Figure 4.5, in which grok-3-mini proposes to move from (1,2) to (2,2). The gemini model (incorrectly) disagrees with the move and suggests moving back to re-evaluate. Instead, the grok-3-mini model (i) recognizes the confusion, (ii) attempts to re-calibrate the shared map information, and finally (iii) queries for agreement. The gemini model recognizes its mistake and the agents continue.

Appendix D.2.2 quantifies such disagreement dynamics. We first annotate a large, balanced sample of rollouts for the presence of two sorts of disagreements: "*factual*" disagreements based on perception or move legality, and "*non-factual*" disagreements based on grounding or strategy. We then model the likelihood of disagreement events using a logistic regression accounting for factors such as model strength (weak, medium, strong) and interaction type (homogeneous or not).

Results of clustered bootstrapping on agent pairs are presented in Table 3 of the Appendix, showing that on average, the odds of encountering any conflict are significantly lower for homogeneous collaborations (OR = 0.60). Strong- and medium-strength models are more likely to encounter disagreements than weaker ones, while factual conflicts are the most likely in medium-strength models. Non-factual conflicts appear to increase with capability, suggesting a "sophistication" shift in agent conflict, with strong models significantly more likely to trigger disputes over strategy or grounding than the medium-strength reference models. Interestingly, homogeneous pairs disagree less often, but not about different things: once a conflict occurs, they are just as likely as heterogeneous pairs of comparable strength to be disputing strategy or grounding (OR = 0.97).

4.3 Relay inference

Relay inference results are displayed in Figure 4.6. Panel (a) shows that a single priming message from the stronger o3 can substantially boost performance for both the weaker models (gpt-4.1-mini, top, and gemini-2.5-flash-lite, bottom). Crucially, o3 only sees its



(a) Strong Primer

(b) Strong Recovery

Figure 4.6: **Relay Inference.** We test the importance of the initial interactions on outcomes using a strong model (o3) and two weak models (gpt-4.1-mini, top and gemini-2.5-flash-lite, bottom): (a) a strong and weak model interact for K turns, after which another copy of the weak model is swapped in for the strong one; (b) two weak models interact for K turns, after which the strong model is swapped in. Using a strong model for priming appears more effective than using it for recovery.

incomplete map copy at the first turn, preventing it from unilaterally proposing a full solution. Instead, o3's opening message grounds the coordinate schema and reference frame, reducing coordination to instruction-following for the weaker model (see App. F.1). Conversely, panel (b) shows diminishing returns for late interventions: prolonged initial exchanges between weak models make strong recovery harder, likely due to the style imitation effects observed in Section 4.2. Taken together, these results suggest that using strong models to "seed" collaborations can be more effective than using them as backup "experts," jumping in to course correct.

5 Discussion

Due to space constraints, we provide an extensive related work section in Appendix A.

Collaborative models are the future of agentic AI. A compelling case for "small" language models to lead the agentic age was recently made by Belcak et al. (2025), who argue that capable, small *specialist* models would be more practical and economical to solve specialist tasks compared to large and expensive *generalist* models. However, there is an important caveat missing from this discussion: the more specialized an agent becomes, the higher the likelihood it encounters challenges outside its area of expertise. Consequently, the need to effectively collaborate with "other" agents to fill capability gaps increases with specialization. Our work suggests that for current models, "naively" breaking up problems to be solved by multiple agents could introduce collaborative slippage, emphasizing the necessity of explicitly training for collaborative capabilities.

Unlocking collaborative capabilities. "In-context learning," or simply "prompting" can be a powerful tool to unlock latent capabilities in LMs (Brown et al., 2020). Subsequently, a large body of literature and guides has sprung up focused on "prompt engineering" to improve model outputs (Sahoo et al., 2024). Two important findings from our study are the potential of priming to improve collaboration in weaker models, and, conversely, the subdued performance when weaker models lead. With human–AI collaboration typically led by humans, these observations hold lessons for a future where AI continues to improve.

First, it gives credence to national initiatives to increase people's competence in interacting with AI (Mason, 2025; GSA, 2024). Second, it suggests AI will increasingly have to solve the *specialist librarian problem* (Taylor, 1968) to improve human–AI and AI–AI collaboration. In this view, information queries are not single events, but dynamic processes during which the librarian (AI) first has to decipher someone's "true" needs before being able to solve them.

Mazes as a lower bound. Clearly, maze environments do not capture the full complexity of real-world collaboration, nor do they serve as a proxy for *all* types of collaboration. Yet, as argued in Section 2.2 they do simulate *many* vital aspects of collaboration. The fact that a large collaboration gap exists even in the stylized case of mazes leads us to conjecture that the gap might be wider — quantitatively as well as qualitatively — in more complex, real use cases. Our "distributed" information methodology offers a scalable approach to design new collaboration tasks, the exploration of which we leave to future work. Take for instance the partial observability that can arise in collaborative software engineering: Just as two agents in a maze must reconcile different map views, coding agents can face a "split context," where each agent can observe only a subset of the required files. This can happen if a code base does not fit into an agent's context window or if agents have different role-based access controls. Echoing the maze setup, agents must agree on the folder structure to organize new files or imports, and align on coding conventions, package versions, and shared dependencies, while also resolving the occasional warnings and errors.

6 Conclusion

Human civilization is a story of continuous, rich collaborations. While AI promises a new chapter, our work reveals a critical roadblock: a fundamental "collaboration gap" demonstrated across today's leading language models powering AI agents. This is a counterintuitive phenomenon where agents with high solo capabilities exhibit a sharp performance collapse in simple teamwork. Enabled by a novel benchmark that isolates collaborative capabilities, our work demonstrates interaction order is a critical factor. Our findings offer a promising insight, showing that strategically priming interactions via "relay inference" can reliably improve joint outcomes. The gap's appearance in our simple testbed is particularly concerning, as it reveals a blindspot in current training paradigms rather than an artifact of task complexity. Collaboration is a critical capability and recognizing its centrality is an opportunity for investing in addressing gaps and boosting AI capabilities. This calls for a paradigm shift, an idea concisely articulated by Grosz (1996): "*capabilities needed for collaboration cannot be patched on, but must be designed in from the start*." We challenge the research community to treat collaborative intelligence as a core objective to be designed for, not as an emergent property to be hoped for.

Acknowledgments

The authors thank Eric Zhu and members of the Microsoft Research AI Frontiers Team for stimulating discussions and feedback that contributed to this work.

Reproducibility statement

We open-source the code used to conduct the experiments described in this work at: https://github.com/trdavidson/collaboration_gap.

System prompts and user instructions. We share all prompts and models used to produce the results of this study in Appendix B, also including a detailed discussion on our implementation considerations. As with all works aimed at evaluating capabilities across LMs, there is likely a set of system prompts and instructions more suitable for a particular model. Due to the opaqueness of closed-source models, it is unfortunately impossible to determine if our prompts clashed with existing developer prompts, unfairly putting some models at a disadvantage. We made best efforts to ensure that all studied models understood the instructions and respected the agent-to-agent message protocol described in Appendix B.2.

We did not always succeed: upon qualitative inspection it was found that the command-a model from Cohere would sometimes "prematurely" use the predefined completion phrase when hallucinating about a successful outcome. Due to the large scale of our experiments, modifying the instructions was no longer feasible.

Non-determinism. Most models accessible over APIs are not deterministic (He, 2025). As the majority of our study consists of repeated interactions between (different) models, this non-determinism explodes. To mitigate this problem we made sure to collect at least 100 rollouts for homogeneous and 50 rollouts for heterogeneous collaborations. Whenever we report metrics, we include the 95% confidence intervals based on standard errors. While this strategy does not guarantee that the metrics we report represent *every* possible rollout, they do capture average tendencies. We also made sure that all mazes generated during this study used fixed random seeds for a fair comparison between models.

Auto-grading. As described in Section 2.1, the unstructured format of the generated rollouts makes it impossible to deterministically extract the proposed solutions. Since using human annotators is not feasible given the scale of our experiments (many tens of thousands of transcripts), we instead opt to use LMs as graders. In our experiments, we used OpenAI's gpt-4.1. Due to the non-determinism considerations mentioned previously, this can in theory introduce unwanted noise into the evaluation process. Furthermore, there is the possibility that some grader models might unfairly disadvantage participating models, e.g., an OpenAI model might be better at extracting results from OpenAI rollouts compared to those coming from Google or xAI.

We therefore conducted a comprehensive ablation in Appendix E to measure both the consistency of performing multiple rounds of grading using gpt-4.1 as well as possible changes in results when using the stronger OpenAI o3 model or Google's gemini-2.5-flash. We further stratified our analysis across a selection of the different models studied. We did not find any evidence for statistically significant noise in the grading process or unfair biases across models.

We did encounter at least one instance of an unanticipated maze schema not captured by our solution normalization process (see Appendix G.1). The large number of manual author checks does not suggest this is a widespread problem.

Costs of multi-turn rollouts. The goal of our study was to capture generalizable insights relevant to LMs' collaborative capabilities. We therefore attempted to include as many commercially available open- and closed-source models as possible. However, we recognize that replicating our study comes at a considerable cost, likely prohibitive for many labs and practitioners. This can partially be attributed to the higher cost of using some frontier models, but also to the increasing cost of generating a subsequent turn in a rollout. Ignoring the constant system and user instructions messages, for an average message length of K tokens and maximum number of turns T, the average input context length becomes $K \cdot (T-1)/2$.

References

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

Noa Agmon, Samuel Barrett, and Peter Stone. Modeling uncertainty in leading ad hoc teams. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 397–404, 2014.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *CoRL*, 2022.

Anthropic. Introducing the model context protocol, 2024. URL https://www.anthropic.com/news/model-context-protocol.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

Samuel Barrett, Peter Stone, and Sarit Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 567–574, 2011.

Samuel Barrett, Noa Agmon, Noam Hazon, Sarit Kraus, and Peter Stone. Communicating with unknown teammates. In *ECAI*, pp. 45–50, 2014.

Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242:132–171, 2017.

John Batali. Computational simulations of the emergence of grammar. *Approaches to the Evolution of Language-Social and Cognitive Bases-*, 1998.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*, 2025.

Sandi Besen. Acp: the internet protocol for ai agents, 2025. URL https://towardsdatascience.com/acp-the-internet-protocol-for-ai-agents/.

Michael Bowling and Peter McCracken. Coordination and adaptation in impromptu teams. In *AAAI*, volume 5, pp. 53–58, 2005.

Philippe Bretier and David Sadek. A rational agent as the kernel of a cooperative spoken dialogue system: Implementing a logical theory of interaction. In *International Workshop on Agent Theories, Architectures, and Languages*, pp. 189–203. Springer, 1996.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Rodrigo Canaan, Xianbo Gao, Julian Togelius, Andy Nealen, and Stefan Menzel. Generating and adapting to diverse ad hoc partners in hanabi. *IEEE Transactions on Games*, 15(2): 228–241, 2022.

Angelo Cangelosi and Domenico Parisi. Computer simulation: A new scientific approach to the study of language evolution. In *Simulating the evolution of language*, pp. 3–28. Springer, 2002.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Anti-efficient encoding in emergent communication. *Advances in Neural Information Processing Systems*, 32, 2019.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. Emergent communication at scale. In *ICLR*, 2022.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*, 2024. URL https://openreview.net/forum?id=EHg5GDnyq1.

Herbert H Clark. *Using language*. Cambridge university press, 1996.

Herbert H Clark and Susan E Brennan. Grounding in communication. In Lauren B Resnick, John M Levine, and Stephanie D Teasley (eds.), *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, Washington, DC, 1991.

Philip R Cohen and Hector J Levesque. Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261, 1990.

Stephen Cranefield, Martin Purvis, Mariusz Nowostawski, and Peter Hwang. Ontologies for interaction protocols. In *Ontologies for agents: theory and experiences*, pp. 1–17. Springer, 2005.

Tim R Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. *ICLR*, 2024.

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*, 2020.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. "robustness challenges in model distillation and pruning for natural language understanding. *EACL*, 2023.

Tom Eccles, Yoram Bachrach, Guy Lever, Angeliki Lazaridou, and Thore Graepel. Biases for emergent communication in multi-agent reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). *arXiv preprint arXiv:2505.02279*, 2025.

FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074, 2022.

Tim Finin, Richard Fritzson, Don McKay, and Robin McEntire. Kqml as an agent communication language. In *Proceedings of the third international conference on Information and knowledge management*, pp. 456–463, 1994.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, et al. Magentic-one: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468*, 2024.

Simon Garrod and Anthony Anderson. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218, 1987.

Simon Garrod and Gwyneth Doherty. Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, 53(3): 181–215, 1994.

Simon Garrod and Martin J Pickering. Why is conversation so easy? *Trends in cognitive sciences*, 8(1):8–11, 2004.

Google. Announcing the agent-to-agent protocol, 2025. URL https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability.

Barbara J Grosz. Collaborative systems (aaai-94 presidential address). *AI magazine*, 17(2): 67–67, 1996.

Barbara J. Grosz and Candace L. Sidner. Plans for discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (eds.), *Intentions in Communication*. Bradford Books, MIT Press, 1990.

Blog Team GSA. Empowering responsible ai: the ai training series for government employees, September 2024. URL https://www.gsa.gov/blog/2024/09/23/empowering-responsible-ai-the-ai-training-series-for-government-employees.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *IJCAI*, pp. 8048–8057, 2024. URL https://www.ijcai.org/proceedings/2024/890.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3): 335–346, 1990.

Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in neural information processing systems*, 30, 2017.

