DO AGENTS DEVELOP TACIT UNDERSTANDING? FROM REPEATED COLLABORATION TO PARTNER-SPECIFIC COORDINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In long-term agent collaboration, fixed partners gradually accumulate relational knowledge about each other's response patterns and coordination habits. Human long-term partners transform such knowledge into tacit understanding, thereby reducing repeated negotiation in similar situations. Prior work shows that LLM agents can coordinate implicitly, form conventions, and adapt to partner history, yet it remains unclear what behavioral evidence distinguishes genuine dyadic tacit understanding from coordination explained by current task structure, generic experience, partner predictability, or memorization. To address this gap, we formalize Dyadic Tacit Understanding (DTU) as a directional, history-induced shift from a history-neutral coordination-response law, and define Genuine Dyadic Tacit Understanding (GTU) by the joint requirements of Formation, Interaction Specificity, and Cross-Instance Generalization. We further propose a decentralized online relational learning framework that enables each agent to independently learn its partner's coordination-response patterns and transform them into private distributional guidance for its own decisions under limited explicit communication. We validate the proposed theory and framework on multiple collaborative benchmarks, showing that agents can develop reusable partner-specific coordination knowledge from shared experience while maintaining collaboration quality with reduced explicit communication. Code is available at https://anonymous.4open.science/r/tact-agents-F45C.

1 INTRODUCTION

Long-term human partners often develop increasingly natural coordination as shared experience accumulates (Rico et al., 2008; Hawkins et al., 2020). Repeated interaction allows them to learn how one another tends to respond in recurring situations: one partner may prepare a handoff before being asked, while the other completes the preceding task. Coordination that initially requires explicit negotiation can thus become an implicit convention (Brennan & Clark, 1996; Metzing & Brennan, 2003). Such tacit understanding need not involve access to a partner's internal thoughts. It is expressed through relational expectations formed through shared experience and reused in new but structurally similar situations.

LLM agents can likewise operate in persistent pairs, such as a planner and an executor, across structurally related but instance-specific tasks (Zhang et al., 2024b; Qian et al., 2024). This raises a basic question: **Can persistent LLM agents develop reusable, partner-specific tacit coordination through repeated shared experience?**

Figure 1 illustrates the motivating phenomenon: after repeated collaboration, a persistent dyad may reuse relational expectations accumulated from prior interaction to coordinate on a new task with less explicit negotiation. Yet observing such behavior does not establish that the dyad has formed partner-specific tacit understanding. The current state, role assignment, and environmental constraints may already make a partner's behavior highly predictable, so better prediction need not imply newly acquired relational knowledge. Gains from long-context history or retrieval may instead reflect memory of specific tasks or trajectories, while reduced communication may arise from task determinism or omitted information (Shaikh et al., 2024). These observations therefore can-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

not by themselves distinguish partner-specific relational adaptation from generic task structure or memory reuse (Lowe et al., 2019; Wang et al., 2026).

Existing LLM multi-agent research mainly improves collaboration through explicit communication, long-term memory, and partner modeling. Communication methods reduce coordination overhead by modifying message content, timing, or interaction structure (Zhang et al., 2025; 2024a; Wang et al., 2025); memory-based methods leverage long contexts, reflection, or episodic retrieval (Park et al., 2023; Shinn et al., 2023; Packer et al., 2023; Zhang et al., 2026); and partner-modeling methods adapt decisions by reasoning about other agents' actions, intentions, or policies (Cross et al., 2025; Mu et al., 2026). Recent work also studies convention formation in LLM populations and repeated reference games (Ashery et al., 2025; Vaduguru et al., 2026), and tests whether apparent coordination depends on a particular partner history (Wang et al., 2026; Gao et al., 2026). The key issue is whether a persistent dyad develops a directional coordination relation that exhibits formation, interaction specificity, and cross-instance generalization, rather than merely retaining prior interaction history.

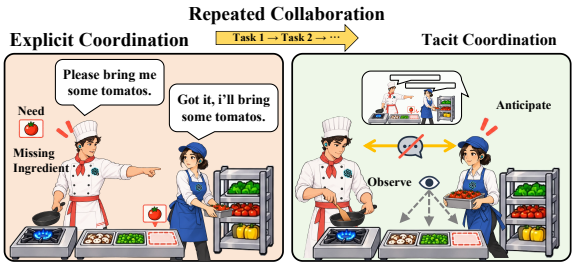


Table 1: From Communication to Tacit Coordination.

