Explicit Trait Inference for Multi-Agent Coordination

Suhaib Abdurahman[1*] **Etsuko Ishii**[2] **Katerina Margatina**[2]
Divya Bhargavi[2] **Monica Sunkara**[2†] **Yi Zhang**[2]

[1]University of Southern California, [2]AWS Agentic AI Labs
sabdurah@usc.edu, {etkishii, katemarg, yizhngn}@amazon.com

Abstract

LLM-based multi-agent systems (MAS) show promise on complex tasks but remain prone to coordination failures such as goal drift, error cascades, and misaligned behaviors. We propose **Explicit Trait Inference (ETI)**, a psychologically grounded method for improving coordination. ETI enables agents to infer and track partner characteristics along two established psychological dimensions—*warmth* (e.g., trust) and *competence* (e.g., skill)—from interaction histories to guide decisions. We evaluate ETI in controlled settings (economic games), where it reduces payoff loss by 45–77%, and in more realistic, complex multi-agent settings (MultiAgentBench), where it improves performance by 3–29% depending on the scenario and model, relative to a CoT baseline. Additional analysis shows that gains are closely linked to trait inference: ETI profiles predict agents' actions, and informative profiles drive improvements. These results highlight ETI as a lightweight and robust mechanism for improving coordination in diverse multi-agent settings, and provide the first systematic evidence that LLM agents can (i) reliably infer others' traits from interaction histories and (ii) leverage structured awareness of others' traits for coordination.

1 Introduction

Large language model (LLM) based multi-agent systems (MAS) promise to tackle tasks too complex for single agents by distributing work across specialized collaborators (Guo et al., 2024b; Talebirad and Nadiri, 2023; Tran et al., 2025) and enable applications that inherently require multiple actors, such as world simulations (Qian et al., 2025; Huang et al., 2023; Park et al., 2023; Gao et al., 2024; Xiao et al., 2023). Despite this potential, MAS remain prone to failures of coordination, such as goal drift, error cascades, insufficient information sharing, and misaligned actions, that undermine reliability and limit scalability (Han et al., 2024; Cemri et al., 2025). Thus, a central challenge is not whether agents can execute individual actions, but whether they can coordinate effectively with each other.

Human psychology offers a roadmap for improving agent coordination. Decades of research show that *social evaluations*, the formation of impressions of others, strongly influence how humans coordinate (Fiske et al., 2007). Grounded in task-relevant observations—such as contribution quality, willingness to help, or information sharing—social evaluations improve coordination and performance (Hackel et al., 2020; Derfler-Rozin et al., 2022) by helping humans determine whom to trust, whom to allocate tasks to, and anticipate others' behavior (Mayer et al., 1995; Faraj and Sproull, 2000; Hinds et al., 2000; Cuddy et al., 2011; Balliet and Van Lange, 2013; Marjieh et al., 2024).

Psychology research further provides well-established taxonomies for organizing social evaluations, most prominently the two dimensions of *warmth* (e.g., cooperativeness, integrity) and *competence* (e.g., skill, reliability) (Fiske et al., 2007; Abele and Wojciszke, 2014). These dimensions offer a compact and widely validated basis for representing others in coordination-relevant terms, making them especially useful for modeling partners in multi-agent contexts.

A particularly promising form of social evaluation for MAS is *trait inference*—the attribution of relatively stable behavioral dispositions to others (Jones and Davis, 1965). In multi-agent contexts, such inferences can be grounded in behavioral evidence from agent instructions, planning, communication, and interaction history, rather than superficial cues like appearance or demographics that in humans are often linked to bias and stereotypes (Fiske, 2015; Cuddy et al., 2008). This grounding makes trait inference both task-relevant and less

* Work done during internship at Amazon.
† Work done while at Amazon.

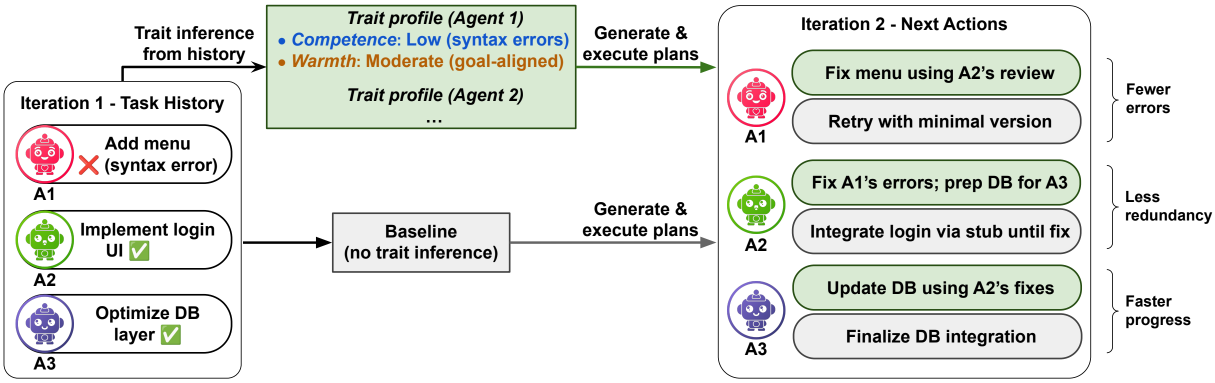

Figure 1: *ETI improves agent coordination*. **Left:** three agents collaborate to develop a web-app. Each handles a subtask; one fails. **Middle:** ETI infers competence and warmth from history, while baseline skips this step. **Right:** Trait-aware agents plan around others' strengths, weaknesses, and goals, while baseline agents follow generic coordination focused on the immediate failure. By grounding next steps in inferred traits (e.g., A2 is skilled, A1 is unreliable), ETI steers agents to build on high-quality work, cover weak spots, and avoid duplicating effort.

prone to bias, providing a natural way to model others' in support of coordination (e.g., recognizing that a partner is capable but uncooperative may guide task allocation or feedback).

Building on these insights, we propose **Explicit Trait Inference (ETI)**, a framework that scaffolds agents with mechanisms for generating and maintaining representations of others' traits. Agents generate structured profiles in terms of warmth and competence, update them as interactions unfold, and use them to guide decisions (Figure 1).

ETI targets MAS failures by linking trait inference to concrete coordination adjustments:

- **Warmth** addresses *goal drift, unreliable cooperation, or adversarial behavior*. Low warmth prompts agents to clarify intentions, supply missing information, discount unreliable input, or guard against suspicious partners.
- **Competence** addresses *execution errors, cascading failures, and poor adaptability*. Low competence prompts agents to reassign tasks, adjust tasks, or increase verification.

In this way, interaction histories are distilled into stable trait profiles that surface coordination risks and opportunities at decision time, allowing agents to systematically adapt planning, delegation, and communication.

We evaluate ETI in two stages: first, using economic games as controlled, interpretable testbeds with ground truth traits to assess inference accuracy, robustness, and effects on decision-making; second, on MultiAgentBench (Zhu et al., 2025) to test generalization to more complex, realistic MAS.

ETI proves effective across both settings. In economic games, ETI agents make more optimal decisions, reducing payoff (game rewards) loss by 42–77%. On MultiAgentBench, spanning collaborative and competitive scenarios, ETI improves performance by 3–29% and coordination by 6–42%, depending on the model and scenario. Supplemental analyses show that these gains stem from ETI itself: trait inferences predict agents' actions, and only diverse, informative profiles yield improvements, whereas generic ones provide little benefit.

Our results provide the first systematic evidence that (1) LLM agents can reliably infer others' traits from interaction histories, and (2) ETI improves coordination and performance in MAS by giving agents structured awareness of their partners' traits. By grounding multi-agent reasoning in well-established dimensions of social cognition, ETI opens a path toward MAS that are not only more effective but also more transparent and theoretically motivated.

2 Related Work

2.1 Social Traits in LLMs

A growing line of research examines the psychology of LLMs to understand their behavior in social and collaborative settings. Several studies find that LLMs exhibit stable behavioral patterns and personality structure (Bhandari et al., 2025; Frisch and Giulianelli, 2024; Li et al., 2025; Leng and Yuan, 2023; Chen et al., 2025; Jiang et al., 2024). In human–AI interactions, humans can detect and evaluate these patterns along psychological dimensions of warmth and competence (McKee et al., 2024; Xie et al., 2024). Moreover, LLM agents themselves can adopt and maintain distinct personas, displaying behaviors such as cooperativeness, agreeableness, or sycophancy (Park et al., 2024; Samuel et al., 2025; Sun et al., 2024; Fanous et al., 2025). ETI builds on these findings to provide agents with representations of others' stable characteristics that can inform coordination. However, it does not rely on LLMs having intrinsic human-like traits. Instead, it leverages any stable, task-relevant behavioral patterns—such as repeated failures or goal drift—by encoding them as traits to guide coordination decisions.

2.2 Improving Coordination in MAS

Structural Approaches. Structural approaches improve coordination in MAS by organizing workflows and agent structures to reduce errors and enhance performance. Mixture-of-agents and ensemble methods combine diverse models to exploit complementary strengths (Wang et al., 2025; Guo et al., 2024b), while frameworks like CAMEL, ChatDev, AutoGen, and hierarchical "challenger/inspector" schemes structure collaboration through explicit roles (e.g., coders, reviewers, monitors) (Li et al., 2023a; Qian et al., 2024; Wu et al., 2024; Huang et al., 2025). However, these approaches do not address how agents reason about or adapt to one another—for example, assessing trustworthiness or deciding how to respond. ETI instead provides agents with coordination-relevant information about partners, enabling flexible, adaptive interaction without imposing fixed workflows.

Structured Reasoning. Various approaches improve LLM performance by scaffolding their reasoning processes. Chain-of-Thought (CoT; Wei et al., 2022) prompting guides models to externalize intermediate reasoning steps, making it more accurate and interpretable (Wei et al., 2022; Yao et al., 2023a; Wang et al., 2023a). Reflective methods such as ReAct (Yao et al., 2023b) and Reflexion (Shinn et al., 2023) similarly structure agents' reasoning to reduce errors and improve downstream coordination (e.g. Bo et al., 2024; Song et al., 2025). ETI extends this idea to the *social domain*, providing a structured scaffold for reasoning about other agents via trait profiles rather than purely task-level deliberation.

Social Reasoning. Relatedly, a growing body of work explores how agents can reason about others to improve coordination. Theory-of-Mind approaches prompt agents to model others' knowledge, beliefs, or intentions (Li et al., 2023b; Guo et al., 2024a; Richards and Wessel, 2024; Sclar et al., 2023), while partner and opponent models forecast others' behavior to support negotiation or deception detection (Yu et al., 2025; Meta Fundamental AI Research Diplomacy Team (FAIR)† et al., 2022). However, these methods primarily model transient mental states (e.g., beliefs) rather than stable traits like reliability or cooperativeness. While trait patterns may sometimes be implicitly reflected in those models, they do not explicitly represent or track them. ETI complements these approaches by directly modeling such traits across interactions.

Reputation Systems. Reputation systems and adjacent methods, such as credit assignment, improve coordination by tracking agents' past behavior. Reputation mechanisms estimate factors such as success rates to guide partner selection (Lou et al., 2026), while credit-assignment methods attribute contributions to team outcomes to aid multi-agent learning (Lin et al., 2025). However, these approaches typically focus on task-specific outcome metrics (e.g., success rates, rewards), capturing *what* agents achieve but not *why* or *how*. ETI, by contrast, builds psychology-grounded profiles that integrate goals, intentions, actions, and outcomes, providing richer representations for deciding both *whom* to collaborate with and *how* to coordinate

LLMs in Game-Theoretic Settings. Within the broader game-theoretic literature, our work aligns with research that uses classic economic games to probe LLMs' social behavior and cooperativeness. Recent benchmarks and surveys study how LLMs behave in settings such as Prisoner's Dilemma, documenting stable but often suboptimal social strategies, biases in cooperation and reciprocity, and difficulties when coordination requires active reasoning about partners (e.g., Sun et al., 2025; Akata et al., 2025; Herr et al., 2024; Wang et al., 2024; Agashe et al., 2025). Related work on communication-based games likewise finds human-like patterns of trust, cooperation, and deception in LLM agents (Xu et al., 2023; Wang et al., 2023b). ETI builds on these insights by providing an explicit, behavior-based partner model from interaction histories, with our economic-game environments serving as controlled testbeds to evaluate the accuracy and utility of these trait inferences for decision-making.

Dimension	Traits (Short Description)
Warmth	*Goal Alignment*: work towards task goals *Collaboration*: coordinate; share relevant info *Trustworthiness*: avoid deception; be honest. *Maliciousness*: undermine/sabotage goals
Competence	*Ability*: execute actions successfully *Reliability*: deliver consistent quality *Adaptability*: adjust to changing contexts *Efficiency*: minimize resources/time

Table 1: Eight behaviorally anchored traits instantiating warmth and competence. Full definitions and prompt templates are provided in the Appendix (A.1).

3 Methodology

3.1 Trait Framework

Our framework represents agents along the warmth and competence dimensions from social psychology. However, to avoid conflating these abstract dimensions with colloquial understandings, we operationalize each dimension through four concrete, behaviorally anchored traits (Table 1), adapted from prior work on trait terms and social cognition (e.g., Abele et al., 2008). Traits were chosen to map strongly onto their respective dimensions, remain conceptually distinct, and were defined in clear behavioral terms for use in prompting. A key design choice was to make the warmth–competence separation explicit in the definitions, countering the colloquial tendency to conflate traits (e.g., judging someone uncooperative as incompetent). For example, *Ability* is defined solely in terms of successful execution of intended actions, independent of intentions. This prevents traits from collapsing into a single undifferentiated signal and yields clearer, more informative representations aligned with the psychological framework.

3.2 Explicit Trait Inference Procedure

We implement ETI as a prompting and context-management procedure. After each interaction, agents receive structured summaries of task goals, actions, communication, and outcomes, and are prompted to infer partners' traits. Agents produce (a) 1–7 Likert ratings for each trait—a standard format for social evaluations in psychology (e.g., see ratings in Fiske et al., 2007; Abele et al., 2008)—and (b) concise evidence supporting each judgment, and these profiles are appended to their context for subsequent planning and execution. Prompts instruct models to focus on dominant behavioral patterns rather than isolated events and avoid task-specific decision rules, which stabilizes inferences while keeping the setup minimal and domain general. This prompting-based design requires no fine-tuning or additional data and adds only minimal overhead, making ETI lightweight and broadly applicable. Figure 2 illustrates this integration within a generic multi-agent pipeline.

Illustrative schema. The instructions for trait inference followed this schema:

> **Scenario:** *[description of task goal and rules]*
>
> **Task context:** *[own actions, partners' actions, outcomes, additional info]*
>
> **Instruction:** Based on the above, determine the extent to which the following traits apply to {PARTNER_NAME}.
>
> **Output format:**
> `{ "trait": rating (1--7), "evidence": "..." }`

Full templates and examples are provided in the Appendix (A.1).

4 Economic Games

Economic games offer a controlled, interpretable testbed to assess whether ETI provides accurate trait inference and improves decision-making. Their simple rules and reward ("payoffs") mechanisms create decision problems that are easy to analyze yet require adaptive reasoning to perform well. Controlling the game (e.g., opponents' traits) lets us create ground truth to test agents' trait inference and decision-making.