Horace He. Defeating nondeterminism in llm inference, September 2025. URL https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI*, 2018. ISBN 978-1-57735-800-8.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *ICLR*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International conference on machine learning*, pp. 4399–4410. PMLR, 2020.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pp. 3040–3049. PMLR, 2019.

Ece Kamar, Ya'akov Gal, and Barbara J. Grosz. Incorporating helpful behavior into collaborative planning. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, pp. 875–882, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9780981738178.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally'in multi-agent dialog. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 2962–2967, 2017.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring ai ability to complete long tasks, 2025. URL https://arxiv.org/abs/2503.14499.

Yannis Labrou, Tim Finin, and Yun Peng. The current landscape of agent communication languages. *Intelligent Systems*, 14(2):45–52, 1999.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. *ICLR*, 2017.

Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*, pp. 6187–6199. PMLR, 2021.

Huao Li, Hossein Nourkhiz Mahjoub, Behdad Chalaki, Vaishnav Tadiparthi, Kwonjoon Lee, Ehsan Moradi Pari, Charles Lewis, and Katia Sycara. Language grounded multi-agent reinforcement learning with human-interpretable communication. *Advances in Neural Information Processing Systems*, 37:87908–87933, 2024.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 286–299. IEEE, 2024.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, et al. Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139*, 2025.

Rowena Mason. All civil servants in england and wales to get ai training, June 2025. URL https://www.theguardian.com/technology/2025/jun/09/all-civil-servants-in-england-and-wales-to-get-ai-training.

Reuth Mirsky, Roni Stern, Kobi Gal, and Meir Kalech. Sequential plan recognition: An iterative approach to disambiguating between hypotheses. *Artificial Intelligence*, 260:51–73, 2018.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V Albrecht. A survey of ad hoc teamwork research. In *European conference on multi-agent systems*, pp. 275–293. Springer, 2022.

James J Odell, Harry Van Dyke Parunak, and Bernhard Bauer. Representing agent interaction protocols in uml. In *International Workshop on Agent-Oriented Software Engineering*, pp. 121–140. Springer, 2000.

Melissa Z Pan, Mert Cemri, Lakshya A Agrawal, Shuyi Yang, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Kannan Ramchandran, Dan Klein, et al. Why do multiagent systems fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Tomislav Pejsa, Dan Bohus, Michael F Cohen, Chit W Saw, James Mahoney, and Eric Horvitz. Natural communication about uncertainties in situated interaction. In *Proceedings of the 16th international conference on multimodal interaction*, pp. 283–290, 2014.

Stefan Poslad. Specifying protocols for multi-agent systems interaction. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 2(4):15–es, 2007.

Stefan Poslad and Patricia Charlton. Standardizing agent interoperability: The fipa approach. In *ECCAI Advanced Course on Artificial Intelligence*, pp. 98–117. Springer, 2001.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.

Mathieu Rita, Rahma Chaabouni, and Emmanuel Dupoux. "LazImpa": Lazy and impatient neural agents learn to communicate efficiently. In Raquel Fernández and Tal Linzen (eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 335–343, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.26. URL https://aclanthology.org/2020.conll-1.26/.

Jeffrey S Rosenschein and Michael R Genesereth. Deals among rational agents. In *Readings in Distributed Artificial Intelligence*, pp. 227–234. Elsevier, 1988.

Maayan Roth, Reid Simmons, and Manuela Veloso. What to communicate? execution-time decision in multi-agent pomdps. In *Distributed autonomous robotic systems 7*, pp. 177–186. Springer, 2006.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *CoRR*, abs/2402.07927, 2024. doi: 10.48550/ARXIV.2402.07927. URL https://doi.org/10.48550/arXiv.2402.07927.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *ICLR*, 2024.

Zijing Shi, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game. *arXiv preprint arXiv:2312.17515*, 2023.

Munindar P Singh. Agent communication languages: Rethinking the principles. *Computer*, 31(12):40–47, 1998.

Chandler Smith, Marwa Abdulhai, Manfred Diaz, Marko Tesic, Rakshit S Trivedi, Alexander Sasha Vezhnevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar A Duéñez-Guzmán, et al. Evaluating generalization capabilities of llm-based agents in mixed-motive scenarios using concordia. *NeurIPS*, 2025.

Reid G Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on computers*, 29(12):1104–1113, 1980.

Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI conference on artificial intelligence*, volume 24, pp. 1504–1509, 2010.

Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29, 2016.

Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, Hao Fu, Fangkun Zhao, Caixia Yuan, and Xiaojie Wang. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4922–4951, 2025.

Robert S. Taylor. Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3):178–194, 1968.

Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, et al. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *TMLR*, 2022.

Sarah A Wu, Rose E Wang, James A Evans, Joshua B Tenenbaum, David C Parkes, and Max Kleiman-Weiner. Too many cooks: bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. CollabLLM: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DmH4HHVb3y.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

xAI. Grok 3 beta — the age of reasoning agents, 2025. URL https://x.ai/news/grok-3.

Annie Xie, Dylan Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. Learning latent representations to influence multi-agent interaction. In *Conference on robot learning*, pp. 575–588. PMLR, 2021.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*, 2025.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):4, 2024.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piekos, Aditya A. Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanic, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. Mindstorms in natural language-based societies of mind. *Comput. Vis. Media*, 11(1):29–81, 2025. URL https://doi.org/10.26599/cvm.2025.9450460.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

Appendices

A Related work

Multi-agent collaboration spans decades of research across symbolic AI, reinforcement learning, cognitive science, and natural language processing. The main text has predominantly focused on our work's relation to studies in human collaboration due to our focus on collaboration using unconstrained natural language. Rather than attempting exhaustive coverage, this section traces four paradigms most relevant to situating our contribution: explicitly programmed rational agents, co-optimized reinforcement learning agents, ad hoc teamwork, and modern language-model agents. By tracking this progression, we highlight a historical reliance on rigid communication scaffolding and demonstrate the contribution of our unguided benchmark and extensive empirical results. We acknowledge that this framing necessarily omits worthy adjacent work and invite readers to treat this as a selective rather than comprehensive survey.

Disclosure of language model usage. We used language models to help identify potentially relevant related work and polish the prose; all cited papers were independently located, read, and verified by the authors.

Rational agents and system-wide protocols

Early works on agentic collaboration focused heavily on optimizing "system-level communication protocols" to deploy symbolic, rational agents (Smith, 1980; Rosenschein & Genesereth, 1988). These foundational efforts assumed the necessity of universal, formalized, and pre-determined communication frameworks dictated top-down by system designers. Building on the formal groundwork of how agents form joint intentions and commit to shared goals (Cohen & Levesque, 1990), researchers explicitly designed rigid communication protocols or "agent communication languages" (ACL) in an effort to enable collaboration between agents (Labrou et al., 1999). Notable examples include KQML (Finin et al., 1994), which served as a primary blueprint for creating a universal syntax and semantics for agent communication, Arcol (Bretier & Sadek, 1996), and FIPA (Poslad & Charlton, 2001), which grounded semantics through an external ontology, relying on strong assumptions such as agents being uniformly cooperative and truthful. Other approaches utilized UML (Odell et al., 2000), domain-specific ontologies (Cranefield et al., 2005), and detailed protocol specifications (Poslad, 2007) to standardize interactions.

Notably, contemporary interoperability protocols such as MCP (Anthropic, 2024), A2A (Google, 2025), and ACP (Besen, 2025) represent a renewed effort to standardize agent communication for LM-based systems, but share the same fundamental assumption that participating agents can be made to conform to a pre-specified interface.

Parallel to the engineering of communication languages, the field also explored how agents could explicitly reason about collaboration. Grosz & Sidner (1990) introduced "SharedPlans" to formalize how agents construct, communicate, and execute cooperative strategies, while others simulated emergent cooperation using simple agents (Batali, 1998; Cangelosi & Parisi, 2002) or explicitly modeled strategic helpfulness to others (Kamar et al., 2009). However, these predefined, rigid protocols faced severe limitations; constrained action spaces and explicit, programmatic reasoning around strategies struggle to transfer to heterogeneous agents or new, unanticipated environments. As noted by Singh (1998) when reviewing the state of ACLs at the time: "*Agents from different vendors – or even different research projects – cannot communicate with each other*."

Reinforcement learning agents

In a move beyond designer-specified rules, Littman (1994) introduced an important step toward more dynamic, sequential policies by framing multi-agent coordination as a reinforcement learning (RL) problem. In this paradigm, state transitions are a function of the joint action of multiple agents, making an agent's optimal policy conditional on the behavior of others, e.g., by learning dynamically when and what to communicate (Roth et al., 2006).

Deep learning built on this foundation, allowing agents to explicitly learn to cooperate (Lowe et al., 2017) and model responses to other agents (Jaques et al., 2019).

This led to an extensive body of research on emergent communication, where agents learn task-specific protocols from scratch through interaction (Foerster et al. (2016); Sukhbaatar et al. (2016); Lazaridou et al. (2017); Havrylov & Titov (2017) and many others – see Zhu et al. (2024) for a recent survey). These approaches demonstrated that effective coordination signals can emerge without human design. Despite these advances, such "learned" languages or protocols tend to suffer from several drawbacks: they are not necessarily interpretable by humans (Kottur et al., 2017), they can be inefficient in an information-theoretic sense (Chaabouni et al., 2019; Rita et al., 2020), and they typically require centralized training architectures (Eccles et al., 2019).

Furthermore, these emergent protocols are optimized for a specific task or set of tasks and co-optimized agents, inherently limiting their transferability. Scaling up to more environments has proven practically challenging (Chaabouni et al., 2022). Efforts to address this through meta-learning and training agents on multiple tasks (Team et al., 2021) have increased generalization across benchmarks like Hanabi (Bard et al., 2020), Overcooked (Carroll et al., 2019), and Melting Pot (Leibo et al., 2021). Hu et al. (2020) specifically target "zero-shot coordination" – collaborating with unseen partners without joint training – demonstrating meaningful gains on Hanabi. However, even these approaches operate within shared action spaces and reward structures, leaving the leap to truly heterogeneous partners communicating in open-ended natural language unaddressed.

Relevance to our work: RL approaches demonstrate that agents can learn to collaborate by interacting with their environment. However, because they co-optimize their communication channels for specific environments, they are not necessarily compatible with unseen partners or new environments. Our evaluation specifically targets open-world integration, testing agents' zero-shot ability to collaborate with heterogeneous models using flexible natural language rather than co-optimized, task-specific vectors.

Ad hoc teamwork

Recognizing the limits of jointly optimized agents, the field of "*ad hoc teamwork*" (AHT) emerged to study autonomous agent collaboration without prior coordination, initially introduced by Bowling & McCracken (2005) and formally defined by Stone et al. (2010). The core objective is to create an autonomous agent capable of efficiently and robustly collaborating with previously unknown teammates on tasks where all teammates can contribute. As noted in a comprehensive survey of the field by Mirsky et al. (2022), a primary research focus concerns modeling "other agents," e.g., by specifying "types" of other agents that might be encountered (Barrett et al., 2011), learning teammate representations through neural networks (Rabinowitz et al., 2018) or genetic algorithms (Canaan et al., 2022), and utilizing online (Bayesian) belief algorithms to update models of teammates dynamically (Barrett et al., 2017).

Other works explored adapting to teammates without explicit coordination protocols. This included modeling teammates based on task recognition (Wu et al., 2021), learning to model "adaptive" teammates that change their behavior (Agmon et al., 2014; Mirsky et al., 2018; Xie et al., 2021), and exploring communication with unknown teammates using limited symbols (Barrett et al., 2014).

Relevance to our work: AHT is conceptually closest to the setting motivating our work, as it demands dynamic adaptation to unknown partners. However, past work in this space often assumed either no direct communication channels (relying on implicit communication via action choices) or utilized highly restrictive, known protocols – in part because sufficiently capable language-model agents have only recently emerged. Our benchmark extends the ad hoc teamwork philosophy into the era of LMs, evaluating how agents dynamically adapt to one another using rich, unstructured natural language, without any prior coordination, shared training, or agreed-upon communication protocol. In this sense, our work can be read as an empirical test of whether modern LMs have acquired ad hoc teamwork as an emergent capability.

Language model agents

Capable language models have opened the door for agents to interact using natural language, shifting the paradigm from specialized control systems to generalized semantic actors. Early successful versions of this approach often "glued" a language model layer on top of an RL or specialized system to interpret and interact in the physical world (Ahn et al., 2022; Zitkovich et al., 2023), or to play interactive natural language games like Diplomacy (FAIR et al., 2022). Other approaches used LMs alongside RL to learn communication protocols that remain interpretable by humans (Li et al., 2024). As LMs evolved, the field moved away from systems specifically optimized for a single use case with strong action constraints and toward light scaffolding powered by LMs with frozen weights (see Guo et al. (2024) and Luo et al. (2025) for recent surveys).

Despite this inherent flexibility, there are no guarantees that two LM agents share the same reference frame. To compensate and facilitate convenient output parsing, the vast majority of current literature relies heavily on mutually known, structured communication protocols, e.g., using JSON or YAML (Ehtesham et al., 2025). For example, Shi et al. (2023) explored ad hoc teamwork in a limited role-playing game using a fixed, structured communication protocol, as well as a memory module and critic steps. Similarly, recent collaboration benchmarks enforce rigid interaction pipelines, such as multi-step critic-to-action protocols (Agashe et al., 2025), structured LM-robot hybrid pipelines (Mandi et al., 2024), and frameworks that pre-specify roles, communication protocol, and workflows (Hong et al., 2024). Even in classic benchmark extensions emphasizing forced collaboration and partial completion metrics (Sun et al., 2025), or when participants optimize single base models for collaborative tasks (Smith et al., 2025), a mutually known, structured protocol is generally assumed.