For a persistent LLM dyad $\mathcal{D} = (A, B)$, we formalize genuine directional tacit understanding as $\mathrm{GTU}^{(n)}_{i\to j} = \mathrm{Form}^{(n)}_{i\to j} \wedge \mathrm{Spec}^{(n)}_{i\to j} \wedge \mathrm{Gen}^{(n)}_{i\to j}$. The relation must emerge through repeated collaboration, depend on the actual i–j interaction structure, and remain reusable on unseen instances from the same coordination family. History dependence alone is therefore insufficient. To make this distinction testable, we represent partner behavior through a structured *Coordination Response* $C = (R, G, T^{\mathrm{resp}})$ that is comparable across task instances, and define the *Dyad-Specific Interaction Residual* (DSIR) as the response-law shift induced by real shared history relative to a history-neutral reference. Building on this characterization, we propose a decentralized online relational learning framework for frozen black-box LLM agents. Each agent independently maintains a directional *Tacit State* from legally observable interactions and converts relations repeatedly supported across task instances into private self-side coordination priors. The two learners share neither writable adaptive states nor gradients, and primitive actions remain selected by the frozen LLMs.

Our contributions are as follows:

- We formalize dyadic tacit understanding as directional response shifts from a history-neutral reference, identified through Formation, Interaction Specificity, and Cross-Instance Generalization.
- We introduce a decentralized relational learner for frozen black-box LLM agents that learns coordination-response distributions from shared experience, prospectively qualifies their use, and converts them into private coordination guidance.
- We conduct extensive experiments across multiple collaborative benchmarks and LLM backbones, validating dyadic tacit formation and showing that the learned relational guidance improves coordination under constrained communication.

2 Related Work

LLM collaboration builds on communication design, experience reuse, and partner-aware decision-making. AgentPrune and AgentDropout reduce interaction overhead through message-graph pruning and agent elimination (Zhang et al., 2025; Wang et al., 2025). Memory-based approaches retain, reflect on, or retrieve prior experience (Park et al., 2023; Shinn et al., 2023; Packer et al., 2023), while Experiential Co-Learning extracts reusable shortcuts from collaborative trajectories for unseen tasks (Qian et al., 2024). Hypothetical Minds and A-ToM use interaction evidence to refine hypotheses about a partner's strategy or estimate its reasoning depth (Cross et al., 2025; Mu et al., 2026). In complementary MARL settings, Other-Play targets zero-shot compatibility with unfamiliar partners (Hu et al., 2020), whereas few-shot coordination studies adaptation after limited interaction

(Nekoei et al., 2023). These methods establish ways to exploit experience and partner information, but improved coordination alone does not identify whether the resulting behavior reflects relational knowledge formed within a particular pair.

Related behavioral studies examine syntactic adaptation (Kandra et al., 2025), communication-free focal-point coordination (Aharon et al., 2026), and convention formation in repeated reference games (Vaduguru et al., 2026). Crucially, Wang et al. (2026) find similar lexical alignment in real and matched pseudo-dyad corpora, showing that surface convergence need not depend on shared partner history. Complementarily, Gao et al. (2026) report increased communication costs after role-matched agent swaps on held-out tasks following repeated collaboration. Therefore, it remains unresolved whether repeated collaboration induces a reusable, partner-specific directional coordination relation in persistent LLM dyads, rather than a behavioral shift explained by generic task structure, history reuse, or partner predictability. Addressing this question requires jointly establishing formation, interaction specificity, and cross-instance generalization, while separating these identification criteria from state attribution and downstream communication utility.

3 Background and Problem Formulation

3.1 Problem Definition

Consider a persistent dyad $\mathcal{D} = (A, B)$, where both agents are implemented by frozen black-box LLMs. They repeatedly collaborate on distinct task instances $T_1, \ldots, T_n \sim \mathcal{F}$ from a pre-specified coordination family $\mathcal{F}$, whose instances share high-level role, dependency, and handoff structure while differing in surface realization.