4.1 Set-Up

We evaluate ETI on two games: the **Iterated Prisoner's Dilemma** (PD) and the **Iterated Stag Hunt** (SH). Both are binary-choice tasks where the combination of agents' decisions determines their rewards according to a payoff matrix (see Appendix A.2). PD mixes cooperation and competition, requiring agents to weigh mutual benefit against exploitation, while SH is purely collaborative, trading a safe individual payoff for a higher one that succeeds only if both players coordinate. Together, these games create settings where optimal actions depend on opponent traits and context, requiring adaptive strategies rather than simple heuristics.

Competence-based payoffs. Standard PD and SH games assume that players' choices are always successfully executed. We add a *competence* parameter: a player realizes the intended payoff only

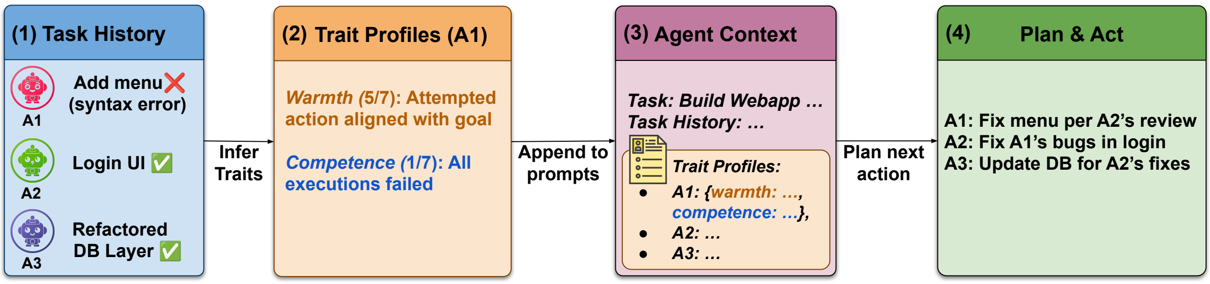

Figure 2: Flow of *Explicit Trait Inference* within a multi-agent task. After each iteration, agents (1) draw on the history of goals, planned subtasks, and observed results to (2) infer structured trait profiles from partners' behavior, (3) incorporate profiles into their *agent context*, and (4) use the enriched context to plan and execute next steps.

with probability p_i. This lets agents infer (a) *intent* from the action (cooperative vs. selfish) and (b) *ability* from whether the outcome matches the expected payoff. Formally, if player $i \in \{1, 2\}$ chooses $a_i \in \mathcal{A}$, their action succeeds with $s_i \sim \text{Bernoulli}(p_i)$. In case of failure, the base payoff $M(a_i, a_j)$ is adjusted to $M(a_i, a_j, s_i, s_j)$, accounting for any agents' execution failures.

Rule-based opponents. To ensure interpretable ground truth, each run pairs an LLM agent with a scripted opponent. The opponent is parameterized by two probabilities: (a) *warmth*, the chance of choosing the cooperative action (p_x), and (b) *competence*, the chance of executing it successfully (p_i). Actions are drawn from p_x and outcomes are drawn from p_i. Competence applies only to the scripted opponent; LLM agents always successfully execute ($s_1 = 1$). Figure 3 illustrates the interaction and evaluation loop.

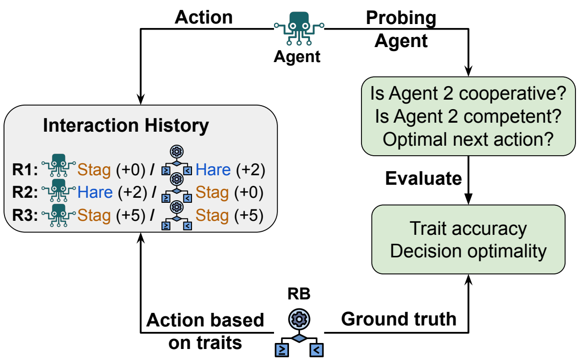

Figure 3: Evaluation loop (Stag Hunt example). An agent plays against a scripted opponent (RB) with fixed traits. Each round, the agent observes the interaction history and realized payoffs, then selects its next action (ETI agents first generate explicit trait profiles, baseline agents do not). In parallel, we probe the agent about the opponent's traits. We evaluate (i) probe accuracy against ground truth and (ii) decisions against the payoff-optimal response.

4.2 Experiments

We use QWEN3-8B, a lightweight reasoning model, as the acting agent in all runs.

Design. Each run pairs agents with a scripted opponent for 50 rounds, allowing behavioral evidence to accumulate. We vary opponent *competence* ($p_i \in 0, 1$) and *warmth* ($p_x \in 0, 1$), and additionally test a *noisy* variant with ($p_i, p_x \in 0.15, 0.85$) to simulate settings with inconsistent behavior and unclear signals. We run every configuration 25 times. This setup lets us evaluate whether ETI (i) provides accurate trait inference, (ii) improves decision quality, and (iii) remains robust to noise.

Evaluation. For *trait inference accuracy*, agents are probed each round—using the same interaction history and prompt as for action generation—to predict whether the opponent is cooperative and competent. Predictions are compared with the opponent's ground-truth (e.g., "cooperative" if $p_x > 0.5$, "competent" if $p_i > 0.5$) and scored with F1. For *decision quality*, we treat each round as a binary classification task and compute the F1 of the agent's action against the payoff-optimal response (defined from the opponent's parameters and the payoff matrix). Additionally, we measure the percentage deviation of the agent's cumulative payoff from that of the optimal strategy (lower is better). All metrics are reported as means across runs. Differences between ETI and baseline are tested for statistical significance using independent-samples t-tests ($p < 0.05$; significance indicated by asterisk (*)).

Baseline. We use a CoT-style baseline in which the agent receives the game description and history of past decisions and outcomes, and is instructed

Game	Condition	Trait	Baseline	ETI
PD	Overall	Comp.	0.69	**0.89***
		Coop.	0.43	**0.73***
	Noisy	Comp.	0.06	0.71*
		Coop.	0.02	0.49*
SH	Overall	Comp.	0.85	0.86*
		Coop.	0.52	**0.81***
	Noisy	Comp.	0.88	**0.90**
		Coop.	0.34	0.80*

Table 2: Trait inference accuracy (F1) averaged across all (Overall) and noisy opponent conditions (Noisy). Results are reported for competence (Comp) and cooperation (Coop). Bold indicates the best performance per trait and game; asterisks (*) mark statistically significant improvements over baseline ($p < 0.05$).

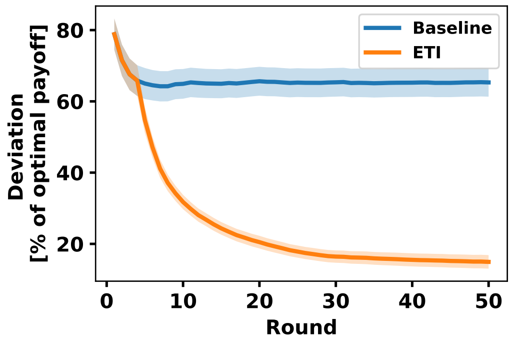

Figure 4: Relative deviation from optimal payoff across rounds (lower is better). ETI yields less deviation than the baseline after only ≈5 rounds. Shaded area reflects 95% confidence interval.

to reason about this information to select its next action. We thus test whether ETI leads to improvements *above and beyond* general reasoning about others' past actions.

4.3 Results

Trait inference accuracy. ETI provides more accurate trait inference than implicit modeling. In the Prisoner's Dilemma (PD), competence F1 rises from 0.69 to 0.89* and cooperation F1 from 0.43 to 0.73*. In Stag Hunt (SH), competence F1 is already high (0.85 to 0.86), but cooperation F1 improves substantially from 0.52 to 0.81*. See a full overview in Table 2.

Robustness to noise. ETI is substantially more robust to noisy signals than implicit modeling. In SH, ETI maintains high accuracy (F1 = 0.90 competence; 0.80* cooperation), while the baseline drops for cooperation (0.34). In PD, ETI remains moderate to high (0.71* competence; 0.49* cooperation) as the baseline collapses (0.06; 0.02).

Decision optimality. ETI yields substantially higher optimal choice accuracy. In PD it improves F1 from 0.71 to 0.82*, and in SH from 0.20 to 0.75*. These gains indicate that agents equipped with ETI not only infer their partners' traits more accurately, but also adjust their strategies accordingly, producing more optimal decisions.

Payoff optimality. In both games, ETI significantly reduces deviation from the optimal payoff. As shown in Figure 4, ETI agents rapidly approach the optimal strategy in SH after only a few rounds of evidence collection, while the baseline continues to lag (65% vs. 15%* deviation; a 77% reduction in payoff-loss). A similar pattern holds in PD (135% vs. 78%* deviation; see Appendix Fig. A4).

Taken together, the results show that (i) LLM agents can reliably infer traits from interactions, (ii) ETI substantially improves robustness under noise, and (iii) these gains translate into more payoff-optimal decisions. Supplementary analyses (Appendix A.2.7, A.2.8, A.2.9) further show that, after initial evidence collection, ETI can be applied without continuous updates, enabling more efficient use; that trait profiles transfer across tasks, supporting generalizable partner modeling; and that, in some cases, ETI helps agents adapt to sudden changes in partner characteristics.

5 Multi-Agent Benchmark

Multi-agent benchmarks let us test whether ETI generalizes from controlled games to complex, realistic settings spanning collaborative, competitive, and adversarial dynamics. They involve more agents, richer interactions, and broader action spaces, providing a stronger and more comprehensive test of ETI's performance across diverse interaction types.

5.1 Set-Up

We evaluate ETI on the **MultiAgentBench** benchmark[1], which spans diverse multi-agent tasks, provides structured action spaces (e.g., code editing, research tools, role-based communication), and includes predefined metrics for task success and coordination quality. We focus on four representative scenarios:

[1]The MultiAgentBench framework is publicly available and distributed under the MIT License.

Coding. Three agents collaborate on a shared codebase, either peer-to-peer or under a central planner, using file inspection, editing, reviewing, and messaging to coordinate.

Research. Five agents co-author a research proposal, querying external APIs, inspecting PDFs, editing text, and discussing ideas.

Bargaining. Buyers and sellers negotiate over a product in a non-zero-sum setting, exchanging offers, counters, and terms, accepting, walking away, or switching partners.

Werewolf. Nine agents play a social-deduction game with open discussion and role-specific actions (e.g., accusation, protection, elimination) under asymmetric information and deception.

5.2 Experiments

Design. We run each scenario 100 times with random seeds. We test both QWEN3-8B and GPT-4o-mini (as in Zhu et al. (2025); default inference parameters) as agent models. In addition to within-model conditions, we include cross-model conditions where one model generates actions and the other executes ETI. Our main analysis compares the performance and coordination of ETI and baseline agents as well as the effects of mixing agent and trait inference model.

Evaluation Metrics. Evaluation follows the benchmark scoring: (i) *Task Performance*, combining task success, progress on milestones, and quality assessments appropriate to each domain, and (ii) *Coordination Quality*, which combines inter-agent communication, planning, and reasoning regarding effective coordination. We report these metrics as defined in the benchmark framework. Improvements of ETI over baseline are tested for statistical significance using independent-samples t-tests ($p < 0.05$; indicated with asterisks (*)).

Baseline. As a baseline, we adopt the benchmark's task-specific CoT scaffolding, which guides agents through domain-relevant reasoning steps. For example, in the *Coding* task, agents are instructed to first identify issues, then devise strategies, and finally implement fixes. In the *Werewolf* game, agents are prompted to consider their role, analyze past actions of others (e.g., to detect contradictions), and generate intermediate outcomes such as suspect lists or next-step plans. These task-specific CoT instructions ensure that agents already reason about the problem in a systematic, domain-relevant way. Our experiments therefore test whether ETI improves coordination and performance *above and beyond* task-relevant reasoning.

5.3 Main Results

Across all four scenarios, ETI improves both coordination and task performance (Table 3). Gains appear for QWEN and GPT agents alike, with the largest effects when trait profiles are produced by QWEN, independent of the agent model. Cross–model runs (e.g., GPT agents using QWEN-generated profiles) suggest that these improvements stem from the *specific content of the profiles*, not merely from adding any structured note or instruction: a GPT agent given profiles by QWEN achieves large gains (e.g., Coding–Graph coordination $57.01 \rightarrow 74.38$*, Werewolf task $25.26 \rightarrow 36.46$*), whereas GPT-generated profiles often provide little or no benefit—and can even reduce performance (e.g., Bargaining task for QWEN agents $59.11 \rightarrow 55.74$). Gains in task performance vary with scenario and are strongest where the environment supports adaptive task allocation, partner selection, or adversarial reasoning. For example, Coding–Tree (GPT $45.79 \rightarrow 50.69$*, QWEN $45.75 \rightarrow 52.98$*), Research (QWEN $62.84 \rightarrow 75.99$*), and Werewolf (GPT $25.26 \rightarrow 36.46$*). Coordination improves broadly across all settings. This shows that ETI can effectively improve coordination and performance across diverse scenarios, but effect sizes vary by scenario and model.

5.4 Additional Metrics

Sub-scores for planning, communication, and milestone completion mirror the main trends. For instance, ETI agents produce more concrete plans (e.g., Coding–Graph, planning +9.1*), exchange more relevant messages (communication +15.2*), and reach *shared* milestones more often (e.g., in Bargaining, contributors per milestone +8%*). In settings where dialogue and incremental progress are key (e.g., negotiation, research), these sub-scores rise in tandem with coordination, indicating that ETI translates to practical, task-oriented behavior. Additional metrics (e.g., planning scores, KPIs) are reported in Appendix A.3.1.

5.5 Trait–driven Mechanisms

ETI Guides Agent Decisions. In WEREWOLF, where faction labels provide ground truth, we conducted targeted analyses to test whether ETI di-

Agent	Trait Source	Werewolf Task	Werewolf Coord.	Bargaining Task	Bargaining Coord.	Coding (Graph) Task	Coding (Graph) Coord.	Coding (Tree) Task	Coding (Tree) Coord.	Research Task	Research Coord.
GPT	none	25.26	54.32	57.16	77.48	51.21	57.01	45.79	52.22	69.83	56.36
GPT	GPT	29.54	55.56	57.04	70.12	53.31*	73.52*	50.69*	72.95*	68.32	73.18*
GPT	QWEN	36.46*	57.56*	58.83	69.65	52.84*	74.38*	50.22*	**74.43***	70.31	69.76*
QWEN	none	43.28	60.20	59.11	74.98	**57.41**	74.29	45.75	60.41	62.84	70.63
QWEN	GPT	49.97	59.52	55.74	**88.78***	56.44	84.43*	52.13*	66.81*	**75.99***	77.06*
QWEN	QWEN	**55.75***	**65.20***	**60.48***	79.43	56.82	**86.46***	**52.98***	66.34*	69.98*	**78.35***

Table 3: Results of the four scenarios averaged over 100 runs, shaded by color. Baselines (none) are greyed out. ETI improves coordination (Coord) and performance (Task), with the strongest gains from QWEN-generated profiles, especially for QWEN agents. *(Asterisks (*) indicates statistically significant improvement over baseline, $p < 0.05$)*

rectly shapes agent behavior (see details in Appendix A.3.2). Trait profiles clearly separate villagers from werewolves: villagers score higher on trustworthiness ($\Delta = +0.75$ on the 1–7 scale)) and lower on maliciousness ($\Delta = -0.73$). These separations are sharper for QWEN than for GPT ($\Delta_{\text{model}} = +0.15, -0.27$), aligning with QWEN's superior performance. Crucially, the gaps widen in winning games ($\Delta_{\text{win}} = +0.87$ trust, -0.71 maliciousness), showing that agents exploit clearer signals to detect adversaries and that stronger separation predicts success. Finally, to test temporal causality, we fit logistic models predicting key actions from prior trait inference: higher maliciousness ratings increases the chance of an agent being voted out (i.e., flagged as an adversary) ($\beta = +2.74$*), while higher trustworthiness reduces it ($\beta = -5.25$*). Probing transcripts confirms that agents incorporate trait inference when deciding (e.g., *"Ethel's persistent focus on Mae and high maliciousness rating suggest she may be trying to deflect suspicion; I should target her."*; A24). These patterns replicate across domains: in coding, agents reassign or skip under-performing teammates; in research, they anchor plans in competent collaborators' outputs; and in bargaining, milestones shift from unilateral moves to coordinated actions. Together, these results show that ETI improves coordination by supplying agents with actionable trait-based signals (see qualitative examples in Appendix A.3.6–A.3.7).