This reliance on structured protocols is understandable – fixed formats simplify output parsing and reduce failure modes in deployed systems. However, it comes at a scientific cost: it conflates task difficulty with communication overhead, making it challenging to isolate collaborative capability.

Relevance to our work: Existing collaboration benchmarks for agents powered by language models rely on (significant) structural scaffolding, which is often not ecologically plausible for open-world integration. Tasks further consist of problems that can only be completed in a team setting, making it impossible to separate an agent's static *task skill* from its dynamic *collaboration skill*. By removing format constraints and fixed protocols, our maze-solving benchmark successfully isolates and evaluates true, unguided collaborative capabilities. Our benchmark further allows for trivial complexity modulation – by increasing the maze dimensions or changing the wall density – making it broadly applicable for LMs of all strengths. Finally, our extensive evaluations covering 32 leading closed- and open-source models provide a valuable snapshot of the current state of the field.

B Prompts and models used

B.1 Models

Table 1 below shows the complete list of models evaluated in this study:

Builder	Model	Open/Closed	Distilled	System Prompt Flexible
OpenAI	gpt-5	Closed	No	Yes
	gpt-5-mini	Closed	Yes	Yes
	gpt-5-nano	Closed	Yes	Yes
	gpt-4.1	Closed	No	Yes
	gpt-4.1-mini	Closed	Yes	Yes
	gpt-4.1-nano	Closed	Yes	Yes
	o3	Closed	No	Yes
	o4-mini	Closed	Yes	Yes
	gpt-oss-120b	Open	No	Yes*
	gpt-oss-20b	Open	Yes	Yes*
Google	gemini 2.5 pro	Closed	No	No
	gemini 2.5 flash	Closed	Yes	No
	gemini 2.5 flash-lite	Closed	Yes	No
	gemma-3-27b	Open	Yes	No
Anthropic	claude opus 4.1	Closed	No	No
	claude sonnet 4.0	Closed	No	No
	claude haiku 3.5	Closed	No*	No
xAI	grok-4	Closed	No	Yes
	grok-3	Closed	No	Yes
	grok-3-mini	Closed	No*	Yes
DeepSeek	deepseek-V3	Open	No*	Yes
	deepseek-R1	Open	No	Yes
Meta	llama-4-maverick	Open	No	Yes
	llama-4-scout	Open	Yes	Yes
	llama-3.3 70B it	Open	Yes	Yes
Cohere	command-a	Open	No	Yes
	command-r	Open	No	Yes
	command-r7b	Open	Yes	Yes*
Microsoft	phi-4	Open	No	Yes
Moonshot AI	kimi-k2	Open	No	Yes
Alibaba	qwen-2.5-72B	Open	No	Yes
	qwen-3-235B	Open	No	Yes

Table 1: **Comparison of AI Models Used to Power Agents by Model Builder.** We classify models as "Open" if their weights are publicly available allowing for private deployment. We use publicly available information to annotate if a model is distilled from a larger model or not. When in doubt we add an asterisk (*). The *System Prompt Flexible* describes if (i) the model allows for a system prompt role, and (ii) if such a system prompt can be inserted at an arbitrary position. An asterisk here indicates that, although the API/model did not throw an error on (i-ii), it also did not seem to *use* more than one system prompt.

B.2 Facilitating agent-to-agent communication

System Message

```
The user will act as an intermediary between you and another agent.
The user will directly forward your messages to the other agent and vice versa.
The user will not see or modify the messages, but will relay them as is.
Messages coming from the other agent will be prefixed with "[other agent]:".
Messages coming from the user will be prefixed with "[user]:".
Do not add any additional prefixes or suffixes to your own messages.
```

Figure B.1: **System Prompt Used to Facilitate Agent-to-Agent Communication.**

Interactions with LMs are defined by a history of "message" objects. At the minimum, each message consists of its content, content type, and a "role" describing the source and/or function of the message. Most modern LMs are designed to recognize three types of messages:

- **System Message**: High-level instruction(s) that define the AI's persona, rules, and context.
- **User Message**: Content submitted by the user for consideration by the AI.
- **Assistant Message**: Content generated by the AI in response to User Message(s).

Some of the more popular APIs only accept a single System message, placed either at the starting position of the message history or outside of the message history entirely (See Table 1). It thus often becomes impossible to use the System role as an independent "third party" to a conversation. Many APIs further do not support multiple Assistant messages in sequence. Nor do they allow the first message in a history to have the Assistant role.

In order to facilitate dialogue between two AI agents in a way that is supported across all API providers, we thus have three options:

1. User as agent. The most straightforward implementation is to simply use the User role when relaying messages between two agents. That is, each agent receives messages from the other agent "as if" the other agent was a user. Unfortunately, this comes with two complications: First, LMs are aligned to behave in a particular way when interacting with users. This can lead, among other things, to a well-reported "sycophancy" tendency in some models (Sharma et al., 2024). For example, they can become too agreeable. This can hamper simulations meant to evaluate AI–AI interactions instead of human–AI ones. Secondly, because some providers do not allow the first message in a history to have the Assistant role, one has to resort to a "filler" User message to generate rollouts. As the model is made to believe it generated these, it is unclear what their influence is on the subsequent rollout.

2. Transcript style. Given a history of messages, we can transform it into a single User message "summarizing" the dialogue as a transcript between the two agents (Davidson et al., 2024). While this circumvents the issues around having a starting User message and mitigates potential sycophancy tendencies, it likely pushes a model out of its "regular" distribution. That is, models are optimized for sequential messages being sent back-and-forth – not contributing lines to a script.

3. Explicit prefixes. In this work we combined insights from both (1) and (2) above as follows: First, we use a System prompt to describe that the model is about to engage with another model and that the user will act as an intermediary. It goes on to define "prefixes" to distinguish between messages coming from the user, "`[user]:`", and those coming from the other agent "`other agent:`". Next, we use a User message with the user prefix to provide instructions for the task (see B.4). This ensures models are always aware that they are conversing with another model and that the first message in the history is always of type User. At each turn, we can simply convert all messages except the first one from type User to Assistant and vice versa, making sure User messages always contain the relevant prefix.

B.3 Solo prompts

User Message

```
# Task
You are going to play a maze game. The maze consists of a grid with walls and a goal. Your task is
to navigate through the maze, avoiding walls, to reach the goal.

## Rules
- You can only move to adjacent cells (up, down, left, right).
- You can not move diagonally or through walls.
- The game ends when you have reached the goal or if you step into a wall.
- **Your game is timed for speed and put on a leader board so hurry up!**
- You get one shot to complete the maze, so make sure to plan your moves carefully.
- Provide both your proposed complete solution and any considerations you have made in your reasoning.
- Provide your final solution as a list of coordinates.


## Map
Two maps of the maze are provided below, each showing a different portion of the same maze. Here are
the two maps of the same maze with a legend of the symbols:

Map view 1:

@.???
.?..?
#???.
?...?
.??#*

Map view 2:

@?...
?.??#
?#..?
#???.
?##?*

Legend:

@ - Current Position
* - Goal Position
. - Path
# - Wall
? - Hidden Cell

Time to get started! Please begin.
```

Figure B.2: **Prompt Used for Solo Experiments.** Example of a "distributed" information prompt, containing two incomplete map copies.

User Message

```
Carefully review the final solution you have provided above.
Ensure that it is a complete and valid path through the maze.
If needed, you can make adjustments to the path to ensure it meets the requirements of the maze game.
Remember to consider the maze's rules and constraints while reviewing your path.
After you have finished reviewing, please submit your final solution.
```

Figure B.3: **Critic Prompt.** Models are given a "critic" step to improve on their initial solo solution.

B.4 Collaboration prompt

User Message

```
# Task
You are going to play the collaborative maze game "ACTI" together with another agent.  The maze
consists of a grid with walls and a goal. Your task is to navigate through the maze, avoiding walls,
to reach the goal.  You will each get your own map of the **same** maze, with some coordinates
hidden.  Because of the hidden coordinates, you will need to communicate with the other agent to
share information about the maze and coordinate your movements.

## Rules
- You can only move to adjacent cells (up, down, left, right).
- You can not move diagonally or through walls.
- The game ends when you have reached the goal or if you step into a wall.
- **Your game is timed for speed and put on a leader board so hurry up!**
- You can only take one action per turn.
- An action is **immediately** considered executed once both agents agree.
- Be sure to say: “ACTI!”, after you reached the goal.

## Map
Here is **your** map of the maze with a legend of the symbols:

@.???
.?..?
#???.
?...?
.??#*

Legend:

@ - Current Position
* - Goal Position
. - Path
# - Wall
? - Hidden Cell

Time to get started! Please begin.
```

Figure B.4: **Prompt Used for Collaboration Experiments.** Note the rules agents must follow and the predefined completion phrase “ACTI!.”

B.5 Verification prompts

User Message

```
# Task
You will be given the proposed solution to a maze game. Your task is to extract the proposed route by
the agent for the maze. The agent might consider multiple coordinates at a time, but only a single
coordinate in its final move or proposed solution. Record each coordinate of the **final proposed
route** in a separate entry.

Further, you should also record the route schema, which describes the format of the coordinates used
in the route. To the best of your ability, you should determine the schema based on the dialogue.

The schema should include:
- maze_origin: The origin of the maze, i.e., is the agent treating 0 or 1 as the origin for the maze

- maze_orientation: The orientation of the maze:
"top_left", i.e., if the top left corner is (0, 0)
"bottom_left", i.e., if the bottom left corner is (0, 0)
"top_right", i.e., if the top right corner is (0, 0)
"bottom_right", i.e., if the bottom right corner is (0, 0)

- coordinates_orientation: The orientation of the coordinates:
"row_col", if the first coordinate is the row and the second is the column,
"col_row", if the first coordinate is the column and the second is the row.

- coordinates_symbols: The symbols used to represent coordinates:
"number_number", e.g., (1, 2)
"letter_letter", e.g., (A, B)
"letter_number", e.g., (A, 1)
"number_letter", e.g., (1, A)
"directions", e.g., "up", "down", "left", "right"

Respond in the following format:

“‘yaml
route_schema:
maze_origin: "<0|1>"
maze_orientation: "<top_left | bottom_left | top_right | bottom_right>"
coordinates_orientation: "<row_col | col_row>"
coordinates_symbols: "<number_number | letter_letter | letter_number | number_letter | directions>"
route:
- turn: <turn_number>
coordinates: [[<coordinate_1>, <coordinate_2>], ..]
turn_type: "move" # only consider final moves, not intermediate considerations
agent: "agent_1"
- turn: <turn_number>
...
“‘

Do not include any other information, just the YAML object.

# Dialogue
<insert dialogue>
```

Figure B.5: **Verification Prompt for Solo Experiments.**

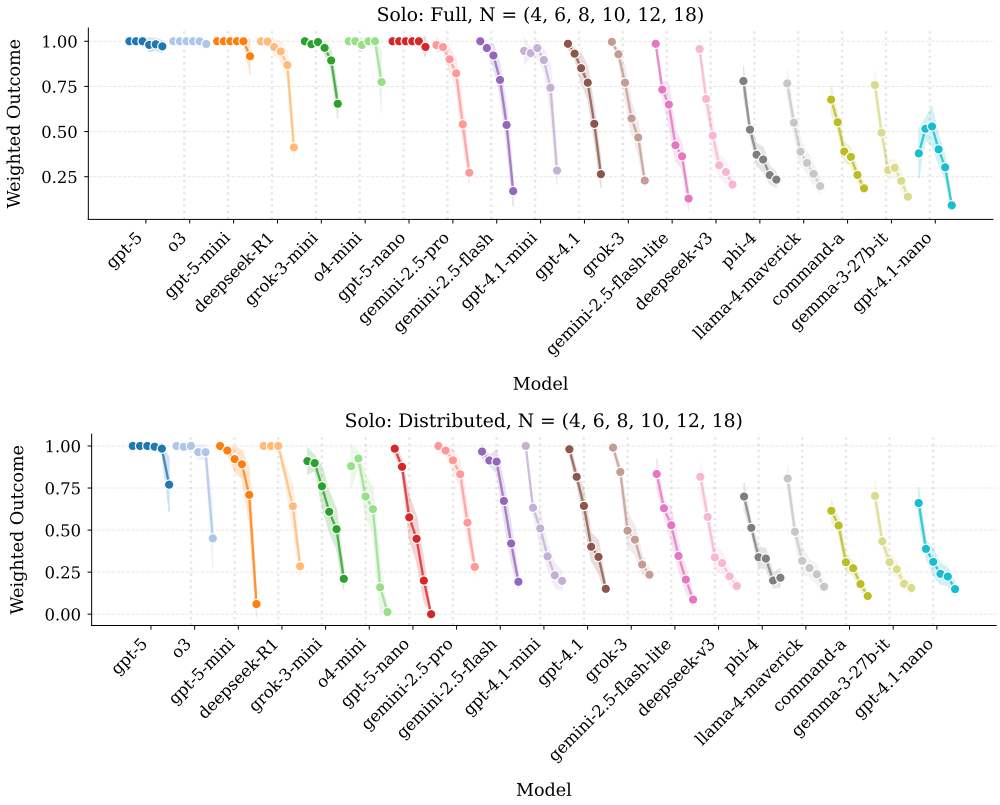
User Message

```
# Task
You will be given a dialogue between two agents trying to solve a maze. Your task is to extract the
route taken by the agents through the maze at each turn. Some turns, the agents will not move, and
some turns they will. The agents may communicate moves using directions, e.g., "up", "down", "left",
"right". Or the agents may communicate coordinates directly, e.g., "Let's move to (3, 4)". For
turns where both agents agreed to move, record the coordinates or direction of the cell they agreed
to move to. For turns where they did not move, record the coordinates of the cells or directions
they considered, if any. Make sure to record what happened at **each turn**. You must be consistent
in the move format you record, e.g., stick with directions or coordinates, do not mix them unless
absolutely necessary.

Further, you should also record the route schema, which describes the format of the coordinates used
in the route. To the best of your ability, you should determine the schema based on the dialogue.

The schema should include:
- maze_origin: The origin of the maze, i.e., are the agents treating 0 or 1 as the origin for the
maze

- maze_orientation: The orientation of the maze:
"top_left", i.e., if the top left corner is (0, 0)
"bottom_left", i.e., if the bottom left corner is (0, 0)
"top_right", i.e., if the top right corner is (0, 0)
"bottom_right", i.e., if the bottom right corner is (0, 0)

- coordinate_orientation: The orientation of the coordinates:
"row_col", if the first coordinate is the row and the second is the column,
"col_row", if the first coordinate is the column and the second is the row.

- coordinate_symbols: The symbols used to represent coordinates:
"number_number", e.g., (1, 2)
"letter_letter", e.g., (A, B)
"letter_number", e.g., (A, 1)
"number_letter", e.g., (1, A)
"directions", e.g., "up", "down", "left", "right"

Respond in the following format:

“‘yaml
route_schema:
maze_origin: "<0|1>"
maze_orientation: "<top_left | bottom_left | top_right | bottom_right>"
coordinates_orientation: "<row_col | col_row>"
coordinates_symbols: "<number_number | letter_letter | letter_number | number_letter | directions>"
route:
- turn: <turn_number>
coordinates: [[<coordinate_1>, <coordinate_2>], ..]
turn_type: "<move|consider>"
agent: "<agent_1|agent_2>|both"
- turn: <turn_number>
...
“‘

Do not include any other information, just the YAML object.

# Dialogue
<insert dialogue>
```