For either direction $i \to j$, let $H_{i\to j}^{(n)}$ denote the legally observable interaction history accumulated by agent i about partner j after n collaborations. The histories need not coincide because observations are agent-specific. We define tacit coordination over an observable structured *Coordination Response*, rather than the partner's next primitive action:

$$C_{j|i,t} = \Psi_{\mathcal{F}}\Big(X_t, \Gamma_{ij,t}^{(H_c)}\Big) = \Big(R_{j|i,t}, G_{j|i,t}, T_{j|i,t}^{\text{resp}}\Big). \tag{1}$$

Here, X_t contains information legally available to the focal agent before the relevant decision, and $\Gamma_{ij,t}^{(H_c)}$ is the subsequent bounded observable response window, where $H_c \subseteq H_{i\to j}^{(n)}$ denotes the bounded window of prior i–j interaction history that is legally observable and used to condition the response. R describes the relation to the current coordination need, G the coordination target, and T^{resp} the temporal or initiative mode, inferred from observable evidence rather than unexecuted partner intentions.

3.2 Core Theory of Genuine Dyadic Tacit Understanding

The current coordination context may already explain much of the partner's behavior. We therefore distinguish history-neutral and history-enabled response laws:

$$\begin{aligned} p_0^{\star}(c \mid x) &= P(C_{j|i} = c \mid X = x, \eta = 0), \\ p_{i\to j}^{\star(n)}(c \mid x, h_n) &= P\Big(C_{j|i} = c \mid X = x, H_{i\to j}^{(n)} = h_n, \eta = 1\Big), \end{aligned} \tag{2}$$

where $\eta = 0$ excludes information from prior i–j interactions, whereas $\eta = 1$ allows the realized dyadic history to affect the response. Their response-specific shift is the *Dyad-Specific Interaction Residual* (DSIR):

$$\rho_{i\to j}^{\star(n)}(c; x, h_n) = \log \frac{p_{i\to j}^{\star(n)}(c \mid x, h_n)}{p_0^{\star}(c \mid x)}. \tag{3}$$

Let $\mu_{\mathcal{F}}^{i\to j}$ be a fixed directional audit distribution shared across checkpoints and history regimes. We define the *Directional Dyadic Response-Shift Strength* as

$$S_{i\to j}^{(n)} = \mathbb{E}_{X\sim\mu_{\mathcal{F}}^{i\to j}}\Big[D_{\text{KL}}\Big(p_{i\to j}^{\star(n)}(\cdot \mid X, h_n)\Big\|p_0^{\star}(\cdot \mid X)\Big)\Big]. \tag{4}$$

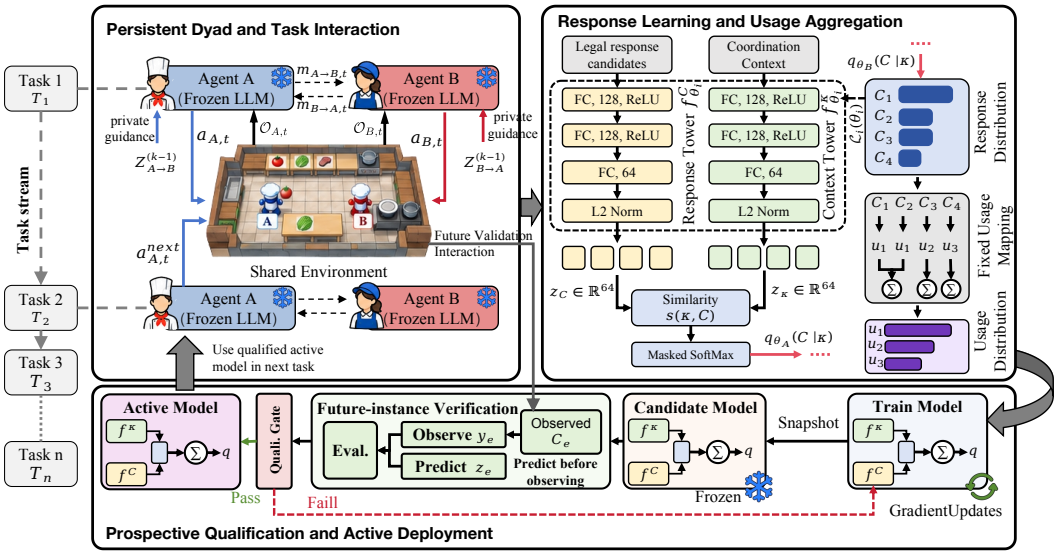


Figure 1: From Communication to Tacit Coordination. Repeated same-family collaboration may allow Assistant B to combine the current context with its directional Tacit State to coordinate proactively on an unseen task without routine explicit communication.