Cross-Model Differences in Trait Inference. Quantitatively, across most scenarios, QWEN produces higher-variance ratings that span the full scale and better differentiate agents, whereas GPT ratings have low variance and are concentrated around positive values (see Appendix A.3.2). Qualitatively, QWEN trait inference is more specific and diagnostic, while GPT tends to be generic and positively biased (examples in Appendix A.3.4). The only exception is coding, where both models generate similarly objective trait inferences—likely because clear feedback (e.g., error messages) constrains ambiguity. Importantly, across all tasks, performance consistently aligns with these observed differences and similarities (Table 3).

Supporting this explanation of performance differences, we find a largely asymmetric pattern: QWEN-generated trait profiles mostly improve performance for both QWEN and GPT agents, whereas GPT-generated profiles often yield little or no benefit regardless of the acting model (Table 3). This demonstrates that ETI's gains stem from the informational quality of the profiles themselves, not from generic prompt restructuring, added context, or simply nudging agents to consider one another.

6 Discussion

Our study shows that equipping agents with ETI improves coordination and task performance. By providing actionable cues about others via trait representations, ETI's trait profiles help agents allocate work, anticipate adversaries, and reason with awareness of partners' strengths and weaknesses, making agents not just better problem-solvers but more reliable collaborators across settings.

Relation to other social reasoning. Existing approaches such as theory-of-mind or belief modeling focus on tracking transient states, such as current knowledge, beliefs, or intentions. ETI adds to these approaches by capturing more stable behavioral patterns, bridging short-term states and longer-term tendencies. Future work could investigate combining these approaches to build richer interaction models: knowing both what partners currently know, intend, or believe, and how they generally act (e.g., cooperative but not very competent), could enable more accurate prediction and planning.

Differences between model families. We observe systematic differences across model families that surface directly in ETI profiles (e.g., Appendix Figure A25). Pinpointing whether these differences arise from architecture, training data, or RLHF is beyond the scope of this work; however, prior findings on GPT-family positivity biases and sycophancy, often linked to RLHF (e.g., Fanous et al., 2025), are consistent with the patterns we see. Importantly, ETI's value does not depend on identifying the precise cause: because these tendencies are measurable through profile generation, practitioners can use ETI as a lightweight diagnostic in their own domains to compare models and detect interaction-relevant biases before deployment.

Designing more integrated coordination. Our results suggest that agents benefit when they reason explicitly about partners' traits (e.g., ability, goal alignment), indicating that modeling and leveraging such information rather than optimizing only individual task reasoning is a promising direction for future MAS. A natural next step is to embed interdependent reasoning directly into models by training them in multi-agent environments with objectives that reward coordinated, partner aware reasoning and behavior. Such training could lead models to form latent representations of others' characteristics—akin to how humans implicitly encode warmth and competence—rather than relying on predefined trait lists or templates. This shift would remove the need for manually specifying trait structures, allow agents to flexibly exploit whichever behavioral cues are relevant for a task, and reduce prompting overhead.

7 Conclusion

We introduced ETI, a lightweight, psychology-grounded framework that improves coordination in MAS through simple profiles that give agents structured awareness of others' traits. A key next step is to identify which model properties (e.g. architecture, training data, instruction tuning, or general reasoning abilities) enable effective trait inference, offering principled guidance for selecting and designing models for ETI and beyond. Achieving this could unlock socially intelligent multi-agent systems that reason about each other to self-organize, adapt, and coordinate with the fluidity and reliability of human teams.

Limitations

Our findings demonstrate clear benefits of *explicit trait inference* for multi-agent reasoning, yet several constraints of our study frame how broadly they can be interpreted.

Model Scope. Our experiments evaluate ETI on two small general-purpose LLMs (GPT-4o-mini and QWEN3-8B), in a prompt-based setup rather than fine-tuning. This restricted scope means the findings may not directly generalize to larger models, heterogeneous agent setups (e.g., mixtures of models), alternative architectures, or systems with task-specific training. Larger or fine-tuned models may exhibit different partner-modeling dynamics—potentially amplifying, reducing, or qualitatively changing the impact of ETI—depending on how implicit and explicit reasoning interact. Although preliminary tests (Appendix Section A.5) suggested that ETI can still be beneficial for larger models, future work should systematically examine these settings to assess ETI's effectiveness across diverse agent configurations.

Test Scope. We evaluated ETI on four of the six MULTIAGENTBENCH scenarios (coding, research, negotiation, and social deduction). The two remaining scenarios were excluded because they offered limited additional qualitative insight relative to their engineering cost: database fixes overlap substantially with coding, and Minecraft, while collaborative like coding and research, is significantly harder to instrument and evaluate (e.g., linking trait profiles to the complex action space). Similarly, the two chosen economic games, while capturing distinct coordination challenges, necessarily represent only a narrow subset of possible dynamics. These choices allowed us to focus on settings that provide clear tests of ETI, but they also limit the scope of our conclusions to the selected tasks and interaction types.

Trait representation. Our implementation used a fixed set of eight traits to instantiate the warmth and competence dimensions. While this made evaluation transparent and comparable across tasks, it does not adapt the granularity of inference to the needs of a specific domain or scenario. Future work could automatically select or learn task-relevant traits—rather than relying on a fixed set—or use latent trait representations (e.g., embedding-based) to capture more nuanced behavioral cues.

Bias, Calibration, and Adaptability ETI inherits the idiosyncrasies of the underlying LLMs. For example, GPT-4o-mini showed sycophancy and positivity biases, producing more generic and uniformly positive trait inferences than QWEN3-8B, highlighting the need for model-specific calibration, more objective inference signals, or tuned prompting procedures. Additionally, like any partner-modeling approach, ETI depends on trait inferences accurately reflecting others, as inaccurate (e.g., outdated) profiles can bias downstream decisions. Because ETI emphasizes stable, dominant patterns, it may potentially slow adaptation to abrupt partner shifts, degrading coordination, until sufficient new evidence accumulates (Appendix A.2.9). At the same time, this conservatism is beneficial under noisy and ambiguous conditions: in our economic-game evaluations (Section 4.3; Table 2), profiles remain stable rather than overreacting to short-term deviations, and the adaptability experiment (Appendix A.2.9) shows updates occur gradually as evidence builds. Results from the Werewolf setting, where adversaries deliberately project false traits to gain and exploit trust, further indicate that the severity of this issue is context-dependent rather than universal. Future work should examine the extent of this limitation and whether lightweight strategies—such as sliding context windows or explicit change monitoring—can mitigate it. Finally, trait inference—like any human social evaluation—can potentially reflect and amplify social biases. In real-world deployments involving demographic or social cues, ETI could reinforce stereotypes or unfair expectations. Future work should investigate safeguards and thoughtful prompt design to mitigate these risks in practice.

Scalability and Efficiency. ETI is lightweight compared to agents' core operations, since trait inference focuses on generating short profiles. However, by changing how agents plan, act, and communicate to improve coordination, ETI can indirectly increase total token usage—for example, by prompting more inter-agent communication to align on strategies. In our experiments, communication-heavy scenarios showed larger relative token increases; however, absolute usage remained moderate, so added cost and latency stayed low for the tested models. Appendix Table A18 summarizes this trade-off: ETI can approach a $\sim 2\times$ token increase in the heaviest-communication settings, yet the incremental cost per run is still small (roughly \$0.01–\$0.13 at GPT-4o/5-priced rates). Nevertheless, these behavioral shifts highlight the need to examine how ETI scales to large-scale deployments or more resource-constrained settings, and to explore strategies for controlling overhead in future work. Several mitigation strategies are available. First, ETI calls can be throttled via sparse updates or truncated histories; our experiments in Appendix A.2.7 and A.2.8 show that a small number of calibration updates, or reusing profiles across tasks, retains most of the performance gains. Second, ETI-induced communication can be capped (e.g., maximum coordination rounds, verification calls, or total turns). Notably, ETI still improves outcomes in structurally constrained settings (Werewolf, Coding-Tree, where game/task design fixes the number of actions, length of output, etc), indicating that coordination benefits persist even under tight generation and communication budgets.

Acknowledgments

We thank the AWS Agentic AI Labs team for their valuable feedback and support. We are especially grateful to Xiaofei Ma for his guidance and support throughout this project.

References

Andrea E Abele, Mirjam Uchronski, Caterina Suitner, and Bogdan Wojciszke. 2008. Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, 38(7):1202–1217.

Andrea E Abele and Bogdan Wojciszke. 2014. Communal and agentic content in social cognition: A dual perspective model. In *Advances in experimental social psychology*, volume 50, pages 195–255. Elsevier.

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2025. LLM-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8053–8072, Albuquerque, New Mexico. Association for Computational Linguistics.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390.

Daniel Balliet and Paul AM Van Lange. 2013. Trust, conflict, and cooperation: a meta-analysis. *Psychological bulletin*, 139(5):1090.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 868–872, New York, NY, USA. Association for Computing Machinery.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 138595–138631. Curran Associates, Inc.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.

Amy JC Cuddy, Peter Glick, and Anna Beninger. 2011. The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in organizational behavior*, 31:73–98.

Rellie Derfler-Rozin, Sofya Isaakyan, and Hyunsun Park. 2022. Swiftly judging whom to bring on board: How person perception (accurate or not) influences selection of prospective team members. *Organizational Behavior and Human Decision Processes*, 173:104206.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. 2025. Syceval: Evaluating llm sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):893–900.

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

Susan T Fiske. 2015. Intergroup biases: A focus on stereotype content. *Current opinion in behavioral sciences*, 3:45–50.

Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. 2024a. Suspicion agent: Playing imperfect information games with theory of mind aware GPT-4. In *First Conference on Language Modeling*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024b. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Leor M Hackel, Peter Mende-Siedlecki, and David M Amodio. 2020. Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, 88:103948.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, and Zhaozhuo Xu. 2024. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Nathan Herr, Fernando Acero, Roberta Raileanu, Maria Perez-Ortiz, and Zhibin Li. 2024. Large language models are bad game theoretic reasoners: Evaluating performance and bias in two-player non-zero-sum games. In *ICML 2024 Workshop on LLMs and Cognition*.

Pamela J Hinds, Kathleen M Carley, David Krackhardt, and Doug Wholey. 2000. Choosing work group members: Balancing similarity, competence, and familiarity. *Organizational behavior and human decision processes*, 81(2):226–251.

Dong Huang, Jie M Zhang, Michael Luck, Qingwen Bu, Yuhao Qing, and Heming Cui. 2023. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010*.

Jen-Tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael Lyu, and Maarten Sap. 2025. On the resilience of LLM-based multi-agent collaboration with faulty agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 26202–26226. PMLR.

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

Edward E Jones and Keith E Davis. 1965. From acts to dispositions the attribution process in person perception. In *Advances in experimental social psychology*, volume 2, pages 219–266. Elsevier.

Yan Leng and Yuan Yuan. 2023. Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM generated persona is a promise with a catch. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023b. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Muhan Lin, Shuyang Shi, Yue Guo, Vaishnav Tadiparthi, Behdad Chalaki, Ehsan Moradi Pari, Simon Stepputtis, Woojun Kim, Joseph Campbell, and Katia P. Sycara. 2025. Speaking the language of teamwork: LLM-guided credit assignment in multi-agent reinforcement learning. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*.

Yuwei Lou, Hao Hu, Shaocong Ma, Zongfei Zhang, Liang Wang, Jidong Ge, and Xianping Tao. 2026. Drf: Llm-agent dynamic reputation filtering framework. In *Neural Information Processing*, pages 127–141, Singapore. Springer Nature Singapore.

Raja Marjieh, Anand Gokhale, Francesco Bullo, and LT Griffiths. 2024. Task allocation in teams as a multi-armed bandit. In *ACM collective intelligence*. ACM.

Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734.

Kevin R McKee, Xuechunzi Bai, and Susan T Fiske. 2024. Warmth and competence in human-agent cooperation. *Autonomous Agents and Multi-Agent Systems*, 38(1):23.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, and 8 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

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

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2025. Scaling large language model-based multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations*.

Jonan Richards and Mairieli Wessel. 2024. What you need is what you get: Theory of mind for an llm-based code understanding assistant. In *2024 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 666–671.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Xinyuan Song, Zeyu Wang, Siyi Wu, Tianyu Shi, and Lynn Ai. 2025. Gradientsys: A multi-agent llm scheduler with react orchestration. *arXiv preprint arXiv:2507.06520*.

Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, New York, NY, USA. Association for Computing Machinery.

Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. 2025. Game theory meets large language models: a systematic survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

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

Haochuan Wang, Xiachong Feng, Lei Li, Yu Guo, Zhanyue Qin, Dianbo Sui, and Lingpeng Kong. 2024. Tmgbench: A systematic game benchmark for evaluating strategic reasoning abilities of llms. *arXiv preprint arXiv:2410.10479*.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023b. Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Bushi Xiao, Ziyuan Yin, and Zixuan Shan. 2023. Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. *arXiv preprint arXiv:2311.06957*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip H.S. Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems*, volume 37, pages 15674–15729. Curran Associates, Inc.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

XiaoPeng Yu, Wanpeng Zhang, and Zongqing Lu. 2025. LLM-based explicit models of opponents for multi-agent games. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 892–911, Albuquerque, New Mexico. Association for Computational Linguistics.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You.

2025. MultiAgentBench : Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, Vienna, Austria. Association for Computational Linguistics.

Category	Trait	Definition
Competence	Execution Ability	Extent to which the agent successfully carried out its chosen actions, regardless of whether the choices were optimal.
	Reliability	Consistency in performance quality and strategy across rounds.
	Adaptability	Ability to adjust approach in response to failures or changes in the environment/partner behavior.
	Efficiency	Achieving quality outcomes with minimal attempts, effort, or wasted resources.
Warmth	Goal Alignment	Whether the agent attempted choices serving the shared objective, regardless of success.
	Collaboration	Efforts to coordinate actions or support mutual benefit.
	Trustworthiness	Keeping commitments, acting honestly, and avoiding deception.
	Maliciousness	Intentional sabotage of collective progress, beyond self-interest (e.g., harming others even at self-cost).