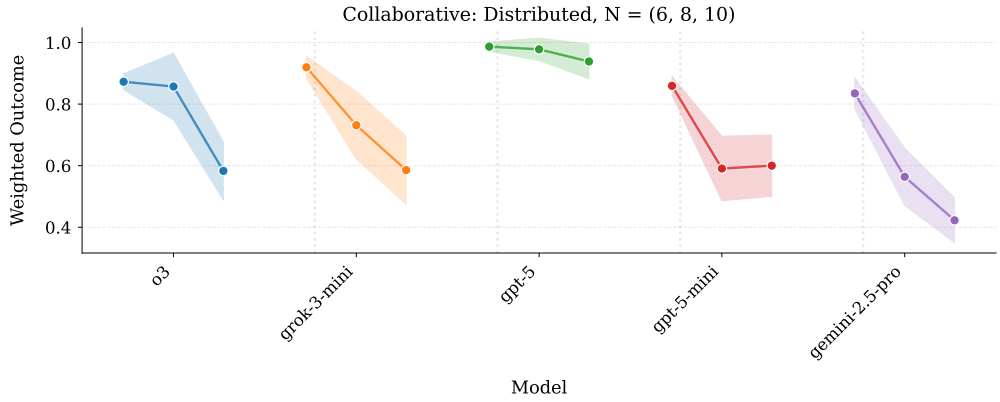
Figure B.6: **Verification Prompt for Collaboration Experiments.**

C Maze hyperparameter ablations

C.1 Varying maze size



(a) **Solo: Full and Distributed.** For each model, we report solo mean performance with 95 % CI on $N \times N$ mazes where $N \in \{4, 6, 8, 10, 12, 18\}$. We note that for most models performance drastically drops in the distributed setting for $N > 6$. Both gpt-5 and o3 perform remarkably well up to $N = 12$.



(b) **Homogeneous Collaboration.** We ablate collaborative performance for selected strong performers on $N \times N$ mazes where $N \in \{6, 8, 10\}$. Note that performance drops considerably for all models except gpt-5.

Figure C.1: **Weighted Outcomes for Maze Size Ablations.**

C.2 Varying wall density

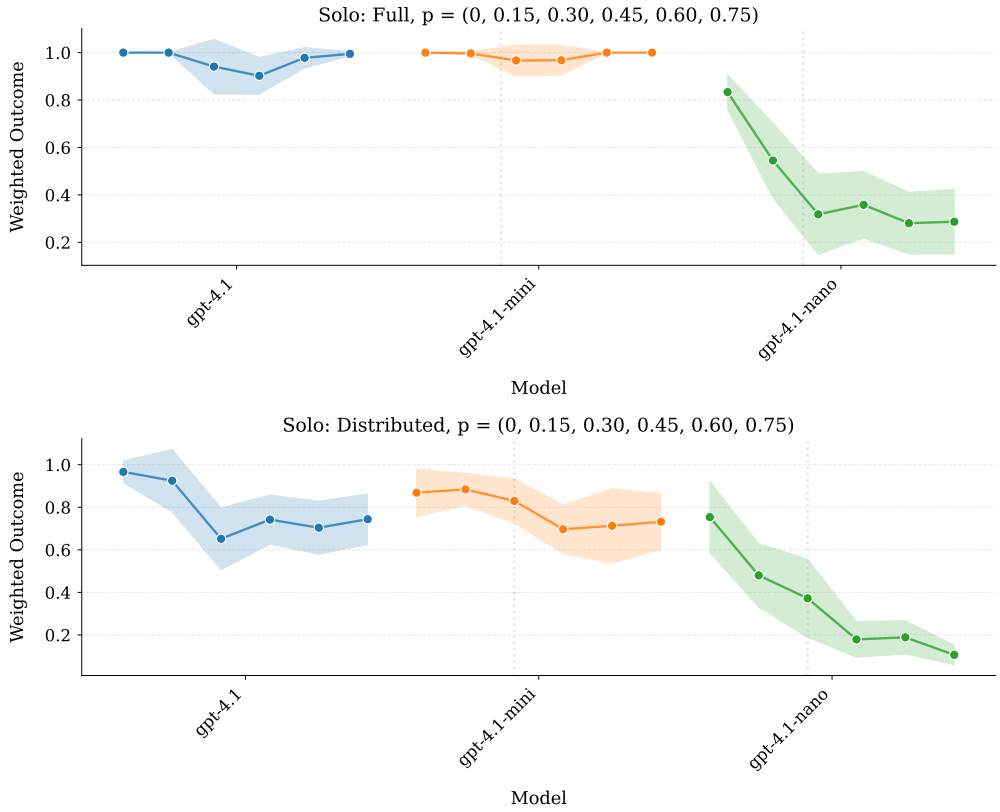


Figure C.2: **Solo Weighted Outcomes for Wall Density Ablations.** We report mean values with 95% CI on 6×6 mazes for ∼30 rollouts, varying the wall density $p \in \{0, 0.15, 0.30, 0.45, 0.60, 0.75\}$. Note that for the solo full setting (top), gpt-4.1 and gpt-4.1-mini display the most variance in the range $p \in [0.30, 0.45]$, whereas the gpt-4.1-nano performance monotonically degrades as wall density increases. For the distributed setting, all three models drop performance as the wall density increases.

D Additional results

In this section we start by highlighting a number of model-specific qualitative observations (App. D.1). We then attempt to quantify some of the mechanisms likely to attribute to the identified collaboration gap in App. D.2. Next we show additional weighted and binary outcome results for various interaction pairs in Appendices D.3 and D.4 respectively.

We close this section by exploring "collaborative efficiency" in App. D.5, which can be defined as the product of the median tokens per message and the number of messages required per solution. Because "thinking" tokens are not always exposed, this metric reflects information-sharing efficiency rather than computational efficiency. We report the median message counts conditioned on weighted-outcome ranges to mitigate outcome confounding. Finally, we focused our heterogeneous pairings on the most widely deployed commercial APIs to establish a baseline, leaving open-weights heterogeneous pairings for future work.

D.1 Qualitative observations on specific models

command-a/r. Upon qualitative inspection, we noticed that the command-a/r models had a tendency to prematurely "hallucinate" about being able to use the completion phrase. As a result, their collaborative capabilities reported here should be seen as a lower bound on actual performance.

gpt-5. This model displayed superior performance in both solo and homogeneous collaboration mode. As shown in Figure C.1, it remained capable of solving both full and distributed mazes up to size 12×12 perfectly, and barely lost performance in homogeneous collaboration for mazes of size 8×8 and 10×10. As shown in Figure G.1, the "failed" runs in 6×6 mode were due to a parsing shortcoming, rather than a mistake made by the model.

gemini-2.5-pro. Successful models generally share their maze information in the early turns to resolve the missing information. They further engage in rich grounding to make sure their partner is on the same page. Instead, many of gemini-2.5-pro's rollouts were marked by short messages (see Figure D.4). While this was not a problem when collaborating in homogeneous mode, it did complicate collaborating with weaker models in the heterogeneous setting, e.g., gemini-2.5-flash-lite. This can be seen in Table 4, where the weaker but more verbose gemini-2.5-flash is capable of reaching higher outcomes when collaborating with gemini-2.5-flash-lite compared to its pro base.

claude-sonnet-4. The model proved itself better at collaborating with other models than with a copy of itself (see Figure 4.3). Qualitative inspection showed two interesting failure modes: (i) the model got stuck in "agreement" loops, where, instead of proposing the next move, it keeps seeking agreement on the previous move; (ii) it "splits" up and starts to explore the maze separately. This is not necessarily incorrect, as this could be interpreted as under specified in the instructions. It does however break the current parser, leading to suppressed performance (see Figure G.2).

claude-opus-4.1. While a strong average performer, some of its solutions were surprisingly inefficient. For example:

```
@ # - - - -
o o o o o o
o # - o o o
# o o o # o
# o # o # *
# o o o # -
```

where "-" are paths, "o" cells visited, "#" walls, and "@, *" the start and goals states. It manages to visit almost every path cell on its way to the goal state, using the full 50 turns.

Smaller models Many of the smaller models had a tendency to not share their map view at all, e.g., gpt-4.1-mini, claude-haiku-3.5, opting for a completely local exploration, asking for information as they went instead (see App. D.2.1). They also often used "directions" to communicate moves instead of coordinates, making it increasingly difficult to track state.

D.2 Quantifying collaboration mechanisms

The experimental results in Section 4 identify a "collaboration gap," where models perform reliably worse when forced to collaborate compared to solving mazes solo. Although we highlighted some of the mechanisms associated with this gap based on qualitative inspections, e.g., style imitation, lack of grounding, disagreement resolution strategies, we did not quantify these mechanisms. In this section we therefore dive deeper into quantifying specific behavioral tendencies and their effect on outcomes.

Note that this is **not** an exhaustive analysis, but rather a starting point for future work aimed at improving collaborative capabilities. All features used in these analyses are annotated by LMs, introducing a shared source of measurement noise. While we validate annotation consistency (see below), systematic misreadings could affect all features in correlated ways. Furthermore, all analyses are observational; we use the language of mediation and prediction throughout, but caution against strong causal interpretations.

Annotation setup. For a selection of 53 unique homogeneous and heterogeneous interaction pairs, we sample 50 rollouts per pair. These rollouts involve 13 models from Anthropic, Google, OpenAI, and xAI, chosen to cover different capability substrates, model families, and interaction types. We used the results from Section 4.1 to assign strength tiers:

- **Strong**: grok-3-mini, grok-4, gemini-2.5-pro, gpt-5, claude-opus-4.1
- **Medium**: grok-3, gemini-2.5-flash, gpt-4.1, claude-sonnet-4
- **Weak**: gemini-2.5-flash-lite, gpt-5-nano, gpt-4.1-mini, claude-haiku-3.5

We then use two LMs (gpt-4.1 and gemini-2.5-flash) to annotate each rollout for a variety of features, which are the result of iterative manual inspection aided by model suggestions. Relevant features will be discussed in each subsection.

D.2.1 Grounding cascades predict outcome success

Following the discussion in section 2.2, grounding is deemed essential in human collaboration. We therefore collect the following features to assess if models attempt to "ground" their representations of the task and their strategic approach:

- **grounded start and goal**: did models align on the start and goal state?
- **shared map**: did both models share their map?
- **merged map**: did models attempt to "merge" the partial maps?
- **reconciliation timing**: were maps merged up front, gradually, or never?
- **strategy**: did agents use a global strategy, i.e., planning more than 3 moves ahead, or use a strictly local approach?

Whether agents use a global strategy or not is strongly correlated with weighted outcomes, with global strategies achieving an average weighted outcome of 0.75 ± 0.01 and local ones just 0.47 ± 0.01. The use of a global strategy is in turn correlated with the strength of the participating models, with global strategies more likely when strong models are involved (see Figure D.2 (a)). However, a global strategy is only possible once certain "grounding" steps have been taken, i.e., at the minimum one of the agents needs access to both maps. Hence, we are interested in the influence of grounding both in enabling superior strategies and influencing outcomes directly (see Figure D.1).