Under response measurability and support overlap, $S^{(n)}_{i\to j} \geq 0$, with equality iff the two response laws coincide for $\mu^{i\to j}_{\mathcal{F}}$-almost every audited context. At cold start, $p^{\star(0)}_{i\to j}(\cdot \mid x, \varnothing) = p^{\star}_{0}(\cdot \mid x)$, hence $S^{(0)}_{i\to j} = 0$.

A positive response shift alone is insufficient for genuine tacit understanding because it may reflect generic same-family experience, response marginals, or instance memorization. We therefore require three joint conditions. **Formation** requires the shift to emerge through repeated real collaboration. **Interaction Specificity** requires the induced relation to be distinguishable both from matched neutralized histories and from a role-matched relation of another dyad: $\mathrm{Spec}^{(n)}_{i\to j} = \mathrm{Corr}^{(n)}_{i\to j} \wedge \mathrm{Pair}^{(n)}_{i\to j}$. **Cross-Instance Generalization** requires the relation to remain detectable on unseen instances from the same coordination family. Accordingly,

$$\begin{aligned} \mathrm{GTU}^{(n)}_{i\to j} &= \mathrm{Form}^{(n)}_{i\to j} \wedge \mathrm{Spec}^{(n)}_{i\to j} \wedge \mathrm{Gen}^{(n)}_{i\to j}, \\ \mathrm{GTU}^{(n)}_{AB} &= \mathrm{GTU}^{(n)}_{A\to B} \wedge \mathrm{GTU}^{(n)}_{B\to A}. \end{aligned} \tag{5}$$

GTU is defined independently of task utility, attribution, and external-state interventions; communication reduction, task gains, and latency are downstream outcomes. The starred response laws are frozen-audit behavioral quantities, distinct from the online predictor $q_\theta(C \mid \kappa)$, and prediction or qualification alone does not establish GTU.

4 Proposed Method

We propose a decentralized framework that converts shared experience into private coordination guidance (Figure 1). Each directional Tacit State $Z^{(k)}_{i\to j} = (\theta^{\mathrm{act}(k)}_i, \mathcal{E}^{\mathrm{act}(k)}_i)$ contains an active model and version-matched qualification and monitoring summaries. Learners share neither writable state nor gradients, and the base LLMs remain frozen. GTU is assessed separately under the criteria in Section 3.

4.1 Learning and Representing Coordination Relations

Agent i encodes its predecision context as $\kappa_{i,t} = \Phi_{\kappa,\mathcal{F}}(X_{i,t})$ and models the complete response C defined in Eq. equation 1. The fixed key $\omega = K_{\mathcal{F}}(\kappa)$ groups comparable contexts. Unit-normalized

encoders define $s_{\theta_i}(\kappa, C) = \langle f^{\kappa}_{\theta_i}(\kappa), f^{C}_{\theta_i}(C)\rangle$, yielding the conditional model and training objective:

$$q_{\theta_i}(C \mid \kappa) = \frac{\exp(s_{\theta_i}(\kappa, C)/\tau)}{\sum_{C' \in A_{\mathcal{F}}(\kappa)} \exp(s_{\theta_i}(\kappa, C')/\tau)}, \quad C \in A_{\mathcal{F}}(\kappa), \tag{6}$$

$$\mathcal{L}_i(\theta_i) = \mathbb{E}_{e \sim \mathrm{Bal}(\mathcal{D}_i)} \left[-\log q_{\theta_i}(C_e \mid \kappa_e) \right]. \tag{7}$$

Here, $\tau > 0$ and the distribution is supported on $A_{\mathcal{F}}(\kappa)$, determined by predefined semantics and current legal information. After each instance, training uses $\mathcal{D}_i$, containing complete observable records from that instance and bounded rehearsal, regardless of task success. Bal samples uniformly over valid instances, then their keys, then records within each key. Instance IDs remain metadata. Raw dialogue is reset across instances, and rehearsal never enters prompts.