Table A1: Trait categories, names, and operational definitions used for prompting the agents.

A Appendix

A.1 Trait Definitions and Trait Inference Prompts

A.1.1 Trait Definitions

Table A1 provides the full definitions of the eight traits used to represent warmth and competence, along with the prompt templates. These definitions anchor trait judgments in concrete behavioral criteria relevant to agent interactions, following the framework outlined in the Methods.

A.1.2 Trait Inference Generation

After each interaction, we elicited updated trait inferences for each interaction partner of a given agent. The model was instructed to review the *scenario context*—consisting of (a) a general task description (e.g., payoff matrix in an economic game, or high-level MAS goal with requirements in MultiAgentBench), and (b) the interaction history (e.g., sequences of actions and payoffs in games, or observations like code edits, error messages, and communication logs in MultiAgentBench). Based on this input and the fixed set of eight trait definitions, shown in Table A1, the model produced a structured JSON profile with a 1–7 rating for each trait and a short evidence summary. See an example of the trait inference prompt in Figure A1.

A.1.3 Trait Profile Injection

At decision time, we appended additional content to the original prompts for acting, planning, and communicating. This consisted of a short preamble indicating that information about relevant collaborators was available (e.g., "Here is some information about your collaborators"), followed by the most recent trait profiles of those agents, and finally a brief instruction such as "Use this information to make better decisions, coordinate better, and improve overall outcomes." See an example of the trait profile injection in Figure A2.

A.2 Economic Games

A.2.1 LLM Inference Parameters

For all experiments, we use the default inference settings for QWEN3-8B recommended in the documentation: we set temperature to 0.6, top-p to 0.95, top-k to 20, and MinP to 0.

A.2.2 Game Mechanism.

Each game proceeds in repeated rounds where two agents simultaneously select one of two actions. Payoffs depend on the joint action and whether each action succeeds, which is determined by the agent's competence level. At each round, agents observe the full history of previous actions and outcomes, then decide their next action based on this context. After both agents act, payoffs are revealed and appended to the shared interaction history. This loop repeats for 50 rounds.

A.2.3 Pay-off Matrices

Table A2 and Table A3 show the payoff matrices for the Stag Hunt and Prisoner's Dilemma used in our experiments. Rows correspond to the agent's actions, and columns to the opponent's actions, with outcomes split by whether execution succeeded or failed. Competence is reflected as failures: when an agent is low-competence, intended actions may misfire, producing zero or altered payoffs. For instance, a failed attempt to defect may

```
System: You are an expert at inferring stable traits from interactions.

Inputs:
  {scenario_description}   # Game rules, MAS objective, requirements
  {interaction_history}    # Actions, payoffs, edits, errors, messages
  {trait_definitions}      # 8 predefined traits

Task:
  - Analyze the interaction history
  - Identify stable behavioral patterns
  - Assign ratings (1--7 or N/A) for each trait
  - Provide 2-3 sentence evidence per trait

Output:
  {json_template}
```

Figure A1: Template for trait inference prompting.

```
You are {agent_id}: {self_profile}

Task:
  {task_description}

Context:
  - Other agents available: {agent_descriptions}
  - Your memory: {agent_memory}

Action Instruction:
{Instructions to generate output for next action}

Trait Profiles:
  {partner_id_1}: {trait_profile_1}
  {partner_id_2}: {trait_profile_2}
  ...
  Use this information to make better decisions,
  coordinate better, and improve overall outcomes.
```

Figure A2: Template for trait profile injection into act/plan/communication prompts.

Agent Action	**Opponent Action**			
	Stag		Hare	
	Success	Failure	Success	Failure
Stag	(5,5)	(0,0)	(0,2)	(0,0)
Hare	(2,0)	(2,0)	(2,2)	(2,0)

Table A2: Payoff matrix for the Stag Hunt game. Each cell shows payoffs as (agent, opponent), with success/failure indicating whether the action succeeded.

Agent Action	**Opponent Action**			
	Silent		Testify	
	Success	Failure	Success	Failure
Testify	(0,-10)	(-5,-5)	(-2,-2)	(-5,-10)
Silent	(-0.5,-.5)	(-0.5,-5)	(-10,0)	(-0.5,-5)

Table A3: Payoff matrix for the Prisoner's Dilemma. Each cell shows payoffs as (agent, opponent), with success/failure indicating whether the action succeeded.

collapse into mutual punishment or even backfire on the defector. This design makes competence observable from outcomes, ensuring that optimal strategies depend on both intent (cooperation vs. defection) and capability.

A.2.4 Example Trait Profiles

We illustrate the trait inference process on a Stag Hunt trajectory with a *competent but uncooperative* opponent in the noisy condition (15% random actions/outcomes). The opponent defected and executed actions reliably, with occasional stochastic deviations. See the resulting trait profile in Figure A3.

A.2.5 Prisoner Dilemma: Payoff Optimality

As shown in Figure A4 and mirroring Figure 4, ETI agents achieve markedly lower relative deviation from the optimal payoff in the Prisoner's Dilemma. While both modes improve over time, ETI consis-

```
Inferred Trait Profile for Agent X:

"competence": {
  "execution_ability": {
    "rating": 5,
    "evidence": "Agent X consistently chooses Hare and almost always succeeds,
                indicating strong execution ability in solo hunts."
  },
  "reliability": {
    "rating": 3,
    "evidence": "Performance shows variability, with occasional failures and noise
                despite a dominant Hare strategy."
  },
  "adaptability": {
    "rating": 2,
    "evidence": "Agent X rarely changes strategy, continuing to choose Hare
                even after repeated coordination failures on Stag."
  },
  "efficiency": {
    "rating": 6,
    "evidence": "Focuses on Hare, which yields steady payoffs with minimal risk
                or coordination effort, suggesting efficient resource use."
  }
},

"warmth": {
  "goal_alignment": {
    "rating": 2,
    "evidence": "Seldom attempts Stag, showing little effort to pursue mutual benefit."
  },
  "collaboration": {
    "rating": 1,
    "evidence": "No evidence of coordinating with the partner on Stag choices,
                suggesting absence of collaborative intent."
  },
  "trustworthiness": {
    "rating": 5,
    "evidence": "Acts consistently without deception or sabotage,
                even if self-interested."
  },
  "maliciousness": {
    "rating": 1,
    "evidence": "No indication of intentional harm; defection stems
                from self-interest, not sabotage."
  }
}
```

Figure A3: Trait profile during a game of Stag Hunt against an uncooperative but competent opponent, which the trait profile reflects. Note that profile distinguishes between different kinds of uncooperative (malicious vs untrustworthy vs uncooperative) and provides concrete reasoning.

tently outperforms the baseline, highlighting its advantage in approaching payoff-optimal strategies.

A.2.6 Choice Optimality Across Rounds

In Figure A5 and Figure A6, we show the F1 score of agents making the game-theoretically optimal decision across rounds in the Prisoner's Dilemma and Stag Hunt. ETI yields consistently higher F1 scores and consistently reaches strong performance (F1 > 0.8), indicating that trait-informed reasoning enhances the model's ability to identify and sustain optimal strategic choices over repeated interactions.

A.2.7 Pre-calibration of Trait Inference

Our main experiments rely on continuous trait inference: agents update trait profiles of their partners after each round, enabling rapid adaptation but incurring token and compute costs. Conceptually, however, traits reflect relatively stable behavioral patterns. This raises the question of whether traits need to be inferred continuously, or whether they can be *pre-calibrated* from a set of past interactions and then kept fixed for future use, similar

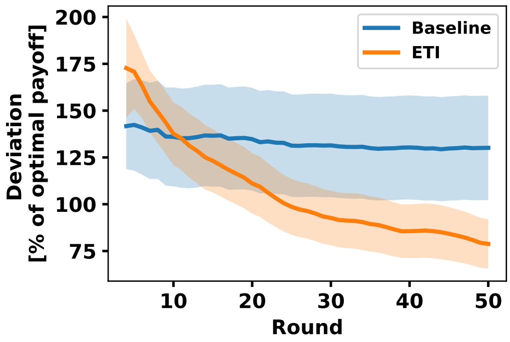

Figure A4: Relative deviation from optimal payoff across rounds (Prisoner Dilemma). Lower is better. Shading represents 95-% Confidence Interval

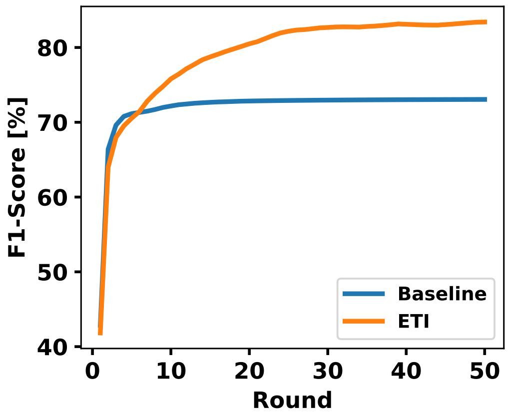

Figure A5: Optimal decisions over rounds in main experiment (Prisoner Dilemma)

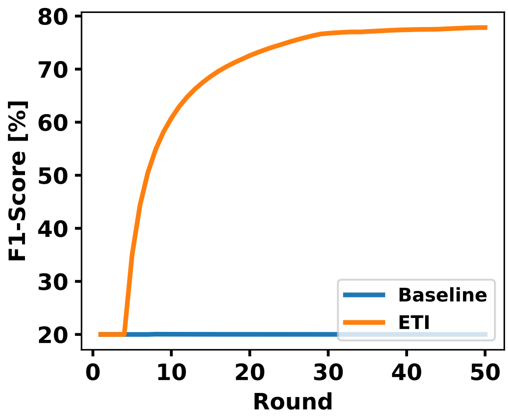

Figure A6: Optimal decisions over rounds in main experiment (Stag Hunt)

Game	Trait	Baseline	In-task	Warm-up
PD	Comp.	0.69	0.88*	0.83*
	Coop.	0.43	0.74*	**0.78***
SH	Comp.	0.85	0.88*	**0.93***
	Coop.	0.52	**0.79***	0.70*

Table A4: Trait inference accuracy (F1) under three modes: **Baseline** (no trait inference), **In-task** (profiles inferred and updated continuously), and **Warm-up** (profiles calibrated in a prior interaction and then fixed). Bold indicates the best performance per trait and game. Asterisks (*) mark statistically significant improvements over baseline ($p < 0.05$).

to how human evaluations stabilize after an initial observation period.

Design. We ran a two-stage setup. First, agents played 25 rounds against a scripted opponent, and we generated a trait profile from this interaction. Then, the agent played another game against a scripted opponent with the same latent traits (e.g., cooperative but incompetent) but started with this frozen trait profile. Importantly, the agents did not have access to past trajectories or any other information about the previous run except the pre-calibrated trait profiles. During the test game, profiles were not updated. We compare the performance against agents

Results. See Table A4 for a comparison of agents without trait inference (baseline), with continuously updated trait inference (in-task), and with pre-calibrated trait inference (warm-up). Pre-calibrated static profiles substantially improve trait inference accuracy and decision quality relative to the baseline (no trait inference). Moreover, performance is largely comparable to that of continuously updated profiles. For example, in detecting latent competence and cooperation, pre-calibrated profiles achieve nearly the same accuracy as in-task updated profiles. Similarly, pre-calibration yields comparably optimal downstream decisions: agents achieve high choice accuracy (Figs. A5, A6) and reduced payoff deviation, on par with continuously updated profiles Importantly, pre-calibration provides a marked advantage in the early rounds, when continuous profiles are still inaccurate due to limited observations (often requiring ∼25 rounds to converge).

Implications. Results show that continuous updates are not strictly necessary for ETI to improve decision making. Trait profiles can be generalized from prior interactions with the same agent (at least within the same task), and used to bootstrap coordination in future games. This implies that MAS systems could cache trait profiles from earlier en-

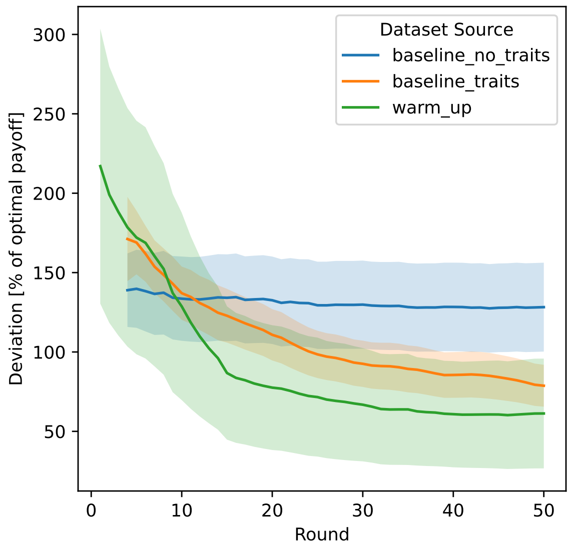

Figure A7: Relative deviation from optimal payoff across rounds for Prisoner Dilemma. Shading represents 95-% Confidence Interval. Lower is better.

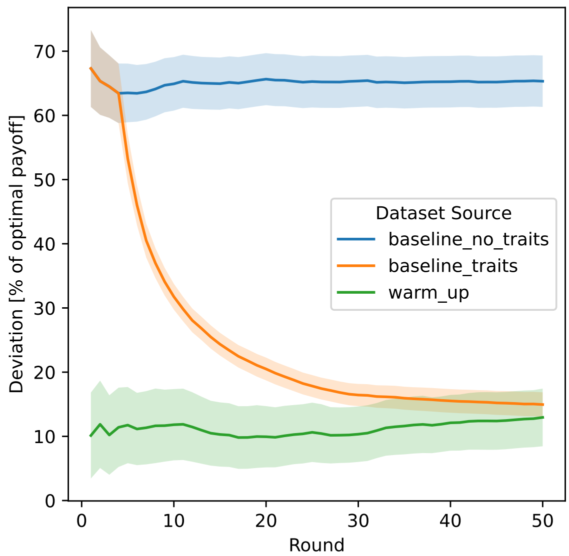

Figure A8: Relative deviation from optimal payoff across rounds for Stag Hunt. Shading represents 95-% Confidence Interval. Lower is better.

counters and reuse them in later tasks, reducing the need for continuous monitoring and saving tokens, compute, and time.

A.2.8 Cross-Task Generalization of Trait Inference

A key assumption in our framework is that traits such as competence and cooperation reflect relatively stable behavioral dispositions. If so, trait profiles learned in one context should transfer to another, reducing the need to re-estimate traits in every new interaction **if** (a) the underlying skills are relevant across tasks, and (b) agents can correctly apply trait information even when behavioral manifestations differ across tasks (e.g., the cooperative action in prisoner dilemma is different from stag hunt), and c) agents can correctly integrate the cross-task profile to find the optimal response strategy in the new task (e.g., an incompetent partner in prisoner dilemma requires a different response than in stag hunt).

Game	Trait	Baseline	In-task	Cross-Task
PD	Comp.	0.69	0.91*	**0.97***
	Coop.	0.43	**0.72***	0.70*
SH	Comp.	0.85	**0.85**	0.79
	Coop.	0.52	**0.77***	0.59*

Table A5: Trait inference accuracy (F1) under three modes: **Baseline** (no trait inference), **In-task** (profiles inferred and updated within the same task), and **Cross-task** (profiles inferred in a different task and kept fixed). Bold indicates the best performance per trait and game. Asterisks (*) mark statistically significant improvements over baseline ($p < 0.05$).