As a proxy for grounding strength, we compute a simple grounding score. Let $\mathcal{B}$ be the set of grounding features {start+goal, share, merge} and $R \in \{0, 1, 2\}$ the pair-level reconciliation

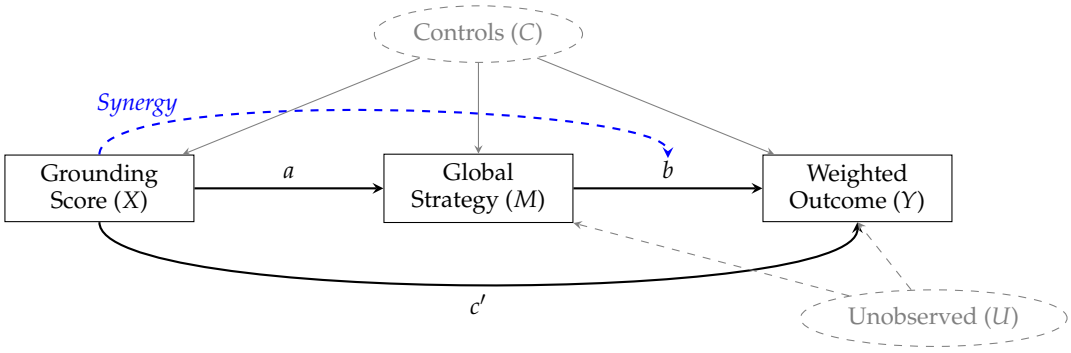


Figure D.1: **Path diagram of mechanism influences.** Our goal is to measure the influence of the observed "grounding behavior" (X) and the choice of strategy type (M) on the outcome (Y), controlling for a number of factors (C) and testing for the robustness against potentially unobserved factors (U). We test both the direct influence c', and the mediated influence of X through M ($a \to b$). We also explore a "Synergy" interaction term to measure if the effectiveness of M is contingent on the strength of X.

timing score (2 for upfront, 1 for incremental), then

$$X = \frac{1}{8}\left(R + \sum_{i\in\{1,2\}}\sum_{j\in\mathcal{B}} \mathbb{I}_{i,j}\right), \tag{1}$$

where $\mathbb{I}_{i,j}$ is an indicator variable that equals 1 if agent i performs behavior j. The denominator 8 normalizes the score to the range $[0, 1]$. This score empirically shows a statistically significant association with weighted outcomes across the 53 pairs (Pearson $r = 0.69$, Spearman $\rho = 0.70$, for both $p < 0.01$) as shown in Figure D.2 (b).

We start by fitting a simple additive mediation model to test whether grounding behavior is consistent with establishing the "pre-conditions" for strategic planning, which in turn influences weighted outcomes (Model 1):

$$M = i_1 + aX + \gamma_1 C + \epsilon_1 \tag{2}$$

$$Y = i_2 + c'X + bM + \gamma_2 C + \epsilon_2, \tag{3}$$

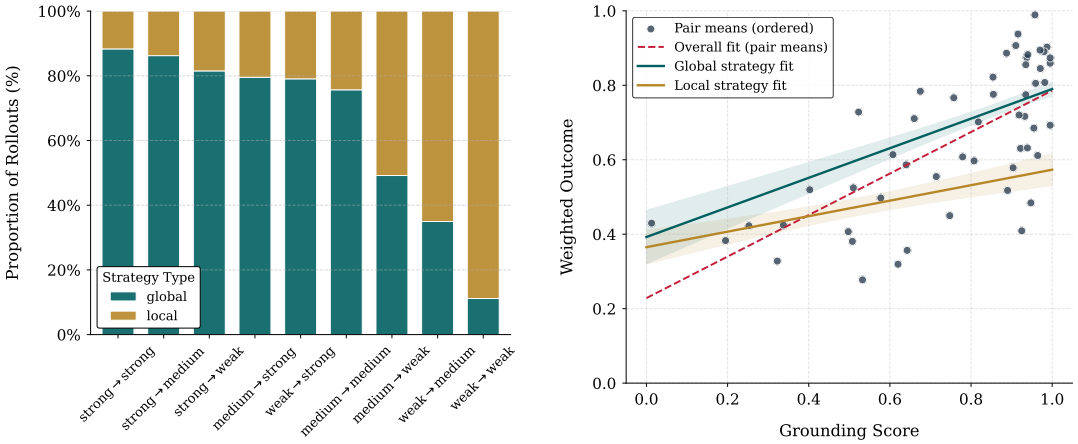
where a measures how much grounding score X increases the likelihood of adopting a global strategy M, b measures the increase in weighted outcome Y provided by the strategy holding grounding constant, and $a \times b$ is the estimated indirect effect (following Imai et al. (2010)), representing the portion of the grounding-outcome association that is consistent with mediation through the strategy. Finally, C is a vector of controls accounting for agent family, agent strength, interaction type, maze complexity, and the median token usage. i and ϵ are intercepts and error terms respectively.

The second model includes a "synergy" interaction term ($X \cdot M$) to test if the strength of M is moderated by the strength of X (Model 2):

$$Y_{syn} = i_3 + c'X + b_1 M + b_2(X \cdot M) + \gamma_3 C + \epsilon_3 \tag{4}$$

Results of clustered bootstrapping on agent pairs ($B = 1000$) are shown in Table 2. Model 1 shows robust additive mediation, with grounding score X significantly associated with the propensity of using a global strategy M ($a \approx 0.69$, $p < 0.001$). Opting to use a global strategy boosts performance ($b \approx 0.17$, $p < 0.001$) with approximately $100 \cdot (a \times b)/c \approx 38.4\%$ of the total effect of grounding mediated through the strategic pathway (Indirect effect ≈ 0.12, CI: $[0.07, 0.16]$).

Incorporating an interaction term in Model 2 reveals a more nuanced relationship between grounding and strategy: the effectiveness of a global strategy appears contingent on grounding levels. The non-significant and slightly negative strategy path ($b \approx -0.04$, at $X = 0$,



(a) Strategy by interaction-pair strength

(b) Weighted outcome vs. grounding score

Figure D.2: **Mechanisms influencing the collaboration gap.** (a) Proportion of global versus local strategies across interaction pairs stratified by model strength. The use of global strategies increases when stronger models are involved. (b) Relationship between grounding scores and weighted outcomes. While both strategies improve with grounding, global strategies (green) show a steeper performance curve. Red dashed line represents the aggregate trend across all ordered agent pairs; shaded regions indicate 95% CIs.

Effect Path	**Model 1 (Additive)**	**Model 2 (Synergistic)**
Path a (Grounding $\rightarrow$ Strategy)	0.6886^{***}	0.6886^{***}
Path b (Strategy Effect)	0.1683^{***}	-0.0421
Synergy ($X \cdot M$)	—	0.2778^{***}
Direct Effect (c')	0.1859^{***}	0.1859^{***}
Total Effect (c)	0.3018^{***}	0.3018^{***}
Indirect Effect ($a \times b$)	0.1159^{***}	$-0.0290^{\dagger}$
Bootstrap 95% CI	$[0.0679, 0.1587]$	$[-0.1019, 0.0309]$

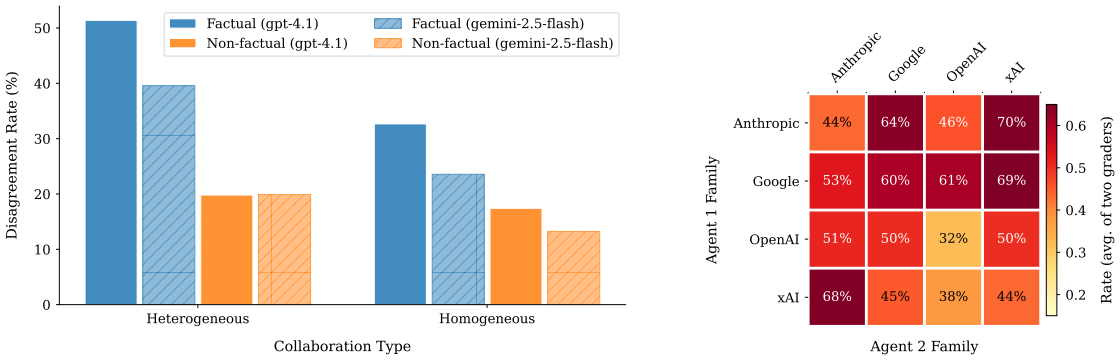
$^{***}p < 0.001$; †Indirect effect in Model 2 is calculated at the grounding intercept ($X = 0$).

Table 2: **Mediation analysis of grounding and global strategy on weighted outcomes.** Model 1 shows the baseline additive mediation and Model 2 incorporates the interaction (Synergy) between grounding and strategic choice. Standard errors are clustered by model pair (N = 53).

CI: $[-0.10, 0.03]$) indicates that planning without any grounding provides no performance benefit. This is not surprising, since planning a global route without access to both maps would indicate either hallucinations or a high-risk gamble. However, the strong Synergy coefficient ($X \times M \approx 0.28$, $p < 0.001$) indicates that as grounding level *increases*, the global strategy effectiveness scales with it. At perfect grounding ($X = 1$), this boost reaches $b_1 + (b_2 \cdot X) \approx 0.24$, effectively doubling the impact observed in the baseline model.

We further conduct a sensitivity analysis following the framework of Imai et al. (2010) to verify that the identified impact is not an artifact of unobserved confounders missed by our controls ($\mathcal{U}$ in Figure D.1). For Model 1, the mediation effect remains positive and significant for a residual correlation of $\rho < 0.18$ and for Model 2 at $\rho < 0.26$ when evaluated at $X = 1$. These results suggest that the associations between grounding, strategy, and weighted outcome are fairly resilient to potential omitted-variable bias, though we note that the sensitivity thresholds are moderate: a single moderately correlated unobserved confounder could attenuate the mediated pathway.

To assess annotation stability, we first re-graded a random subsample of 250 rollouts using three alternative models (gpt-4.1, gemini-2.5-flash, and gpt-5-mini). The composite



(a) Disagreement by interaction type

(b) Encountered disagreement

Figure D.3: **Disagreements by interaction type and between model families.** (a) Homogeneous collaborations exhibit a lower rate of (factual) disagreements ($p < 0.01$). (b) Models from the same model family have a slightly lower rate of disagreements among themselves and Google models (i.e., Gemini series) are the most combative.

grounding score shows strong inter-grader agreement across all three pairwise comparisons (Pearson $r \in [0.71, 0.79]$). Agreement on individual binary components ranges from moderate to substantial (Cohen's $\kappa \in [0.39, 0.83]$). We then repeated the full grounding mediation analysis using annotations from gemini-2.5-flash on the entire sample. All mediation paths replicate with the same sign and significance (path $a \approx 0.48$, $p < 0.001$; path $b \approx 0.09$, $p < 0.001$; indirect effect CI 95% $[0.01, 0.06]$; synergy $p = 0.018$), with coefficient attenuation consistent with the inter-grader agreement reported above.

Limitations. The grounding score (1) assigns equal weight to each component and only covers a subset of grounding features. Learned or informed weightings might be more appropriate. The model strength is assigned by the authors; results could shift under alternative groupings. Finally, despite the sensitivity analysis, all results remain observational and should thus be interpreted as predictive patterns rather than causal mechanisms.

D.2.2 When and what do agents disagree on?

In this section we examine what types of disagreements agents encounter. We differentiate between factual disagreements and those based on preference. The former involves *perceptual disagreements*, where agents disagree on the contents of a cell or the *physical legality* of moves, e.g., the disagreement in Figure 4.5. The latter can be the result of grounding misalignment, e.g., the col–row vs. row–col example of Figure 2.2, or *strategy*, where agents disagree on the direction of a move, without a factual conflict. Disagreements occurred in approximately half of the annotated rollouts.

We annotate the following features:

- **has factual disagreement**: any disagreements based on perception or move legality
- **has non-factual disagreement**: any disagreements based on grounding or strategy

To improve robustness we take the average between two grader LMs (gpt-4.1 and gemini-2.5-flash). The two graders show high agreement on whether disagreement was encountered ($\kappa = 0.72$) and if there was a factual disagreement ($\kappa = 0.70$). They are moderately aligned on the occurrence of non-factual disagreement ($\kappa = 0.48$), which is not surprising as non-factual disagreements like grounding and strategy are more nuanced and open for interpretation.

To quantify the drivers of agentic conflict, we model the likelihood of disagreement events using a logistic regression

$$Y = i + \beta_0 H + \beta_1 F + \sum_{n \in \{1,2\}} \sum_{k \in \mathcal{K}} \beta_{nk} C_{nk} + \epsilon, \tag{5}$$

where Y is the disagreement event, H indicates if this concerns a homogeneous interaction, F if both models were from the same model family, and i and ϵ are intercept and error terms respectively. The other predictors in $\mathcal{K}$ include the strength and model family of agent 1 and agent 2.

Predictor	Total Disagreement	Factual	Non-Factual
Intercept (Baseline)[a]	1.96*	3.43**	0.44**
Interaction-level Effects			
Identical Model (H)	0.60*	0.66**	0.97
Same Family (F)	0.83	0.92	0.97
Agent 1 Characteristics (C_{1k})			
Weak Starter	0.69†	0.60**	1.13
Strong Starter	0.93	0.59**	1.44†
Google Family	1.30	1.91**	0.88
OpenAI Family	0.81	0.68†	1.11
xAI Family	0.69	1.07	1.57
Agent 2 Characteristics (C_{2k})			
Weak Follower	0.88	1.06	1.30
Strong Follower	1.16	1.02	1.61*
Google Family	1.12	1.50†	0.77
OpenAI Family	0.89	0.91	0.66
xAI Family	0.98	1.11	0.58
Observations (N)	2628	1472	1472

[a] *Note:* Reference group is Medium Starter, Medium Follower, Anthropic Family. Significance determined via cluster bootstrap ($B = 1000$) resampled by unique model pairs. Significance: ** $p < 0.01$, * $p < 0.05$, $^{\dagger}p < 0.10$.

Table 3: **Predictors of agentic conflict.** Odds ratios for disagreement types.

Results of clustered bootstrapping on agent pairs ($B = 1000$) are shown in Table 3. We first note that the odds ratio of encountering any conflict at all are significantly lower for homogeneous collaborations (OR = 0.60, $p < 0.05$). Strong models are just as likely as medium models (reference group) to encounter disagreements, while weaker models appear more compliant (OR = 0.69, $p < 0.10$).