A fixed usage mapping $U_i = \mathcal{U}_i(\kappa, C)$ groups responses with equivalent implications for agent i's participation. We sum their probabilities rather than select a single response:

$$q^{U}_{\theta_i}(u \mid \kappa) = \sum_{C \in A_{\mathcal{F}}(\kappa)} \mathbf{1}\{\mathcal{U}_i(\kappa, C) = u\}\, q_{\theta_i}(C \mid \kappa). \tag{8}$$

Its image $A^{U}_i(\kappa)$ defines the usage labels shared by forecasts and guidance. Training and audits retain complete C.

As a concrete illustration, in Overcooked, $R \in \{\textit{accept}, \textit{initiate}, \textit{defer}\}$ records j's handoff response mode; G identifies the target station or ingredient; and T^{resp} indicates whether j's action precedes i's. $\mathcal{U}_i$ retains (R, T^{resp}) and discards G, since the target identity does not affect i's participation timing.

4.2 Prospective Qualification and Private Guidance

Each learner maintains a *train* model, a frozen *candidate*, and at most one *active* copy. After W warmup instances, a train snapshot becomes candidate v and is frozen for the next L_{val} instances. Candidate forecasts precede responses, never enter prompts, and are evaluated before train is updated.

Within each cell $\gamma = V_i(\kappa)$, the first eligible opportunity per instance is selected before its response; missing probes remain in the coverage denominator and are not replaced. Forecasts $z_e = q^{U}_{\theta_{i,v}}(\cdot \mid \kappa_e)$ are scored against $y_e = \mathcal{U}_i(\kappa_e, C_e)$ using $\ell_B(z_e, y_e) = \frac{1}{2}\sum_u (z_{e,u} - \mathbf{1}\{u = y_e\})^2$. Let $\widehat{L}_{v,\gamma}$ be the mean loss over the $m_{v,\gamma}$ fully observed probes, $\widehat{\mathrm{cov}}_{v,\gamma}$ their coverage among selected probes, and $n^{\mathrm{train}}_{v,\gamma}$ the number of distinct training instances in the cell. At each fixed block boundary,

$$Q_{v,\gamma} = \mathbf{1}\left\{ \begin{array}{c} n^{\mathrm{train}}_{v,\gamma} \geq m_{\mathrm{train}}, \quad m_{v,\gamma} \geq m_{\mathrm{val}}, \quad \widehat{\mathrm{cov}}_{v,\gamma} \geq \mathrm{cov}_{\min}, \\ \widehat{L}_{v,\gamma} \leq \min\left\{ L^{\max}_{\gamma}, \widehat{L}^{\mathrm{unif}}_{v,\gamma} + \epsilon_{\mathrm{ref}} \right\} \end{array} \right\}. \tag{9}$$

Here, $\widehat{L}^{\mathrm{unif}}_{v,\gamma}$ is the loss of uniform forecasts over each probe's valid usage set, not the behavioral reference $p^{\star}_0$. Thresholds and mappings are fixed before verification, and use also requires valid support and matching model/semantic versions. Qualification is empirical rather than a risk guarantee or GTU test.

If any cell qualifies, candidate v replaces active as a whole for the next instance but is usable only in its qualified cells; otherwise, a valid incumbent remains or guidance is empty. Each block starts a fresh candidate from train. The active model and enabled-cell set $\mathcal{G}^{\mathrm{act}(k-1)}_i$ remain fixed within T_k, with no active copy initially. For an enabled cell,

$$g^{\mathrm{tac}}_{i,t} = \Pi^{\mathrm{dist}}_i \left(\kappa_{i,t}, q^{U}_{\theta^{\mathrm{act}(k-1)}_i}(\cdot \mid \kappa_{i,t}), \mathcal{E}^{\mathrm{act}(k-1)}_i \right), \quad \gamma_{i,t} \in \mathcal{G}^{\mathrm{act}(k-1)}_i. \tag{10}$$

Here, Π^{dist}_i maps the current context $\kappa_{i,t}$, the active usage forecast q^{U}, and the qualification evidence $\mathcal{E}^{\mathrm{act}}_i$ into a natural-language directive that is injected into agent i's prompt. The directive summarizes the available usage alternatives and their relative likelihoods without exposing the partner's private plans. Otherwise, guidance is empty. Guidance expresses usage alternatives and uncertainty without exposing private partner plans or action sequences; the frozen LLM still selects actions and messages under the common rules, with no prescribed message suppression.