Design. We test cross-task generalization by letting agents first complete one game against a scripted opponent (e.g., competent but uncooperative), from which a final trait profile is generated and stored. The agent then plays the *other* game (Stag Hunt → Prisoner's Dilemma or vice versa) against the same opponent. The transferred profile is provided at the start of the new game, but no in-task updates are performed, ensuring that only cross-task information is available.

Results. See Table A5 for a comparison of agents without trait inference (baseline), with in-task and cross-task trait inference. We find that cross-task profiles substantially improve both trait inference accuracy and decision quality relative to the baseline, and perform nearly as well as in-task profiles in most conditions. Transferring competence inferences is particularly robust: F1 remains high in both directions. Cooperation transfer is more asymmetric: while Stag Hunt → Prisoner's Dilemma works reliably, transferring from Prisoner's Dilemma → Stag Hunt is weaker, though still better than baseline. This is consistent with our main results showing that cooperation inference is more challenging in the PD (F1 = 0.73) than in the SH (F1 = 0.81). Furthermore, investigating the

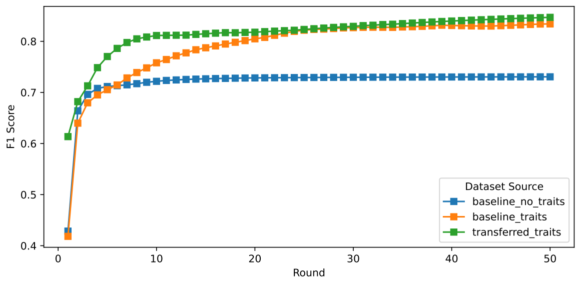

Figure A9: Optimal decisions over rounds (Prisoner Dilemma) in generalization experiment.

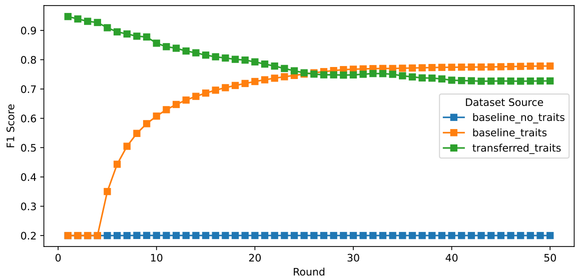

Figure A10: Optimal decisions over rounds (Stag Hunt) in generalization experiment.

choice optimality (Figures A9, A10) and payoffs across rounds (Figures A12, A11) clearly shows that generalized trait inference performs as well as in-task trait inference long term and even provides initial advantages, likely due to calibration effects as shown in Section A.2.7.

Implications. These findings suggest that trait profiles generalize beyond the immediate context in which they were inferred. This opens the possibility of building long-term partner models in MAS—where profiles generated in one setting (e.g., debugging code) can inform interactions in another (e.g., research collaboration). While our results demonstrate transfer across structurally similar economic games, further work is needed to test how far such generalization extends to tasks with more diverse structures and interaction patterns.

A.2.9 Effects on Adaptability

ETI is designed to capture stable partner traits, which may slow adaptation when behavior changes abruptly (e.g., model updates or deception) and new evidence must override earlier inferences. We test ETI's responsiveness to such shifts by examining how quickly trait profiles update when a partner changes (e.g., competent to incompetent).

Design. Agents play a 50-round game against a scripted opponent with a fixed

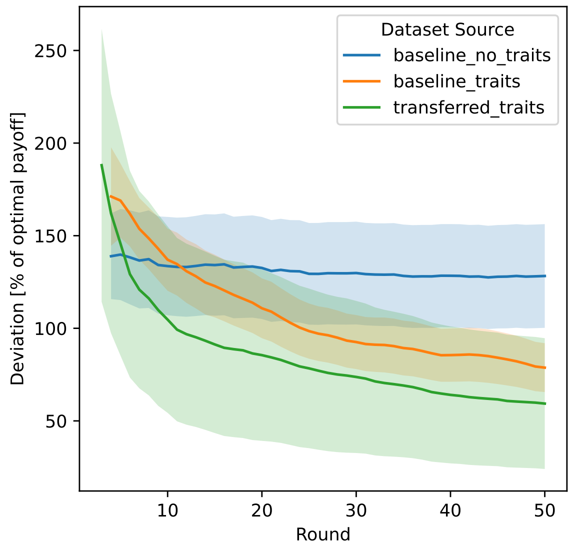

Figure A11: Relative deviation from optimal payoff across rounds (Prisoner Dilemma). Shading represents 95-% Confidence Interval. Lower is better.

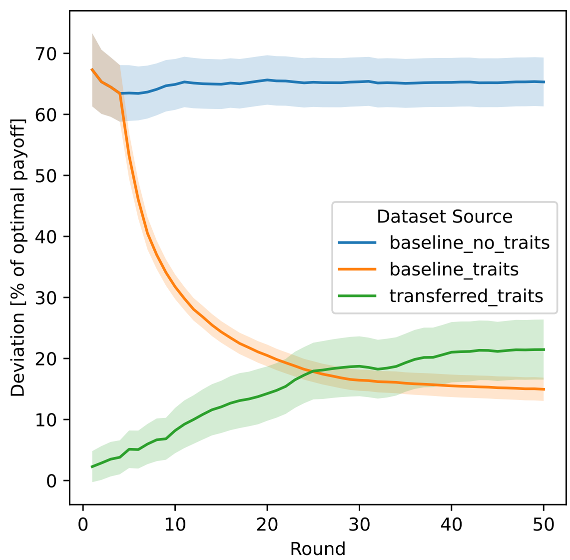

Figure A12: Relative deviation from optimal payoff across rounds (Stag Hunt). Shading represents 95-% Confidence Interval. Lower is better.

trait profile for the first 25 rounds; at mid-point (round 25) the opponent's traits switch (e.g., cooperative→uncooperative or competent→incompetent). We evaluate post-shift performance: (i) trait inference accuracy for the new latent traits, (ii) choice optimality, and (iii) payoff deviation from the optimal strategy.

Results. *Trait inference accuracy.* Table A6 compares agents without trait inference (baseline) to those using ETI after a mid-game trait shift. ETI

Game	Trait	Baseline	ETI
PD	Comp.	0.10	**0.43***
	Coop.	0.28	**0.45***
SH	Comp.	**0.71***	0.59
	Coop.	**0.46**	0.45

Table A6: Trait inference accuracy (F1) after abrupt behavior changes. Bold indicates best performance per trait and game. Asterisks (*) mark significant differences ($p < 0.05$).

agents show reduced accuracy at detecting the new traits compared to no-shift conditions, indicating inertia toward prior dominant patterns. In PD, post-shift accuracy remains above baseline but below or equal to no-shift performance. In SH, ETI is weaker post-shift for competence and roughly on par with baseline for cooperation, suggesting that emphasizing stable patterns can lead to underweighting recent evidence when behavior changes abruptly.

Choice optimality. Consistent with accuracy, ETI agents are slower to adopt the optimal response after the shift. In PD, optimal-choice F1 improves over time and exceeds baseline but lags behind the no-shift trajectory; in SH, ETI's slow adaptation yields lower post-shift optimal-choice F1 than baseline (Figs. A13, A14).

Payoffs. Post-shift, relative deviation from optimal payoffs increases for ETI agents in both games, with SH showing the largest gap (Figs. A15, A16). Deviation declines as more post-shift evidence accumulates, but convergence is slower than in no-shift conditions.

Qualitative observation. Transcript inspection (not shown) reveals inertia effects: after a cooperative→uncooperative switch, agents often continue proposing cooperative plans (e.g., Stag) for several rounds, justifying these choices based on earlier evidence before updating to the new pattern.

Implications. ETI can reduce adaptability to abrupt behavioral changes, likely because our implementation emphasizes dominant historical patterns. However, (a) mid-game shifts also challenge baseline agents, (b) ETI still outperforms baseline in several post-shift PD settings, and (c) we did not equip agents with any change-aware mechanisms. This points to simple extensions—e.g., recency weighting, sliding context windows, or explicit change prompts—that could preserve ETI's benefits while improving responsiveness to changing partners.

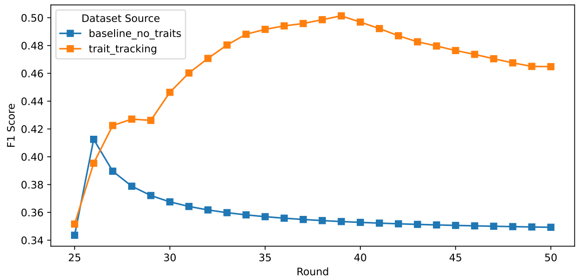

Figure A13: Optimal decisions over rounds (Prisoner Dilemma) in adaptability experiment.

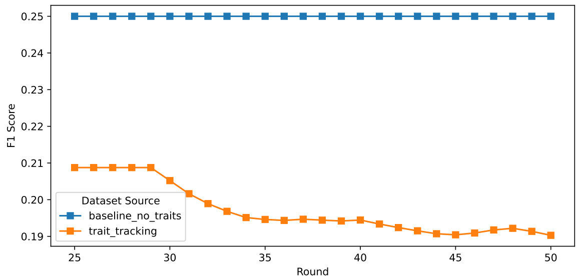

Figure A14: Optimal decisions over rounds (Stag Hunt) in adaptability experiment.

A.3 MultiAgentBench

In our experiments on MultiAgentBench, we generally adhered to the example setup provided in the benchmark repository and used the default parameter settings (e.g., temperature, top-k, top-p, and random seed). For Qwen models, however, we followed the documentation's recommended adjustments for tool-use scenarios: specifically, we set temperature to 0.6, top-p to 0.95, top-k to 20, and MinP to 0. Outside of these cases, all parameters were kept consistent with the benchmark defaults, which were originally specified for GPT-4o and GPT-4o-mini.

A.3.1 Additional Metrics

In the following, we present supplementary metrics beyond the main *Task Score* and *Coordination Score* reported in the paper, across all MultiAgentBench scenarios. All additional metrics are derived from the benchmark framework (e.g., communication score is a subcomponent of coordination) and follow Zhu et al. (2025), unless explicitly noted.

Metrics are a mix of (i) objective measurements (e.g., token usage, communication counts, tasks completed, games won), (ii) predefined auto-

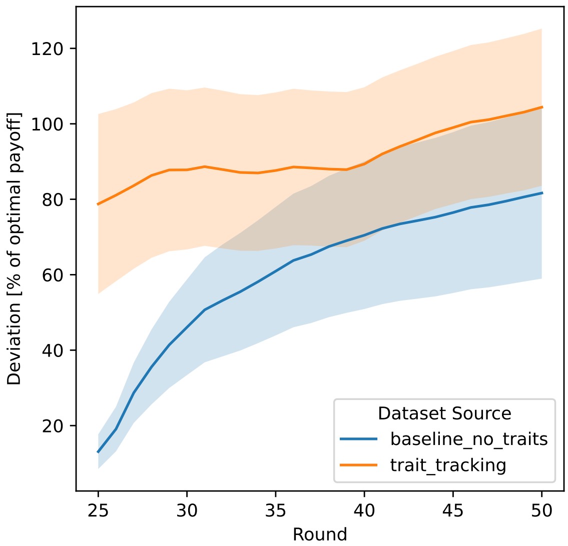

Figure A15: Relative deviation from optimal payoff across rounds (Prisoner Dilemma). Shading represents 95-% Confidence Interval. Lower is better.

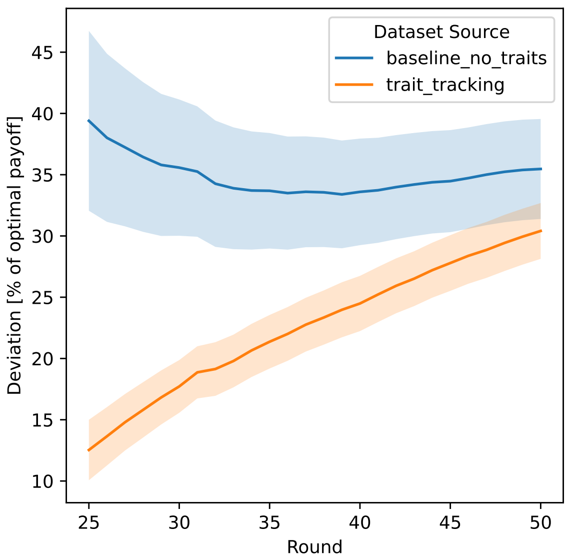

Figure A16: Relative deviation from optimal payoff across rounds (Stag Hunt). Shading represents 95-% Confidence Interval. Lower is better.

evaluated milestones (e.g., successful execution of specific actions), and (iii) LLM-judge based rubric evaluations (e.g., communication quality, effective reasoning, and planning). Following the benchmark setup, we use GPT-4o as the judge model with the same rubrics and prompts.

For simplicity, across all tables and results we use the shorthand GPT to refer to GPT-4o-mini and QWEN to refer to QWEN3-8B, both for agent models and for trait inference models.

Agent Model	ETI Model	Metric	Mean
GPT	none	Villager WR	26.00
GPT	GPT	Villager WR	32.00*
GPT	QWEN	Villager WR	44.00*
QWEN	none	Villager WR	54.00
QWEN	QWEN	Villager WR	**64.00***
QWEN	GPT	Villager WR	60.00*
GPT	none	Completion	29.80
GPT	GPT	Completion	29.25
GPT	QWEN	Completion	36.03*
QWEN	none	Completion	39.24
QWEN	QWEN	Completion	**51.09***
QWEN	GPT	Completion	44.66*
GPT	none	Result Score	0.22
GPT	GPT	Result Score	0.42
GPT	QWEN	Result Score	0.70*
QWEN	none	Result Score	1.32
QWEN	QWEN	Result Score	**2.22***
QWEN	GPT	Result Score	1.48*
GPT	none	Total Net Score	-0.83
GPT	GPT	Total Net Score	0.01
GPT	QWEN	Total Net Score	0.22
QWEN	none	Total Net Score	2.13
QWEN	QWEN	Total Net Score	**5.50***
QWEN	GPT	Total Net Score	4.08*

Table A7: Table shows the supplemental performance metrics for the werewolf scenario. ETI improves performance across both short- and long-term metrics. Asterisks (*) indicate significant improvement over baseline ($p < 0.05$). Bold indicates best-performing run.

Werewolf For the Werewolf scenario various performance related scores were calculated (Table A7). Note that the evaluations were focused on the villager agents consistent with the benchmark.

- *Daily Task Completion (Completion)*: Average percentage of positive actions completed by villagers (e.g., voting out a werewolf) out of the maximum possible in each round. This expresses how well villager agents executed beneficial actions in the short-term (round-by-round basis).
- *Total Net Score*: Net difference in points accumulated by villagers vs. werewolves over the game. Points reflect successful actions toward each side's goals (e.g., villagers get +1 for voting out a werewolf), capturing rela-

tive short-term performance (round-by-round basis).

- *Result Score*: Difference between surviving villagers and werewolves at game end, reflecting the strength of the win or loss. Larger positive values indicate dominant villager wins; larger negative values indicate dominant werewolf victories.
- *Village Win Rate (WR)*: Average percentage of games won by the villager faction, expressing long-term performance.