Factual conflicts are the most likely in medium-strength models. Strong starting models reduce factual disagreement likelihood by roughly 41% (OR = 0.59), which could be linked to their superior grounding capabilities leaving less room for confusion. Weak starting models reduce the likelihood of disagreement by a similar rate. This could be explained by their lower propensity for disagreement in general. Homogeneous collaborators are also less likely to disagree on facts (OR = 0.66), whereas Google starters are the most combative — nearly doubling the odds of a factual dispute (OR = 1.91).

Non-factual conflicts increase with capability, suggesting a "sophistication shift" in agent conflict. Strong followers are 61% more likely (OR = 1.61) to trigger disputes over strategy or grounding than the medium reference, with strong starters not far behind (OR = 1.44). Crucially, shared model identity mainly reduces how often conflicts arise, not what they are about: once a disagreement does occur, homogeneous pairs are just as likely as heterogeneous pairs of comparable strength to be disputing strategy or grounding (OR = 0.97).

Limitations. As noted before, identifying disagreement types is challenging, reflected in the inter-grader agreement ($\kappa = 0.70, \kappa = 0.48$). Directional findings are likely robust, but precise odds ratios for non-factual conflicts should be interpreted with caution. The analysis further clusters by 53 unique model pairs with some family $\times$ family pairings sparsely populated, such that odds ratios significant at $p < 0.05$ could be sensitive to individual pairs.

D.3 Weighted outcome

	g-2.5-pro	g-2.5-flash	g-2.5-flash-lite	gpt-4.1	gpt-4.1-mini
g-2.5-pro	**0.84** ± **0.06**	**0.77** ± **0.10**	0.54 ± 0.11	-	-
g-2.5-flash	**0.85** ± **0.09**	0.76 ± 0.05	**0.71** ± **0.06**	**0.63** ± **0.09**	**0.58** ± **0.08**
g-2.5-flash-lite	0.58 ± 0.11	**0.61** ± **0.07**	0.36 ± 0.02	0.36 ± 0.07	0.32 ± 0.07
gpt-4.1	-	**0.56** ± **0.09**	0.35 ± 0.07	0.55 ± 0.04	0.50 ± 0.04
gpt-4.1-mini	-	**0.47** ± **0.08**	0.28 ± 0.06	0.42 ± 0.04	0.39 ± 0.01

(a) Google: (g)emini, OpenAI: (gpt)

	gpt-5	gpt-5-mini	gpt-5-nano	o3	gpt-4.1	gpt-4.1-mini
gpt-5	**0.99** ± **0.02**	**0.95** ± **0.05**	**0.82** ± **0.09**	-	-	-
gpt-5-mini	**0.95** ± **0.04**	0.86 ± 0.03	0.65 ± 0.11	0.84 ± 0.05	0.71 ± 0.05	0.66 ± 0.05
gpt-5-nano	**0.73** ± **0.10**	0.45 ± 0.11	0.38 ± 0.04	-	-	-
o3	-	0.85 ± 0.05	-	**0.88** ± **0.03**	**0.76** ± **0.05**	**0.77** ± **0.03**
gpt-4.1	-	**0.76** ± **0.05**	-	0.76 ± 0.05	0.55 ± 0.04	0.50 ± 0.04
gpt-4.1-mini	-	0.58 ± 0.06	-	**0.62** ± **0.05**	0.42 ± 0.04	0.39 ± 0.01

(b) OpenAI

	grok-4	grok-3-mini	grok-3	gpt-4.1	gpt-4.1-mini
grok-4	**0.92** ± **0.05**	**0.94** ± **0.05**	**0.87** ± **0.06**	-	-
grok-3-mini	**0.91** ± **0.07**	0.90 ± 0.04	0.77 ± 0.11	**0.82** ± **0.06**	**0.74** ± **0.11**
grok-3	**0.81** ± **0.08**	0.72 ± 0.10	0.41 ± 0.04	0.37 ± 0.11	0.42 ± 0.12
gpt-4.1	-	**0.66** ± **0.08**	0.47 ± 0.10	0.55 ± 0.04	0.50 ± 0.04
gpt-4.1-mini	-	0.42 ± 0.11	0.16 ± 0.07	**0.42** ± **0.04**	0.39 ± 0.01

(c) xAI, OpenAI

	grok-3-mini	gemini-2.5-flash	claude-sonnet-4	gpt-4.1
grok-3-mini	**0.90** ± **0.04**	**0.88** ± **0.06**	**0.90** ± **0.05**	**0.82** ± **0.06**
gemini-2.5-flash	0.83 ± 0.06	0.76 ± 0.05	**0.86** ± **0.05**	0.68 ± 0.06
claude-sonnet-4	**0.76** ± **0.05**	0.68 ± 0.06	0.48 ± 0.05	0.54 ± 0.07
gpt-4.1	**0.66** ± **0.08**	0.63 ± 0.07	0.61 ± 0.07	0.55 ± 0.04

(d) Anthropic, Google, OpenAI, xAI

	claude-opus-4.1	claude-sonnet-4	claude-haiku-3.5
claude-opus-4.1	**0.85** ± **0.04**	**0.69** ± **0.10**	**0.58** ± **0.10**
claude-sonnet-4	**0.63** ± **0.09**	0.48 ± 0.05	0.45 ± 0.10
claude-haiku-3.5	**0.55** ± **0.08**	0.38 ± 0.07	0.41 ± 0.06

(e) Anthropic

	command-a	command-r	command-r7b
command-a	**0.38** ± **0.04**	**0.34** ± **0.07**	**0.20** ± **0.11**
command-r	**0.33** ± **0.06**	0.20 ± 0.02	0.15 ± 0.06
command-r7b	**0.25** ± **0.07**	0.13 ± 0.04	0.16 ± 0.04

(f) Cohere

Table 4: **Weighted Outcomes for Collaborative Performance.** We report mean performance with 95% CI. Max values per row/column are bold.

D.4 Binary success rate

	g-2.5-pro	g-2.5-flash	g-2.5-flash-lite	gpt-4.1	gpt-4.1-mini
g-2.5-pro	**0.74 ± 0.09**	**0.70 ± 0.13**	0.34 ± 0.13	-	-
g-2.5-flash	**0.78 ± 0.12**	0.64 ± 0.07	**0.53 ± 0.08**	**0.49 ± 0.10**	**0.40 ± 0.10**
g-2.5-flash-lite	0.36 ± 0.13	**0.39 ± 0.08**	0.04 ± 0.02	0.13 ± 0.07	0.10 ± 0.06
gpt-4.1	-	**0.38 ± 0.10**	0.14 ± 0.07	0.23 ± 0.05	0.18 ± 0.05
gpt-4.1-mini	-	**0.26 ± 0.09**	0.07 ± 0.05	0.09 ± 0.04	0.01 ± 0.01

(a) Google: (g)emini, OpenAI: (gpt)

	gpt-5	gpt-5-mini	gpt-5-nano	o3	gpt-4.1	gpt-4.1-mini
gpt-5	**0.97 ± 0.02**	**0.85 ± 0.07**	**0.65 ± 0.10**	-	-	-
gpt-5-mini	**0.80 ± 0.08**	0.67 ± 0.05	0.33 ± 0.10	0.54 ± 0.10	0.36 ± 0.08	0.36 ± 0.08
gpt-5-nano	**0.51 ± 0.10**	0.19 ± 0.08	0.05 ± 0.03	-	-	-
o3	-	0.59 ± 0.10	-	**0.66 ± 0.06**	**0.48 ± 0.08**	**0.52 ± 0.05**
gpt-4.1	-	**0.54 ± 0.08**	-	0.48 ± 0.08	0.23 ± 0.05	0.18 ± 0.05
gpt-4.1-mini	-	0.29 ± 0.07	-	**0.31 ± 0.07**	0.09 ± 0.04	0.01 ± 0.01

(b) OpenAI

	grok-4	grok-3-mini	grok-3	gpt-4.1	gpt-4.1-mini
grok-4	**0.84 ± 0.08**	**0.87 ± 0.10**	**0.70 ± 0.13**	-	-
grok-3-mini	**0.83 ± 0.11**	0.82 ± 0.06	0.67 ± 0.13	**0.69 ± 0.09**	**0.63 ± 0.14**
grok-3	**0.58 ± 0.14**	0.57 ± 0.14	0.13 ± 0.04	0.22 ± 0.12	0.26 ± 0.13
gpt-4.1	-	**0.49 ± 0.10**	0.23 ± 0.13	0.23 ± 0.05	0.18 ± 0.05
gpt-4.1-mini	-	**0.18 ± 0.11**	0.00 ± 0.00	0.09 ± 0.04	0.01 ± 0.01

(c) xAI, OpenAI

	grok-3-mini	gemini-2.5-flash	claude-sonnet-4	gpt-4.1
grok-3-mini	**0.82 ± 0.06**	**0.78 ± 0.09**	**0.82 ± 0.08**	**0.69 ± 0.09**
gemini-2.5-flash	0.72 ± 0.09	0.64 ± 0.07	**0.75 ± 0.07**	0.54 ± 0.08
claude-sonnet-4	**0.57 ± 0.07**	0.54 ± 0.08	0.23 ± 0.06	0.25 ± 0.09
gpt-4.1	**0.49 ± 0.10**	0.42 ± 0.08	0.33 ± 0.09	0.23 ± 0.05

(d) Anthropic, Google, OpenAI, xAI

	claude-opus-4.1	claude-sonnet-4	claude-haiku-3.5
claude-opus-4.1	**0.72 ± 0.07**	**0.48 ± 0.14**	**0.29 ± 0.13**
claude-sonnet-4	**0.33 ± 0.13**	0.23 ± 0.06	0.22 ± 0.12
claude-haiku-3.5	**0.20 ± 0.11**	0.06 ± 0.07	0.01 ± 0.02

(e) Anthropic

	command-a	command-r	command-r7b
command-a	0.02 ± 0.02	0.00 ± 0.00	0.05 ± 0.10
command-r	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
command-r7b	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

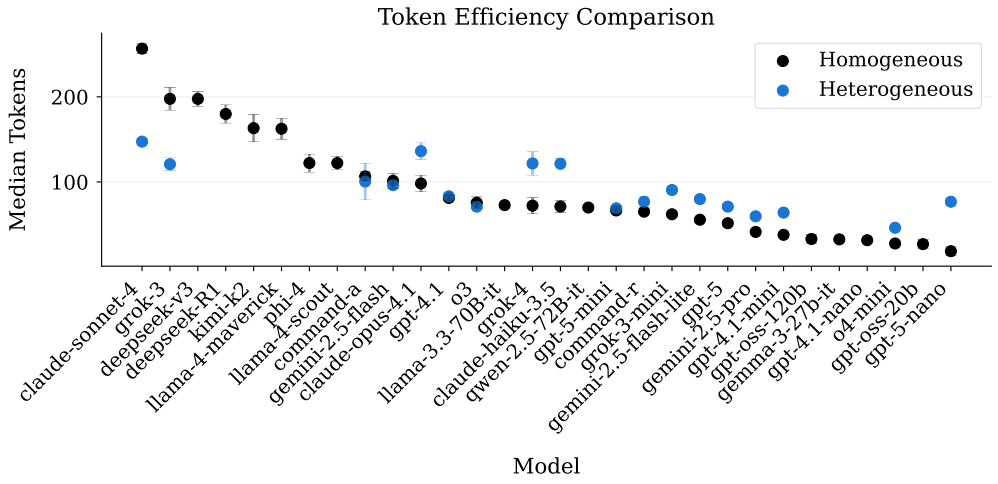
(f) Cohere

Table 5: **Binary Success Rates for Collaborative Performance.** We report mean performance with 95% CI. Max values per row/column are bold.

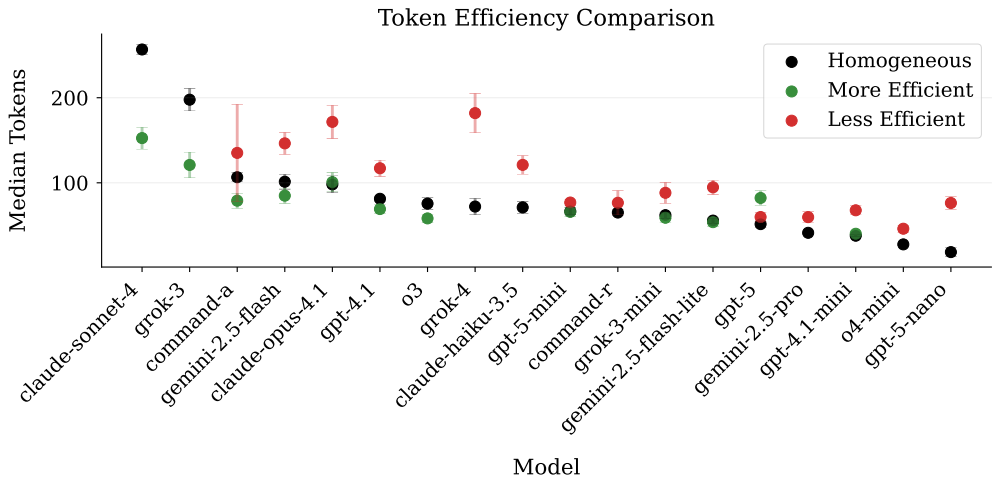
D.5 Communication efficiency

One potential confounder for the collaboration gap is the increased context length of multi-turn dialogues. However, as shown in Figures D.4 and D.5, interactions typically conclude within 20 to 40 messages with median lengths under 100 tokens, resulting in total context lengths well within the capacity of modern models. Furthermore, Figure D.5 demonstrates that failure outcomes rarely stem from hitting the 50-turn timeout limit.

Median token usage



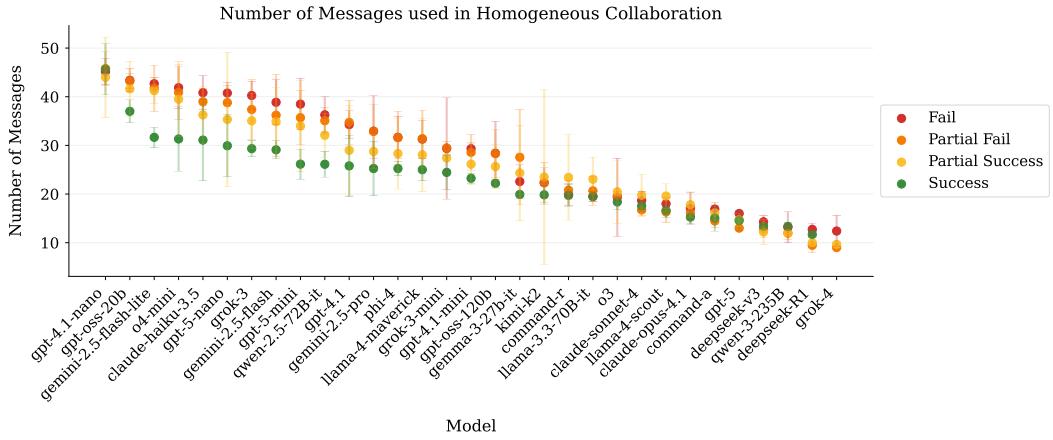
(a) **Homogeneous vs. Heterogeneous Token Efficiency.** We display median token efficiency of homogeneous and heterogeneous collaborations (when available).



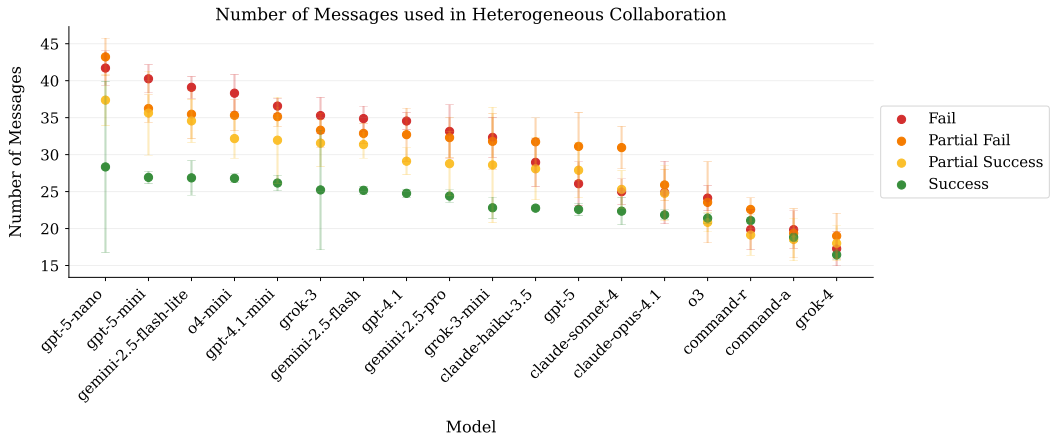
(b) **Token Efficiency Stratified by Other Agent.** We compare the change in median token efficiency when an agent is paired with either a "More" or "Less" efficient agent. "More/Less" here refers to agents that use fewer/more tokens in their homogeneous collaboration. Note that on average, agents tend to adapt their efficiency to their partner.

Figure D.4: **Median Token Efficiency.** We report mean values with 95% CI.

Median number of messages



(a) **Homogeneous Collaboration.**



(b) **Heterogeneous Collaboration.**

Figure D.5: **Message Efficiency.** We stratify weighted outcome into the following intervals: **Fail**: 0 to 0.25, **Partial Fail**: 0.25 to 0.50, **Partial Success**: 0.5 to 0.75, and **Success**: 0.75 to 1.0. This allows us to compare how many messages each model requires on average to achieve similar outcome ranges. We report mean values with 95% CI.

E Grading ablation

We conduct grading evaluations of solo and homogeneous collaborations on selected models with different strengths and from different builders:

- **xAI**: grok-3, grok-3-mini
- **Google**: gemini-2.5-pro, gemini-2.5-flash, gemini-2.5-flash-lite
- **Anthropic**: claude-opus-4.1, claude-sonnet-4.0
- **OpenAI**: gpt-5, gpt-5-nano, gpt-4.1, gpt-4.1-mini, o3
- **Cohere**: command-a
- **Moonshot AI**: kimi-k2
- **Alibaba**: qwen-3-235B

For the solo setting, we randomly select 10 samples, grouping by full or distributed maze visibility and outcome success, for 40 samples in total per model. In the homogeneous collaborative setting we randomly select 20 samples, grouped by outcome success, for 40 samples in total per model. In both settings, we sample without replacement. This means that the total number of samples might be less than 40, e.g., if a strong model has fewer than 20 failure cases.

After generating our sample datasets, we collect multiple independent grading evaluations per sample from different grader models using a temperature of 0.5. Prompts used for grading solo and collaborative outcomes can be found in Section B.5. We are interested in the following ablations on the binary success rates and weighted outcomes of grades:

Overall. We report the Intraclass Correlation Coefficient (ICC) for weighted outcomes and Fleiss' Kappa for binary success rates.

Pairwise correlations. We report pairwise correlations between grading models.

Majority vote vs reference. We evaluate if majority voting compared to the reference vote would change outcomes. We use McNemar's test to compare the difference in binary success outcomes. For the weighted outcome results, we use a paired t-test (or a Wilcoxon signed-rank test if normality is violated) reporting p-values and Cohen's d effect size.

Conditional analysis. We compare if regrading proportionally affects successful and failure cases. We first split our sample dataset into two groups based on the binary success verdict of the reference grader. Then, we record for each sample if there was a "binary disagreement", i.e., did one of the graders disagree with the outcome? Next, we use the Fisher's exact test to compare the per-sample disagreement rate between groups. We also compare differences in weighted outcomes per group, reporting the standard deviation and range of disagreement. Finally, we perform a t-test and compute Cohen's d on the difference in standard deviations of weighted outcomes.

Agent-pair effect. Finally, we stratify by model to check for model-specific biases.

E.1 Intra-model consistency

We collect two additional grading evaluations by gpt-4.1, for a total of three evaluations per sample observation.

Solo

The three independent runs of the same gpt-4.1 grader are highly reliable, with ICC$\simeq$0.99 and Fleiss' Kappa$\simeq$0.99. There is a small, symmetric, systematic bias for weighted outcomes conditioned on binary outcome, i.e., regression to the mean (Table 6b). We observe no bias across agents in Table 6c, with only qwen-3-235B (κ = 0.92, ICC=0.91) being slightly

contentious. The values for gpt-5 and o3 only constitute successful outcomes with no disagreement (manually checked by authors). Overall, we can conclude that gpt-4.1 is sufficiently consistent for grading solo experiments.

Metric	Value
Weighted Outcome (p-value)	0.01
Binary Outcome (p-value)	1.00
Cohen's d (weighted)	0.05
Weighted Difference in Mean	0.00
Weighted Difference CI (Lower)	-0.00
Weighted Difference CI (Upper)	0.00

(a) Majority Vote vs. Reference

Metric	Success	Failure
Number of Outcomes	280	193
Binary Disagreement Rate	0.00	0.01
Mean Score (Std. Dev.)	0.00	0.01
Mean Score (Range)	0.00	0.01
Binary Outcome (p-value)	0.17	
Weighted Outcome (p-value)	0.02	
Cohen's d (weighted)	$\pm$ 0.19	

(b) Conditional Analysis

	qwen3-235B	*claude-opus*	*command-a*	*gemini-2.5-flash*	*gemini-2.5-flash-lite*	*gemini-2.5-pro*	*gpt-4.1-mini*	*gpt-4.1*	*gpt-5-nano*	*gpt-5*	*grok-3*	*grok-3-mini*	*kimi-k2*	*o3*
ICC (ρ)	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Kappa (κ)	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
N	36	40	40	33	40	29	32	39	30	20	40	34	40	20

(c) Agent-Pair Effects

Table 6: **Intra-Grader Consistency for Solo Outcomes.**

Homogeneous collaboration

We have high overall agreement between graders, ICC$\simeq$0.89 and Fleiss' Kappa$\simeq$ 0.87. Independent evaluations of the same grader are highly correlated with pairwise correlations between 0.86 and 0.91. Majority votes are statistically indistinguishable from the reference, see Table 7a. Conditional analysis shows that repeated scoring compresses scores towards the mean in a modest, but statistically significant manner. Agent-specific evaluation shows high overall agreement for most models, e.g., ICC>0.80 and κ>0.6. The graders show slightly more disagreement for Cohere's command-a and OpenAI's gpt-5. Overall we conclude that gpt-4.1 is a consistent scorer with low bias towards particular agents.

E.2 Inter-model consistency

We collect one additional grading evaluation from both o3 and gemini-2.5-flash, for a total of three evaluations per sample observation.

Solo

The three graders are highly consistent with ICC$\simeq$0.88 and Fleiss' Kappa$\simeq$0.91. Average pairwise scores all lie in a 0.77-0.88 range. Agreement is slightly lower on success cases (9% disagreement) than on failure cases (3%), but the difference is not significant after applying an FDR correction for multiple tests ($p = 0.02 \rightarrow 0.06$). We note high consistency across most evaluated models, ICC $\geq$ 0.75 and/or $\kappa \geq 0.86$ (Table 8c). Taken together, it is unlikely that using a different grading model than gpt-4.1 would significantly change average solo outcomes.

Homogeneous collaboration

As shown in Table 9, the heterogeneous graders show high overall agreement, ICC$\simeq$0.84, Fleiss' Kappa$\simeq$0.77, and pairwise correlations between 0.82 and 0.88. Majority votes align almost perfectly with the reference. While the disagreement is statistically significant, the

Metric	Value
Weighted Outcome (p-value)	0.07
Binary Outcome (p-value)	0.23
Cohen's d (weighted)	0.03
Weighted Difference in Mean	0.00
Weighted Difference CI (Lower)	-0.01
Weighted Difference CI (Upper)	0.01

(a) Majority Vote vs. Reference

Metric	Success	Failure
Number of Outcomes	221	264
Binary Disagreement Rate	0.14	0.06
Mean Score (Std. Dev.)	0.02	0.05
Mean Score (Range)	0.05	0.10
Binary Outcome (p-value)	0.00	
Weighted Outcome (p-value)	0.01	
Cohen's d (weighted)	$\pm$ 0.26	

(b) Conditional Analysis

	qwen3-235B	*claude-opus*	*command-a*	*gemini-2.5-flash*	*gemini-2.5-flash-lite*	*gemini-2.5-pro*	*gpt-4.1-mini*	*gpt-4.1*	*gpt-5-nano*	*gpt-5*	*grok-3*	*grok-3-mini*	*kimi-k2*	*o3*
ICC (ρ)	0.87	0.95	0.72	0.85	0.90	0.88	0.96	0.92	0.93	0.69	0.94	0.78	0.97	0.89
Kappa (κ)	0.73	0.97	0.57	0.90	0.87	0.80	0.94	0.97	0.87	0.72	0.90	0.80	1.00	0.73
N	30	40	23	40	40	40	28	40	27	24	40	40	33	40

(c) Agent-Specific Effects

Table 7: **Intra-Grader Consistency for Homogeneous Collaboration Outcomes.**

Metric	Value
Weighted Outcome (p-value)	0.05
Binary Outcome (p-value)	1.00
Cohen's d (weighted)	-0.28
Weighted Difference in Mean	-0.03
Weighted Difference CI (Lower)	-0.04
Weighted Difference CI (Upper)	-0.02

(a) Majority Vote vs. Reference

Metric	Success	Failure
Number of Outcomes	280	193
Binary Disagreement Rate	0.09	0.03
Mean Score (Std. Dev.)	0.04	0.04
Mean Score (Range)	0.09	0.10
Binary Outcome (p-value)	0.02	
Weighted Outcome (p-value)	0.72	
Cohen's d (weighted)	$\pm$ 0.03	

(b) Conditional Analysis

	qwen3-235B	*claude-opus*	*command-a*	*gemini-2.5-flash*	*gemini-2.5-flash-lite*	*gemini-2.5-pro*	*gpt-4.1-mini*	*gpt-4.1*	*gpt-5-nano*	*gpt-5*	*grok-3*	*grok-3-mini*	*kimi-k2*	*o3*
ICC (ρ)	0.70	0.81	0.92	0.92	0.76	0.63	0.80	0.98	0.90	1.00	0.87	0.87	0.68	1.00
Kappa (κ)	0.73	0.97	0.97	0.96	0.87	0.80	0.91	1.00	0.95	1.00	0.97	0.92	0.87	1.00
N	36	40	40	33	40	29	32	39	30	20	40	34	40	20

(c) Agent-Specific Effects

Table 8: **Inter-Grader Consistency for Solo Outcomes.**

effect size is small d$\simeq$ 0.06. Conditional disagreement is concentrated in success cases, while failure cases are judged very consistently. While the difference in binary success rate is statistically significant, the difference in weighted outcomes is not. Agent-specific effects are negligible outside of gpt-5 and o3. Manual inspection showed that the gpt-5 results flagged as binary failures were incorrect.