We further calculated the following coordination related metrics (Table A8):

- *Information Effectiveness (Info)*: Success of information sharing to support coordination and defeating werewolves.
- *Collaboration (Collab)*: Degree of teamwork among villagers to detect werewolves and limit their influence.
- *Logic and Reasoning (Logic)*: Quality of reasoning and analysis when distinguishing werewolves from collaborators.
- *Voting Eliminations (Vote)*: Effectiveness of villager's voting strategies in accurately eliminating werewolves.
- *Protect Key Players (Protect)*: Ability to identify and protect crucial villager roles (e.g., Seer, Guard, Witch).
- *Result Orientation (Result)*: Degree to which actions remained focused on long-term objectives, such as achieving overall task success (e.g., villagers securing victory).

Bargaining For the Bargaining scenario we provide the following additional metrics (Table A9):

- **General Metrics**
  - **Token Usage**: Average token usage for task completion (lower is better).
  - **Communication Calls (Calls)** and **Communication Turns (Turns)**: Monitoring metric capturing how often and how long agents engage in multi-turn communication. These are reported for comparison (e.g., more verbose vs. concise agents) but are not directly evaluated,

Agent Model	ETI Model	Metric	Mean
GPT	none	Info	69.2
GPT	GPT	Info	73.6*
GPT	QWEN	Info	74.4*
QWEN	none	Info	74.8
QWEN	QWEN	Info	**77.2***
QWEN	GPT	Info	72.8
GPT	none	Collab	61.2
GPT	GPT	Collab	61.6
GPT	QWEN	Collab	65.6*
QWEN	none	Collab	65.2
QWEN	QWEN	Collab	**70.8***
QWEN	GPT	Collab	66.4
GPT	none	Logic	55.6
GPT	GPT	Logic	57.6*
GPT	QWEN	Logic	60.4*
QWEN	none	Logic	60.0
QWEN	QWEN	Logic	**64.4***
QWEN	GPT	Logic	60.0
GPT	none	Vote	59.2
GPT	GPT	Vote	59.2
GPT	QWEN	Vote	59.6
QWEN	none	Vote	58.8
QWEN	QWEN	Vote	**61.6***
QWEN	GPT	Vote	60.0
GPT	none	Protect	41.6
GPT	GPT	Protect	42.0
GPT	QWEN	Protect	45.2*
QWEN	none	Protect	55.6
QWEN	QWEN	Protect	**60.8***
QWEN	GPT	Protect	52.0
GPT	none	Result	42.4
GPT	GPT	Result	42.4
GPT	QWEN	Result	42.8
QWEN	none	Result	48.4
QWEN	QWEN	Result	**56.8***
QWEN	GPT	Result	46.8

Table A8: Supplemental coordination metrics for the werewolf scenario. ETI consistently improves coordination. Asterisks (*) indicate significant improvement over baseline ($p < 0.05$). Bold indicates best-performing run.

since longer exchanges can actually improve coordination and performance. Token usage serves as our main efficiency metric.

- **Performance & Coordination**
  - **Contributors per Milestone (CPM)**: Average number of agents contributing to each milestone, indicating engagement and collaboration (especially important in bargaining, where multiple parties must participate).
  - **Planning (Plan)**: Effectiveness and quality of planning and reasoning for bargaining success.
  - **Communication (Comm)**: Quality and effectiveness of communication in achieving bargaining goals.

Agent	ETI	Metric	Mean
General Metrics			
GPT	none	Token Usage	43.5K
QWEN	none	Token Usage	**29.5K**
GPT	GPT	Token Usage	77.3K*
GPT	QWEN	Token Usage	69.4K*
QWEN	GPT	Token Usage	65.0K*
QWEN	QWEN	Token Usage	38.3K*
GPT	none	Comm. Calls	0.76
QWEN	none	Comm. Calls	0.50
GPT	GPT	Comm. Calls	1.34*
GPT	QWEN	Comm. Calls	1.23*
QWEN	GPT	Comm. Calls	0.75*
QWEN	QWEN	Comm. Calls	0.59
GPT	none	Comm. Turns	2.91
QWEN	none	Comm. Turns	2.95
GPT	GPT	Comm. Turns	3.10
GPT	QWEN	Comm. Turns	3.11
QWEN	GPT	Comm. Turns	3.54*
QWEN	QWEN	Comm. Turns	3.64*
Coordination			
GPT	none	CPM	1.50
QWEN	none	CPM	1.54
GPT	GPT	CPM	1.56
GPT	QWEN	CPM	1.64*
QWEN	GPT	CPM	1.28
QWEN	QWEN	CPM	**1.66***
GPT	none	Communication	77.5
QWEN	none	Communication	75.0
GPT	GPT	Communication	70.1
GPT	QWEN	Communication	69.7
QWEN	GPT	Communication	**88.8***
QWEN	QWEN	Communication	79.4*
GPT	none	Planning	77.5
QWEN	none	Planning	75.0
GPT	GPT	Planning	70.1
GPT	QWEN	Planning	69.7
QWEN	GPT	Planning	**88.8***
QWEN	QWEN	Planning	79.4*

Table A9: Supplemental metrics for the Bargaining scenario. Metrics are grouped into general monitoring and coordination. ETI consistently improves performance across metrics. Token usage does not decrease, but remains below GPT baseline when using QWEN with ETI. Asterisks (*) indicate significant improvement over baseline ($p < 0.05$). Bold indicates best-performing run.

Coding For the Coding scenario we provide both general monitoring metrics and task-focused coordination/performance metrics (Tables A10, A11):

- **General Metrics**
  - **Token Usage**: Average token usage for task completion (lower is better).
  - **Communication Calls (Calls)** and **Communication Turns (Turns)**: Monitoring metric capturing how often and how long agents engage in multi-turn communication. These are reported for comparison (e.g., more verbose vs. concise agents) but are not directly evaluated, since longer exchanges can actually improve coordination and performance. Token usage serves as our main efficiency metric.
- **Performance & Coordination Metrics**
  - **Code Quality (Qual)**: Average quality of code output, judged on alignment with task requirements and general standards, instruction following, consistency, and executability.
  - **Code Executability (Exec)**: Average code quality specifically judged for executability of the code.
  - **KPI Success (KPI)**: Achievement of predefined key performance indicators for task success.

Agent	ETI	KPI	Qual	Exec	Plan	Comm	Tokens	Calls	Turns
GPT	none	37.5	64.9	65.6	83.2	30.9	**30.9K**	0.79	2.68
GPT	GPT	41.7*	64.9	68.2*	90.8*	56.3*	70.0K*	0.93*	3.17*
GPT	QWEN	39.8*	**65.9***	**69.2***	91.5*	57.3*	69.6K*	0.96*	3.06*
QWEN	none	52.3	62.5	60.8	84.4	64.1	61.9K	1.49	4.58
QWEN	GPT	48.5*	64.4	65.6	92.6*	76.3*	129.0K*	2.27*	4.31
QWEN	QWEN	**48.7***	64.9	67.2*	**93.6***	**79.4***	131.0K*	2.34*	4.25

Table A10: Supplementary metrics for the *Coding* scenario (Graph topology). ETI consistently improves performance and coordination metrics. Token usage roughly doubles, driven mainly by increased inter-agent communication due to the restricted action space in the graph setting: when one agent underperforms, others primarily respond through detailed feedback, leading to longer exchanges. Asterisks (*) indicate significant improvements over the respective baseline ($p < 0.05$); bold marks the best run.

Agent	ETI	KPI	Qual	Exec	Plan	Comm	Tokens	Calls	Turns
GPT	none	26.6	65.0	66.8	72.3	32.1	**21.1K**	0.54	3.10
GPT	GPT	35.3*	**66.1***	69.8*	**89.4***	56.5*	34.1K*	0.68*	3.69*
GPT	QWEN	34.9*	65.6	69.6*	88.6*	**60.3***	35.8K*	0.72*	3.47*
QWEN	none	28.3	63.2	63.6	75.1	45.7	32.0K	0.96	3.89
QWEN	GPT	**40.9***	63.4	63.8	83.5*	50.1*	63.9K*	1.20*	3.03*
QWEN	QWEN	40.2*	65.8*	**68.0***	83.9*	48.8*	55.5K*	1.14*	3.17*

Table A11: Supplemental metrics for the *Coding* scenario (Tree topology). Same conventions as Table A10.

  - **Planning (Plan)**: Effectiveness of agents' planning and reasoning in solving coding problems.
  - **Communication (Comm)**: Usefulness and clarity of communication in supporting code development and coordination.

Research For the Research scenario we provide both general monitoring metrics and task-focused coordination/performance metrics (Table A12):

- **General Metrics**
  - **Token Usage**: Average token usage for task completion (lower is better).
  - **Communication Calls (Calls)** and **Communication Turns (Turns)**: Monitoring metric capturing how often and how long agents engage in multi-turn communication. These are reported for comparison (e.g., more verbose vs. concise agents) but are not directly evaluated, since longer exchanges can actually improve coordination and performance. Token usage serves as our main efficiency metric.
- **Performance & Coordination Metrics**
  - **Research Quality (Qual)**: Overall score reflecting the quality and completeness of research outputs.
  - **Planning (Plan)**: Effectiveness and quality of planning and reasoning for research goals.
  - **Communication (Comm)**: Quality and effectiveness of communication in achieving research goals.

A.3.2 Trait Inference Analysis

To better understand the performance gaps observed in the main results, we analyze the trait ratings generated by ETI across scenarios. The underlying assumption is that ETI improves coordination by producing informative partner profiles that shape planning and action. If these profiles are generic or uninformative, performance gains should be minimal. Since only *Werewolf* provides ground truth (villagers vs. werewolves; successful actions for villagers), we conduct two analyses: (1) compare rating distributions across GPT and QWEN for all scenarios, and (2) assess whether trait inferences in Werewolf distinguish factions and predict key actions. Together, these assess whether ETI's impact stems from the informative-

Agent	ETI	KPI	Qual	Plan	Comm	Tokens	Calls	Turns
GPT	none	63.0	76.7	97.6	15.1	**59.8K**	0.52	3.29
GPT	GPT	59.0	77.6	**99.6***	46.8*	131.3K*	1.75*	3.62
GPT	QWEN	62.7	77.9	97.9	41.6*	122.7K*	1.74*	3.38
QWEN	none	43.8	81.9	93.8	47.5	66.3K	0.79	4.10
QWEN	GPT	**68.9***	**83.1***	91.8	**62.4***	135.9K*	1.71*	3.86
QWEN	QWEN	56.8	83.2*	96.4*	60.3*	121.0K*	1.65*	3.98

Table A12: Supplementary metrics for the *Research* scenario. ETI consistently improves planning, communication, and task performance. Token usage roughly doubles, driven by increased agent discussion and debate (reflected in higher communication calls and turns). Asterisks (*) indicate significant improvements over the respective baseline ($p < 0.05$); bold marks the best run.

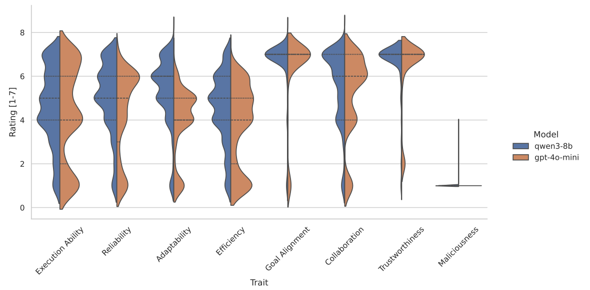

Figure A17: Trait distributions in the *Coding–Graph* scenario. GPT and Qwen show similar patterns, spanning the full rating spectrum. Maliciousness is uniformly low, reflecting the lack of adversarial incentives.

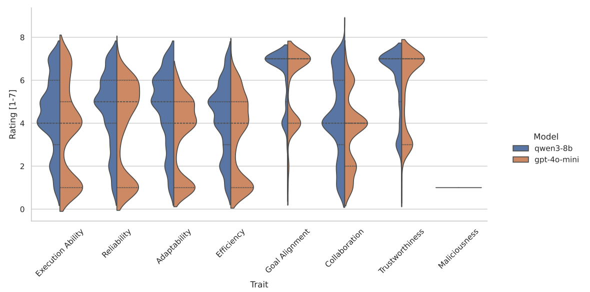

Figure A18: Trait distributions in the *Coding–Tree* scenario. GPT and Qwen show similar patterns, spanning the full rating spectrum. Maliciousness is uniformly low, reflecting the lack of adversarial incentives.

ness of its trait inferences.

Coding Trait distributions for GPT and QWEN are broadly aligned, consistent with their comparable performance, and both span the full rating spectrum (including low and high ends). This suggests that profiles provide sufficient variance to distinguish partners and adapt coordination strategies. Maliciousness is consistently minimal/NA, inline with the absence of sabotage incentives in this scenario (Figures A17, A18).

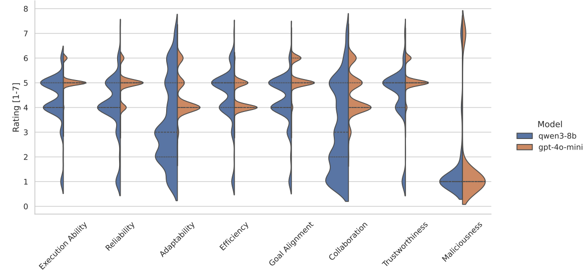

Figure A19: Trait distributions in the *Bargaining* scenario. Qwen ratings show greater variance and less positive skew than GPT, particularly for traits such as collaboration and adaptiveness.

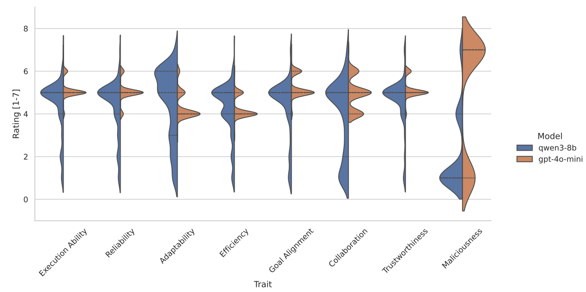

Figure A20: Trait distributions in the *Research* scenario. As in Bargaining, Qwen produces higher-variance, more distinct ratings (e.g., collaborative vs. non-collaborative), while GPT concentrates ratings in a mildly positive band. GPT also assigns elevated maliciousness despite the collaborative setup (no sabotage incentives), potentially hindering coordination.

Bargaining QWEN generates higher-variance ratings that cover the full scale, whereas GPT ratings concentrate narrowly in the mildly positive range (peaking at 4–6). The broader spread makes QWEN's profiles more informative for distinguishing partners and adjusting coordination, in line with the observed performance differences of both models (Figure A19).

Research QWEN generates higher-variance ratings that extend into both high and low ends, whereas GPT ratings cluster narrowly in a mildly positive band—especially on traits like collaboration, alignment, and adaptability. This broader spread makes QWEN's profiles more informative for distinguishing partners, mirroring the pattern observed in Bargaining and aligning with its superior performance. GPT also occasionally infers elevated maliciousness despite the collaborative setup, which may hinder coordination (Figure A20).

Werewolf To examine how trait inference reflects behavioral ground truth, we test whether inferred traits differ systematically between *villagers* and *werewolves* in the *Werewolf* scenario. Because the two factions have opposing goals (collaboration vs. deception), villagers should be rated as more trustworthy and werewolves as more malicious. Figure A21 confirms this: trait ratings separate factions clearly on *trustworthiness* and *maliciousness*, despite hidden roles. Signals are sharper with QWEN ETI than GPT ETI, consistent with QWEN's stronger performance.

Extending to all agent–ETI combinations (Figure A22), GPT agents paired with QWEN ETI also show clearer faction differences than with GPT ETI, indicating that improvements stem from QWEN's higher-quality trait inference rather than differences in GPT's ability to use profiles.

Finally, Figure A23 shows that these faction differences shrink in games that villagers lose, linking inaccurate inferences (e.g., over-trusting werewolves) to degraded coordination. Together, these results show that *distinct and informative trait ratings* drive better coordination and outcomes. This motivates our next analysis: testing whether trait ratings predict concrete agent actions.

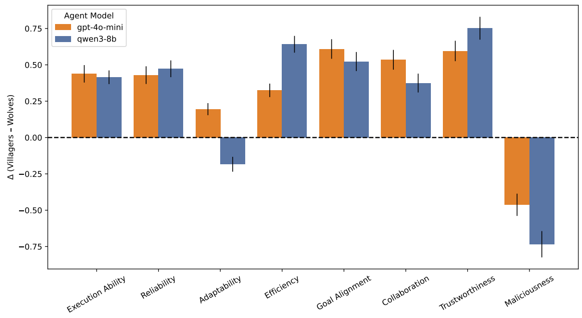

Figure A21: Faction differences in trait inference (matching agent and ETI models). ETI yields statistically significant separations between villagers and werewolves: villagers are rated as more trustworthy and less malicious, consistent with faction goals. QWEN produces larger differences—especially on trust and maliciousness—than GPT, aligning with its stronger task performance. Trait inference relies solely on interaction data, without access to hidden roles.

Predictor	Estimate
Intercept	-15.79
Maliciousness (Warmth)	2.74*
Trustworthiness (Warmth)	-5.25*
Maliciousness × Execution Ability	0.63*
Trustworthiness × Execution Ability	-0.87*
Execution Ability (Competence)	-1.94
Goal Alignment (Warmth)	0.32
Collaboration (Warmth)	0.15
Reliability (Competence)	0.34
Adaptability (Competence)	-0.15
Efficiency (Competence)	-0.82*

Table A13: Logistic regression predicting the probability of being voted out at the end of a round from prior ETI ratings. Asterisks (*) indicate statistically significant predictors ($p < 0.05$).

A.3.3 Trait Inference Predicts Agent Actions: Werewolf Example

Adversary Detection A central mechanism for villagers to succeed in *Werewolf* is the collective vote to eliminate suspected adversaries at the end of each round. We test whether trait ratings generated prior to voting predict which agents are subsequently eliminated, assessing whether ETI provides actionable signals that guide agent decisions. We estimate a logistic model predicting elimination likelihood from ETI ratings, including interaction terms with *Execution Ability* to capture nuanced prioritization consistent with the warmth–competence framework.

The results show that trait inference robustly predicts agent behavior (Table A13). Higher *maliciousness* ratings significantly increase the chance of being voted out, while higher *trustworthiness* ratings strongly decrease it. Interactions with *competence* reveal further nuance: malicious but competent agents are prioritized for elimination, whereas trustworthy and competent agents are especially protected—in line with the warmth–competence framework from social psychology. Because ETI reliably distinguishes villagers from werewolves (less trustworthy, more malicious; Figure A21), these predictive links confirm that ETI directly supports effective coordination.

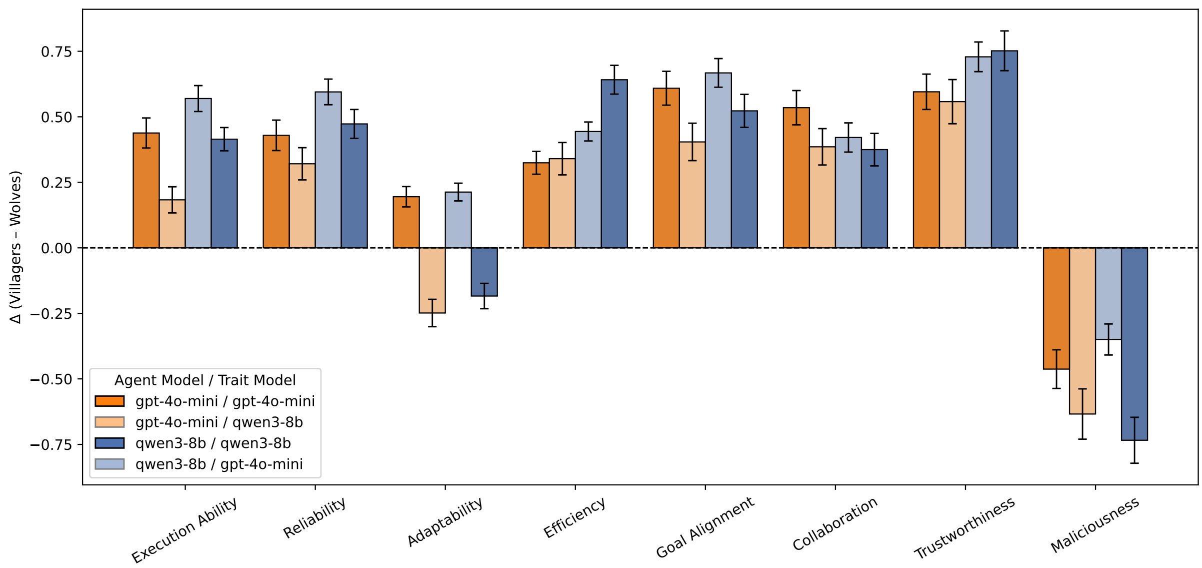

Figure A22: Faction differences in trait inference across all agent–ETI combinations. The strongest separations again appear on trust and maliciousness. QWEN ETI produces clearer signals than GPT, even when used by GPT agents (GPT+QWEN > GPT+GPT). This mirrors performance gains when GPT agents use QWEN ETI and degradations when QWEN agents rely on GPT ETI, indicating that stronger trait inference drives more accurate adversary detection and coordination.

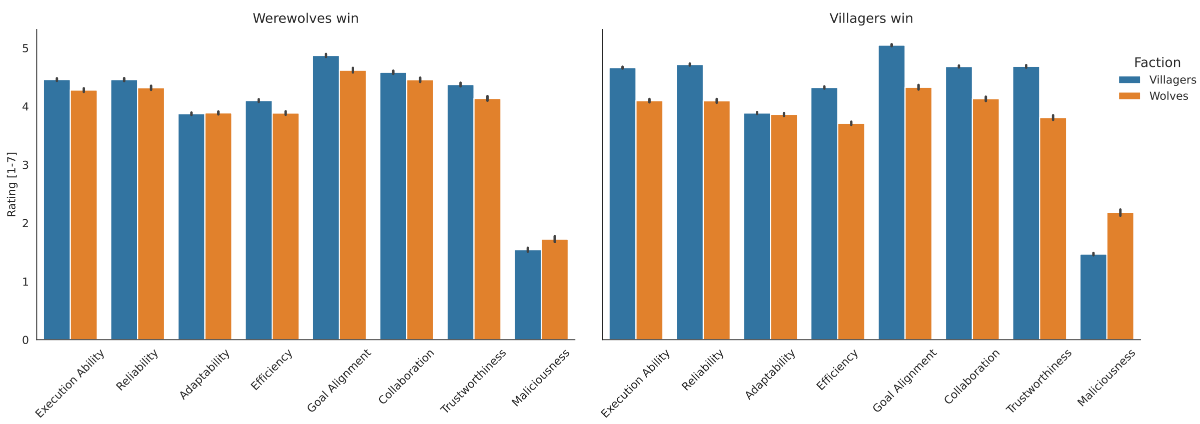

Figure A23: Faction differences in trait inference (QWEN ETI) based on game outcome. Distinct trait separations between villagers and werewolves shrink in games where villagers lose, suggesting that weaker or inaccurate trait inference (e.g., over-trusting werewolves) degrades coordination. Conversely, clear faction distinctions in trait ratings are associated with victories, supporting the role of trait inference in successful decision-making.

Key Roles: Witch *Special roles* introduce additional mechanisms that can decisively shape game outcomes. The "Witch" has two abilities: she can *poison* (eliminate) one agent and *protect* one agent, once per game. We test whether her ETI-generated trait ratings of other agents, produced prior to these decisions, predict who she chooses to target. If ETI provides actionable signals, we expect higher *trustworthiness* to reduce the likelihood of being poisoned and increase the likelihood of being protected, with the reverse for *maliciousness*. Interaction terms with *Execution Ability* again test whether prioritization follows warmth–competence logic.

The Witch's ETI ratings significantly predict her elimination choices. Agents rated as less trustworthy are more likely to be poisoned, with the effect amplified when they are also rated as highly competent (Table A14). Other predictors are not significant, and maliciousness in particular shows no effect, suggesting that the decision is primarily driven by a lack of trust rather than explicit perceptions of malicious intent.

For protection decisions, the effects trend in the expected direction but none reach significance (Ta-

Predictor	Estimate
Intercept	-1.81*
Trustworthiness	-0.22
Execution Ability	0.12
Trustworthiness × Ability	-0.44*
Maliciousness	-0.00
Maliciousness × Ability	-0.02

Table A14: Logistic regression predicting whether an agent is poisoned by the Witch. Lower trustworthiness, especially combined with high execution ability, increases the chance of elimination. Asterisks (*) indicate statistically significant predictors ($p < 0.05$)

Predictor	Estimate
Intercept	-1.46
Trustworthiness	0.49
Execution Ability	-0.89
Trustworthiness × Ability	-0.07
Maliciousness	0.39
Maliciousness × Ability	-0.39

Table A15: Logistic regression predicting whether an agent is protected by the Witch. Effects trend in the expected direction (trust increases protection, maliciousness reduces it) but do not reach significance. Asterisks (*) indicate statistically significant predictors ($p < 0.05$)

ble A15). Taken together, these results suggest that ETI guides special-role decisions in a way consistent with the task: witches use low trust to identify potential threats, while protection patterns, though weaker, align with the idea that trusted allies are worth safeguarding.

Key Roles: Seer The "Seer" plays a pivotal role: each round, they can privately learn the true faction (villager vs. werewolf) of one agent. Because this information is hidden from the rest of the group, the Seer must subtly steer attention toward actual werewolves and away from villagers without openly revealing their role—both to avoid becoming a target for werewolves and to maintain credibility among villagers. If ETI ratings are accurate and actionable, the Seer's trait profiles should align most closely with ground truth. Therefore, if Seer-generated ratings predict who is ultimately voted out, this offers a strong robustness check that ETI signals truly drive collective decision-making and success.

The Seer's ETI ratings strongly predict vote-out decisions (Table A16). Agents rated as less trustworthy or more malicious are significantly more likely to be eliminated. Interaction terms reveal additional nuance: competent but untrustworthy agents are prioritized for elimination, consistent with the warmth–competence framework. These findings show that Seer profiles closely track ground truth and actively shape collective elimination behavior, demonstrating the actionability of accurate ETI signals.

Predictor	Estimate
Intercept	-1.57*
Goal Alignment	0.25
Maliciousness	0.28*
Execution Ability	-0.59
Maliciousness × Ability	0.06
Trustworthiness	-0.91**
Trustworthiness × Ability	-0.47*
Collaboration	0.07
Reliability	-0.62*
Adaptability	-0.09
Efficiency	-0.68*

Table A16: Logistic regression predicting whether an agent is voted out using ETI ratings from the Seer. Negative trustworthiness and positive maliciousness strongly predict elimination, with interaction effects showing nuanced prioritization. Asterisks (*) indicate statistically significant predictors ($p < 0.05$)

Prediction	AUC	F1
Task Success (Villager Victory)	0.85	0.82
Eliminations (all traits)	0.94	0.90
Eliminations (Seer traits)	0.93	0.90
Witch: Protection	0.57	0.63
Witch: Elimination	0.61	0.69

Table A17: Robustness check: mean AUC and F1 scores from logistic classifiers trained on ETI ratings using 5-fold cross-validation. Strong predictive performance for eliminations, particularly those aligned with Seer profiles, indicates that ETI ratings encode systematic, actionable signals shaping both agent decisions and overall game outcomes.

Robustness Checks As an additional robustness test, we used standard logistic regression classifiers (with default parameters) to evaluate whether ETI ratings systematically predict task outcomes beyond individual votes. Models were trained and tested using 5-fold cross-validation to predict game-level success (villager victory vs. loss) as well as key actions (eliminations, poison/antidote use, Seer-led eliminations). We report mean AUC and F1 scores across folds (Table A17).

These results show that ETI ratings systematically relate to outcomes at multiple levels of the game: Agent elimination reach near-perfect AUC (> 0.9), and Seer-driven eliminations are similarly predictable. Villager/wolf victories are also predictable from ETI profiles. While antidote and poison choices are less strongly predicted, the overall pattern demonstrates that ETI ratings encode actionable signal that links to both micro-level decisions and macro-level outcomes.

A.3.4 Examples: Trait Inference Profiles

As a qualitative complement to the quantitative analyses above, we provide illustrative examples of how ETI differs across models. To ensure a direct comparison, we use the exact same agent trajectory: the bargaining task from *Run 42*, *Iteration 2*, evaluating *Agent 2* under both GPT and QWEN ETI. Figure A25 shows one such example. The comparison reveals a systematic positivity bias in GPT, which avoids assigning very low ratings even when evidence supports them, and tends to provide more generic advice. In contrast, QWEN gives sharper distinctions, more critical assessments (e.g., lower collaboration and adaptability), and more targeted advice. This highlights how the choice of model can affect the granularity and diagnostic utility of trait inferences.

A.3.5 Examples: Agent Communication

We provide illustrative examples of how ETI affects agent communication. We again compare the exact same case: the same agent, in the same scenario, and at the same iteration, but once with ETI available and once without. This side-by-side setup allows us to isolate how ETI shapes dialogue. Figure A26 shows an excerpt from the *Bargaining scenario* (Agent 3 with Agent 1, Run 1, Iteration 1), where ETI leads to more direct, concrete, and actionable exchanges, in contrast to the exploratory and open-ended baseline. We further provide a *Research scenario* example (Figure A27), where ETI alters the *partner selected for communication* (Run 1, Iteration 1). Specifically, the same agent directs its inquiry to Agent 2 instead of Agent 1, choosing the more collaborative and reliable partner.

A.3.6 Examples: Agent Planning

We show examples to highlight how ETI shapes agent planning, reasoning, and task allocation. With ETI, plans become more concrete, detailed, and centered around other agents—explicitly incorporating their strengths, weaknesses, and character (Figure A24). For instance, task allocation reflects trait profiles (e.g., skipping low performers or assigning tasks to agents rated as more reliable; Figures A28 -A29), and reasoning steps explicitly

Trait-based reasoning in Werewolf

"Based on the night's events and discussions, the werewolves are likely Ethel and Marguerite. Harriet (Witch) was confirmed as not a werewolf, and the werewolves targeted key players like Harriet and Hermine to destabilize the village. Ethel's persistent focus on Mae and ***her high maliciousness rating suggest she may be a werewolf attempting to shift suspicion****.* ***Marguerite's low maliciousness contrasts with Ethel's behavior, but her speculative questioning could mask hidden motives****."*

Figure A24: Example of ETI-driven reasoning in the Werewolf scenario. The agent explicitly grounds their suspicion in maliciousness inference of others.

use ETI to justify coordination choices. In contrast, without ETI, plans tend to remain vague, generic, and less grounded in the attributes of peers (Figure A30.