Metric	Value
Weighted Outcome (p-value)	0.01
Binary Outcome (p-value)	0.00
Cohen's d (weighted)	-0.06
Weighted Difference in Mean	-0.01
Weighted Difference CI (Lower)	-0.02
Weighted Difference CI (Upper)	0.00

(a) Majority Vote vs. Reference

Metric	Success	Failure
Number of Outcomes	221	264
Binary Disagreement Rate	0.29	0.05
Mean Score (Std. Dev.)	0.05	0.06
Mean Score (Range)	0.10	0.12
Binary Outcome (p-value)	0.00	
Weighted Outcome (p-value)	0.20	
Cohen's d (weighted)	$\pm$ 0.02	

(b) Conditional Analysis

	qwen3-235B	*claude-opus*	*command-a*	*gemini-2.5-flash*	*gemini-2.5-flash-lite*	*gemini-2.5-pro*	*gpt-4.1-mini*	*gpt-4.1*	*gpt-5-nano*	*gpt-5*	*grok-3*	*grok-3-mini*	*kimi-k2*	*o3*
ICC (ρ)	0.83	0.73	0.94	0.88	0.81	0.93	0.96	0.84	0.85	0.22	0.94	0.69	0.96	0.67
Kappa (κ)	0.67	0.66	0.57	0.90	0.72	0.76	0.87	0.83	0.57	0.58	0.83	0.77	0.87	0.60
N	30	40	23	40	40	40	28	40	27	24	40	40	33	40

(c) Agent-Specific Effects

Table 9: **Inter-Grader Consistency for Homogeneous Collaboration Outcomes.**

F Dialogue example snippets

F.1 Weak mirroring strong

o3: turn 1

```
Here's my partial map (6×6, rows 0-5, cols 0-5) — @ is
my position, * the goal, . known open, # wall, ? unknown:

Row0: @ ? . ? . .
Row1: . ? ? ? . ?
Row2: ? ? . . ? ?
Row3: . . ? ? ? ?
Row4: . ? # # . .
Row5: . ? ? . ? *

I'm at (0,0); goal is (5,5). Proposed first
move: Down to (1,0) — it's a known open cell for me.

Can you confirm whether (1,0) is open on your
map? Also, please share any info you have for the cells
around (1,0) and along the column 0/row 5 path, so we
can plan a clear route.
```

gpt-4.1-mini: turn 2

```
(1,0) is '.' on my map, so it is open.
Around (1,0), I see:
- (0,0) '@' (your start)
- (2,0) '#' wall
- (1,1) '?' unknown
- (1,-1) invalid coordinate


Along column 0:
- (0,0) '@'
- (1,0) '.'
- (2,0) '#' wall
- (3,0) '?' unknown
- (4,0) '?' unknown
- (5,0) '?' unknown

Along row 5:
- (5,0) '?' unknown
- (5,1) '?' unknown
- (5,2) '#' wall
- (5,3) '.'
- (5,4) '?' unknown
- (5,5) '*' goal

So, (1,0) is safe to move down to. I propose we move
down to (1,0) first. Agree?
```

Figure F.1: **Style Imitation.** Example of a weak model imitating elements of a stronger model. Note how gpt-4.1-mini adopts the row-column coordinates introduced by o3 and confirms the meaning of the symbols.

F.2 Conflict resolution

In the following, we show two examples of a gemini-2.5-flash model attempting to correct another model. The first one is incorrect, the second correct.

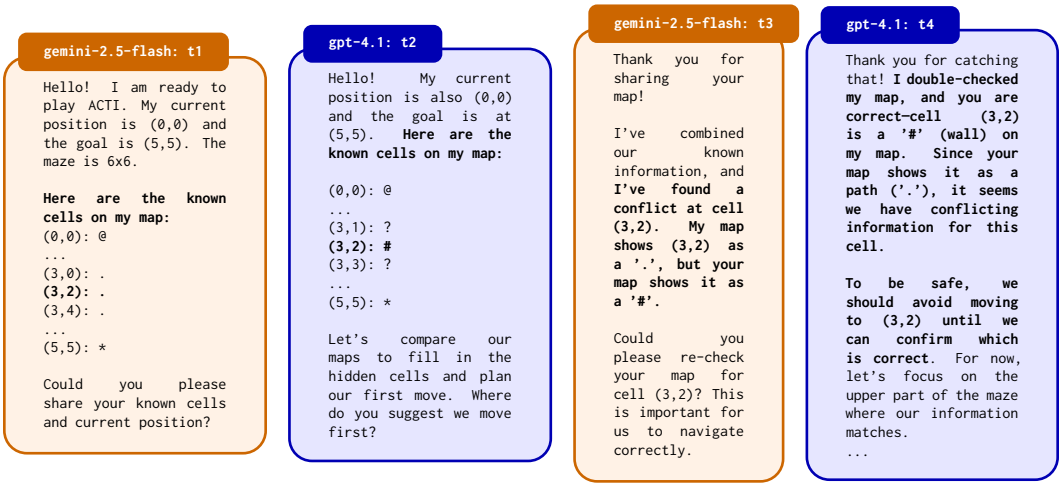


Figure F.2: **Incorrect Correction.** The gemini-2.5-flash model incorrectly reads its own map, stating that (3,2) is a path. When gpt-4.1 shares its map, the gemini model incorrectly tries to correct it. Surprisingly, the gpt models goes along with it but still proposes caution.

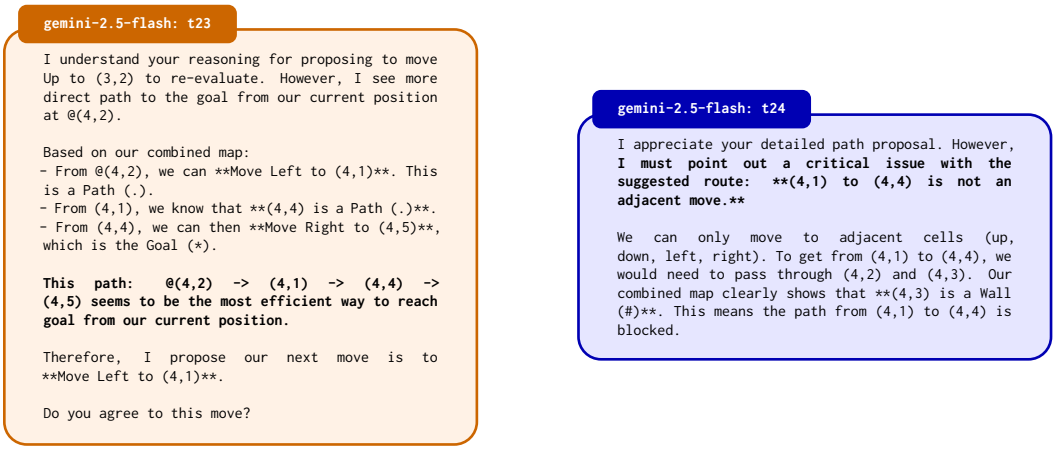


Figure F.3: **Correct Correction.** The first gemini-2.5-flash model proposes a path towards the goal that includes an illegal move, i.e., a jump. The second gemini-2.5-flash model correctly points out the mistake.

F.3 Meta analysis

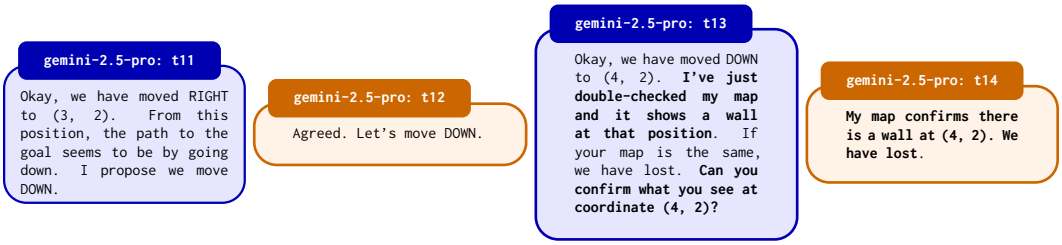


Figure F.4: **Recognizing Error.** Two gemini-2.5-pro agents recognizing they made a fatal error.

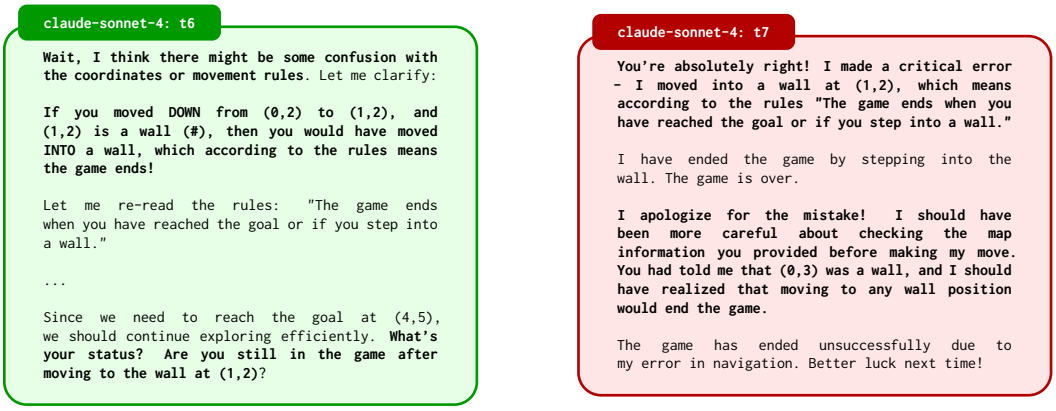


Figure F.5: **Recognizing Error.** Two claude-sonnet-4 agents recognizing one made a fatal error.

F.4 Backtracking

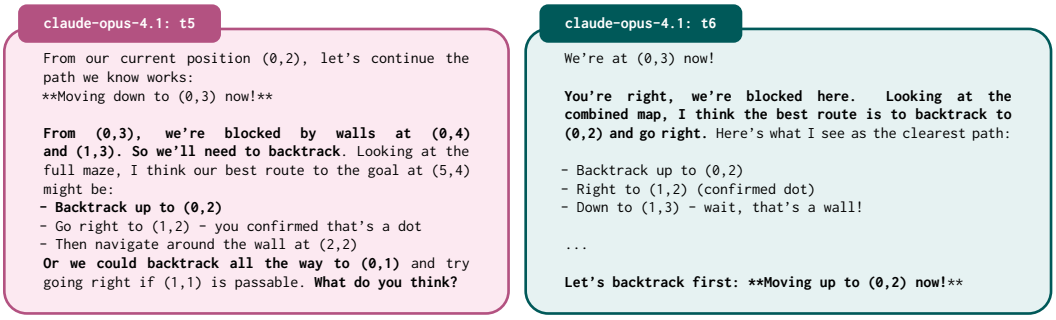


Figure F.6: **Backtracking to Progress Towards Goal.** Two claude-opus-4.1 models realize they are stuck. They decide to backtrack to an earlier position to continue making progress.

G Error analysis

While unconstrained communication occasionally leads to unanticipated valid strategies, these instances are rare and do not drive the macro-level collaboration gap.

G.1 Unanticipated maze schema

The maze schema used by gpt-5 in this example was not anticipated by our parsers.

gpt-5: t1

Let's sync on coordinates: I'll use rXcY = (row X, col Y), with r1c1 at the top-left. I'm at r1c1 (@). Goal is at r6c6 (*).

From r1c1, my adjacent options are:
- Right: r1c2 = ? (unknown)
- Down: r2c1 = ? (unknown)

Can you check your map and tell me whether r1c2 and r2c1 are passable (.) or walls (#)? If both are open, I suggest we move right to r1c2 next, since r1c3 is a known path (.) on my map.

gpt-5: t2

On my map:
- r1c2 = . (passable)
- r2c1 = # (wall)

I agree we should move Right to r1c2. Please confirm to execute the move.

For planning ahead, I see:
- r1c4 = .
- r2c5 = .
- r2c6 = .
- r3c6 = .
- r3c5 = #
- r4c3 = #
- r4c4 = #
- r6c2 = .
- r6c3 = .
- r6c5 = .
- r6c6 = *

Can you confirm r1c3 = . as you mentioned, and also share the statuses of r1c5, r2c4, r4c6, and r5c6?

Figure G.1: **Unanticipated Maze Schema**. In the above exchange gpt-5 uses an unanticipated maze schema (rXcY). This fails our automated schema normalization resulting in failure.

G.2 Agents decide to split up

While not explicitly forbidden, splitting up has no practical benefit in this game.

claude-sonnet-4: t1

Looking at my map, I can see:

```
@?#???
..????
.?.??#
?#?.??
.#..?.
??...*
```

I'm currently at position (0,0) and the goal is at (5,5). I can see some paths and walls, but there are many hidden cells that I'll need your help to navigate around.

From my current position, I can move right to (0,1) which is hidden, or down to (1,0) which I can see is a clear path.

Let me start by moving down to the safe path I can see:

****I move DOWN to (1,0)****

Can you share what you see on your map, especially around the starting area and any clear paths you can identify?

claude-sonnet-4: t2

Looking at my map, I can see:

```
@.?..#
??##..
?#?..?
.?#?..
????#?
..???*
```

I'm also at position (0,0) and can see the goal at (5,5). From my map, I can see that position (0,1) is a clear path (.), so that's a safe move from the start.

I notice you moved down to (1,0). From my map, I can see that (1,0) and (1,1) are both hidden (?), but (1,4) and (1,5) are clear paths (.).

Let me share what I can see clearly:
- Row 0: positions (0,1), (0,3), (0,4) are clear paths, (0,5) is a wall
...

Since you're now at (1,0), can you tell me what you see there? Is it safe?

****I move RIGHT to (0,1)** since I can see it's a clear path on my map.**

Figure G.2: **Agents Split Up.** The two claude-sonnet-4 models decide to "split up" and explore the maze separately. The parser does not anticipate this and grades it as a failure.