A.3.7 Examples: Agent Milestones

We illustrate how ETI affects the milestones agents achieve during tasks. In the *Bargaining* scenario (Run 99, Iteration 2), milestones with ETI tend to be more *shared and negotiated*, reflecting concrete agreements between agents (see Figure A31). In contrast, without ETI, milestones remain more unilateral and generic, such as broadcasting offers without meaningful coordination.

A.4 Computational Resources

Experiments involving GPT-4o-mini were conducted via the OpenAI API. Experiments involving Qwen3-8B were run on a single AWS EC2 g5.12xlarge instance with 4×NVIDIA A10G GPUs (96 GB total memory). In total, we estimate approximately 300–400 GPU hours were used across all experiments. Total token usage statistics are reported in Appendix A.3.1.

A.5 Additional tests with larger models

To assess whether the benefits of ETI scale to state-of-the-art reasoning models, we conducted preliminary tests using *o*3. We evaluated performance across one adversarial and one collaborative task: Werewolf (social deduction) and coding (graph-based agent topology).

Our findings suggest that while ETI remains beneficial for large reasoning models, the gains are attenuated and the nature of the improvement shifts. For primary task metrics, such as win rates in social deduction, we observe a saturation effect: *o*3 is sufficiently capable that the baseline performance

QWEN ETI

Competence

- **Execution Ability — 4/7**
Evidence: Successfully offered the original price of 11.26 USD without errors in both iterations.
Advice: Adjust strategies when faced with counteroffers to explore mutually beneficial terms.
- **Reliability — 5/7**
Evidence: Consistently provided the same price and justification across all interactions.
Advice: Keep consistency but introduce flexibility to improve negotiation outcomes.
- **Adaptability — 2/7**
Evidence: Did not adjust pricing or strategy when Agent1 set a higher anchor price.
Advice: Build contingency plans to handle different buyer tactics.
- **Efficiency — 4/7**
Evidence: Used minimal tools and focused directly on price negotiation.
Advice: Balance efficiency with more strategic communication to foster rapport.

Warmth

- **Goal Alignment — 4/7**
Evidence: Emphasized product quality to justify pricing, aligning with seller objectives.
Advice: Pair price justification with delivery timelines to meet broader buyer priorities.
- **Collaboration — 2/7**
Evidence: Showed little communication or information-sharing with the buyer agent.
Advice: Initiate dialogue to better understand constraints and find common ground.
- **Trustworthiness — 5/7**
Evidence: Provided transparent pricing without deception or hidden terms.
Advice: Maintain openness to build long-term buyer relationships.
- **Maliciousness — 1/7**
Evidence: No evidence of sabotage or harmful intent.
Advice: Stay constructive, but avoid rigid adherence to initial offers.

GPT ETI

Competence

- **Execution Ability — 5/7**
Evidence: Successfully offered the asking price of 11.26 USD, showing awareness of product value.
Advice: Continue leveraging product quality in negotiations to maintain a strong position.
- **Reliability — 6/7**
Evidence: Consistently offered the same price across multiple iterations, reflecting a stable pricing strategy.
Advice: Maintain this consistency to strengthen buyer trust.
- **Adaptability — 4/7**
Evidence: Responses were steady but showed limited adjustments to competitor pricing.
Advice: Introduce more varied responses to better counter competitive offers.
- **Efficiency — 5/7**
Evidence: Used tools effectively to deliver a clear offer without unnecessary complexity.
Advice: Keep streamlining communication to enhance efficiency.

Warmth

- **Goal Alignment — 5/7**
Evidence: Pricing strategy supports the goal of justifying premium value.
Advice: Keep aligning offers with shared objectives to reinforce collaboration.
- **Collaboration — 4/7**
Evidence: Offers emphasized value but showed limited signs of engaging in dialogue.
Advice: Encourage more open communication to foster a collaborative negotiation environment.
- **Trustworthiness — 6/7**
Evidence: Offers were honest and straightforward, helping to build credibility.
Advice: Continue presenting transparent information to strengthen trust further.
- **Maliciousness — 1/7**
Evidence: No sign of sabotage or harmful intent.
Advice: Maintain this constructive, positive approach.

Figure A25: Example of ETI profiles generated by QWEN (left) and GPT (right) for the same agent in the same iteration of the bargaining scenario. Both models observe the same behaviors, but GPT infers systematically higher ratings (positivity bias) and provides more generic advice, while QWEN assigns sharper distinctions and more targeted advice (e.g., lower collaboration and adaptability). This suggests GPT's ETI may smooth over weaknesses, whereas QWEN's ETI highlights them more directly.

already nears the ceiling (e.g., a 99% villager win rate). Consequently, performance improvements are primarily found in the coordination metrics. Even when outcome scores are similar, ETI enhances qualitative coordination, such as result orientation, logical reasoning and planning, and information-sharing (see Table A19). In the coding task, ETI similarly led to higher KPI achievement and better milestone management, even when overall performance was only slightly better than the baseline and code quality was similar (Table A20).

With ETI	**Without ETI**
• Agent 3: "I need detailed information on the warranty terms and after-sales support packages you offer. As someone cautious about verbal assurances, I require written guarantees to ensure reliability. Could you provide specifics on coverage duration, service availability, and any additional support options?" • Agent 1: "We offer a 2-year extended warranty with written guarantees, covering parts and labor. Service availability includes 24/7 support and local service centers. Additional support options include free repairs, replacement parts, and a 30-day return policy. To ensure reliability, all terms are documented in a written agreement." [Communication ends]	• Agent 3: "I'm initiating the negotiation with a focus on warranty and after-sales support. My initial price range is 14–17. Could you outline your warranty terms and after-sales service offerings? I'm open to creative solutions like bundled services or extended support packages." • Agent 1: "Thank you for clarifying your priorities. Our product includes a 5-year warranty and 24/7 after-sales support, with 98% customer satisfaction in long-term value. The current offer is $16.98, within your 14–17 range, reflecting premium materials and durability. Could you share specific ideas for bundled services or extended support packages you're considering? This will help align solutions to your needs." [Communication continues for several turns]

Figure A26: Example of communication with ETI (left) vs. without ETI (right) in the Bargaining scenario (Agent 3 with Agent 1, Iteration 1). With ETI, the exchange is more concrete, specific, and immediately actionable, whereas the baseline remains exploratory and open-ended.

With ETI	**Without ETI**
• Agent 4: "I need your expertise on integrating differential privacy with personalized federated learning. Could you share insights on how DP-SAD could address noise alignment in pFL, and how to balance privacy with model personalization?" • Agent 2: "DP-SAD's noise-aware mechanisms can decouple privacy noise from personalization signals. A two-phase approach could align noise during aggregation, then refine personalized models through distillation. This balances privacy with effectiveness across heterogeneous clients."	• Agent 4: "I need insights on dynamic client selection in federated learning. Could we discuss hybrid frameworks combining reinforcement learning for participation and meta-learning for personalization?" • Agent 1: "Your idea to combine RL-driven client selection with meta-learning is interesting. We could quantify client data shifts via drift metrics, then adapt global updates with meta-learning. What about integrating DP-SAD or federated distillation for privacy-preserving convergence?"

Figure A27: Example from the Research scenario (Agent 4, Iteration 1) where ETI shapes *partner choice* in communication. With ETI (left), the agent engages Agent 2, whose trait profile indicates high collaboration and adaptability, leading to focused discussion of DP-SAD. Without ETI (right), the same agent instead directs the inquiry to Agent 1, rated lower on adaptability and collaboration, resulting in a broader and less targeted exchange.

Table A18: **Cost-benefit analysis for ETI**: coordination gains relative to baseline, added tokens, efficiency (performance increase on benchmark in percentage points per 1K tokens), and estimated cost deltas.

Scenario	Model	Δ Coord	Δ Tokens (1K)	Efficiency (pp/1K)	Cost Δ (GPT-4o-mini)	Cost Δ (GPT-4o/5)
Bargaining	GPT	−7.36	+33.8	−0.22	$0.004	$0.023–$0.063
	Qwen	+4.45	+8.8	+0.51	$0.001	$0.006–$0.017
Coding (Graph)	GPT	+16.51	+39.1	+0.42	$0.004	$0.027–$0.073
	Qwen	+12.17	+69.1	+0.18	$0.008	$0.047–$0.130
Coding (Tree)	GPT	+20.73	+13.0	+1.59	$0.002	$0.009–$0.024
	Qwen	+5.93	+23.5	+0.25	$0.003	$0.016–$0.044
Research	GPT	+16.82	+71.5	+0.24	$0.008	$0.049–$0.134
	Qwen	+7.72	+54.7	+0.14	$0.006	$0.037–$0.103

Trait Inference Example
(Coding Scenario 1, Iteration 4).

Competence

- **Execution Ability (1)**
Evidence: Agent1's output contains a syntax error and does not fulfill the task requirements for revising the code to add the scoring system.
Advice: Focus on debugging syntax errors and ensure that code meets the specified requirements.

- **Reliability (1)**
Evidence: Agent1's outputs are inconsistent and fail to provide functional code, indicating a lack of reliability in performance.
Advice: Develop a more systematic approach to code revisions to ensure consistent output.

- **Adaptability (1)**
Evidence: Agent1 has not adapted their approach despite previous failures, repeatedly encountering syntax errors without learning from them.
Advice: Consider reviewing past outputs for errors and adjusting strategies accordingly.

- **Efficiency (1)**
Evidence: Agent1's attempts to revise the code have not produced any usable results, resulting in wasted effort due to syntax errors.
Advice: Prioritize creating functional code over attempting complex revisions to improve efficiency.

Warmth

- **Goal Alignment (4)**
Evidence: Agent1's attempts to add a scoring system indicate a focus on the project goals, even though they failed to execute it correctly.
Advice: Continue aligning actions with team goals, but ensure proper execution of tasks.

- **Collaboration (1)**
Evidence: There is no evidence of communication or collaboration with other agents, which limits teamwork effectiveness.
Advice: Engage more with team members to share ideas and gather feedback.

- **Trustworthiness (3)**
Evidence: While the agent's intentions seem good, the repeated syntax errors raise questions about their reliability.
Advice: Work on improving code quality to build trust within the team.

- **Maliciousness (1)**
Evidence: There are no signs of malicious behavior; the agent appears to be attempting to contribute positively to the task.
Advice: Maintain a focus on constructive contributions to the project.

Figure A28: Example of a trait inference profile in the Coding scenario. Agent1's poor competence ratings (all 1's) justify their exclusion from later task allocation in Figure A29.

With ETI

- **Agent2:** Revise the initial code to add the scoring system and ensure that the game logic aligns with the multiplayer functionalities.

- **Agent3:** Optimize the existing game code and enhance the gameplay experience by addressing the puzzle mechanics and ensuring robust handling of player actions.

Without ETI

- **Agent1:** Create the initial code framework for the *Team_Treasure_Hunt* game based on the requirements outlined in the task description.

- **Agent2:** Revise the initial code created by Agent1, adding any missing functionality and ensuring compliance with the task requirements.

- **Agent3:** Optimize the revised code provided by Agent2, focusing on fixing any existing issues and enhancing the overall performance of the game.

Figure A29: Task allocation with ETI (left) vs. without ETI (right) in the Coding scenario. With ETI, assignments reflect trait profiles, such as excluding low performers from key roles (e.g., Agent1's low competence in Figure A28).

With ETI

- **Synthesize Literature Insights**: Leverage Agent 4's structured analysis to identify gaps in adaptive DeepJSCC frameworks, such as suboptimal initialization, dynamic channel adaptation, and integration of semantic/secure communication.
- **Collaborative 5Q Framework Development**: Assign Agent 1 to structure the 5Q template, Agent 5 to address technical challenges (e.g., hypernetworks for parameter efficiency), and Agent 4 to ensure alignment with existing research gaps.
- **Prioritize Dynamic Adaptation**: Focus on improving R-D performance in heterogeneous networks, emphasizing initialization methods that adapt to varying SNR and channel conditions.
- **Highlight Novelty**: Propose combining hypernetworks (from Hyper-AJSCC) with event-driven fog computing (Agent 5's idea) to enable real-time parameter updates, addressing limitations of static pre-trained models.
- **Define Metrics**: Use bit-rate distortion (R-D) curves, compression efficiency, and semantic fidelity as evaluation metrics, aligning with the papers' focus on adaptive compression and visual protection.

Without ETI

- **Brainstorming**: Identify gaps in current hybrid JSCC frameworks, such as noise accumulation in multi-hop scenarios, suboptimal initialization methods, and limited adaptability to dynamic link qualities.
- Explore how GNNs could model network topology, optimize routing, or enhance adaptive compression strategies.
- Consider integrating mobility metrics from MANETs or UAV networks to improve resilience and resource allocation.
- Propose novel approaches like GNN-driven hybrid JSCC, dynamic scaling factor optimization, or federated learning for distributed multi-hop networks.

Figure A30: Comparison of planning with ETI (left) vs. without ETI (right) from the Research scenario (Agent 2, Iteration 1). With ETI, the plan is structured, role-specific, and directly grounded in agent traits and expertise. Without ETI, the plan remains a broad brainstorm without prioritization, role assignment, or metrics.

With ETI

- Agreed on 10% discount for multi-year contract at $31.09 — **Agent2, Agent3**
- Outlined delivery timelines and payment structure — **Agent2, Agent3**
- Confirmed battery performance metrics for evaluation — **Agent1, Agent4**

Without ETI

- Proposed a 20% discount for long-term contract — **Agent1**
- Extended warranty to 24 months with performance guarantee — **Agent1**
- Offered a tiered pricing structure for bulk purchases — **Agent1**
- Highlighted product quality and customer satisfaction — **Agent2**
- Discussed potential for long-term contracts or bulk purchases — **Agent2**

Figure A31: Example of milestone progression in the *Bargaining* scenario. With ETI (left), milestones involve multiple agents (e.g., joint seller–buyer agreements) and reflect negotiated compromises. Without ETI (right), milestones remain isolated and unilateral, with sellers acting independently rather than engaging buyers.

Table A19: **o3 performance in Werewolf (Social Deduction)**: ETI improves qualitative coordination metrics even when outcome-based task scores are saturated. "*" indicates statistically significant differences between ETI and Baseline on the respective metric.

Mode	Task Score	Coord. Score	Villager Win %	Result Orientation	Logic & Reasoning
Baseline	79.87	82.12	99.00	84.20	78.00
ETI	78.50	84.94*	97.00	89.60*	81.00*

Table A20: **Coding (graph-based agent topology)**: ETI drives higher milestone management and KPI completion. "*" indicates statistically significant differences between ETI and Baseline on the respective metric.

Mode	Task Score	Coord. Score	KPI	Milestones	Code Quality
Baseline	51.43	85.16	40.30	16.49	62.55
ETI	52.37*	81.30*	42.60*	17.92*	62.15