Table 2: Collaboration quality under standard and constrained communication. Mean ± SD over eight runs per backbone at $n = 48$. Blue/gray indicate NI established/not established for $J(\beta_*)$ against same-backbone full-budget B0, with B0 shown in gray.

		Claude Sonnet 4.6				GPT-5.5				Gemini 3.6 Flash			
	Meth.	$J(\text{fu})\uparrow$	$J(\beta_*)\uparrow$	$L(\text{fu})\downarrow$	$L(\beta_*)\downarrow$	$J(\text{fu})\uparrow$	$J(\beta_*)\uparrow$	$L(\text{fu})\downarrow$	$L(\beta_*)\downarrow$	$J(\text{fu})\uparrow$	$J(\beta_*)\uparrow$	$L(\text{fu})\downarrow$	$L(\beta_*)\downarrow$
Overcooked	B0	48.93±3.51	33.78±4.12	5.00±0.13	2.38±0.06	46.37±3.67	33.53±2.49	5.32±0.11	2.48±0.07	46.31±2.76	32.04±4.38	4.73±0.16	2.41±0.16
	B1	56.77±3.62	42.63±5.03	5.32±0.15	2.63±0.10	53.83±4.72	40.61±3.80	5.40±0.21	2.62±0.23	52.24±4.27	40.01±1.99	5.25±0.36	2.43±0.13
	B2	56.42±1.54	42.74±2.10	5.16±0.27	2.48±0.15	54.86±3.62	40.56±4.14	5.15±0.20	2.65±0.12	53.99±1.74	41.80±3.58	4.89±0.14	**2.22**±0.10
	B3	57.52±4.77	46.78±3.67	5.11±0.38	2.48±0.14	54.90±2.91	45.21±2.18	5.29±0.14	2.48±0.13	52.18±2.95	41.89±4.33	4.84±0.12	2.36±0.06
	B4	56.63±3.45	47.40±2.57	5.15±0.18	**2.30**±0.14	57.22±4.43	47.79±4.56	5.05±0.20	2.43±0.25	52.73±3.44	44.92±3.90	4.72±0.22	2.38±0.19
	B5	58.67±2.09	50.07±2.20	4.82±0.10	2.40±0.18	**59.31**±1.66	51.59±2.26	5.09±0.10	2.61±0.10	53.90±3.21	47.12±5.45	4.72±0.07	2.25±0.20
	B6	**64.36**±4.31	52.89±2.98	5.07±0.23	2.33±0.17	58.83±1.58	53.14±2.01	4.97±0.25	2.49±0.06	56.17±3.50	51.04±3.49	**4.61**±0.21	2.25±0.18
	Ours	59.70±3.99	**57.93**±3.94	**4.74**±0.23	2.39±0.17	56.87±3.06	**54.38**±3.48	**4.76**±0.24	**2.32**±0.14	**57.91**±4.70	**55.99**±4.51	4.79±0.31	2.31±0.23
CoBlock	B0	50.61±2.99	34.07±3.39	4.98±0.12	2.09±0.14	50.24±2.84	29.62±2.12	5.07±0.25	2.26±0.18	47.44±2.54	33.06±4.56	4.53±0.25	2.12±0.15
	B1	52.92±2.17	37.57±1.59	5.02±0.06	2.42±0.10	54.01±1.75	40.49±2.38	5.32±0.20	2.43±0.10	49.06±2.81	36.49±4.05	4.68±0.27	2.19±0.18
	B2	53.15±2.14	40.45±2.31	4.85±0.25	2.23±0.19	53.09±4.35	39.93±4.34	4.82±0.29	2.20±0.13	50.51±3.71	39.28±4.01	4.60±0.28	2.05±0.14
	B3	57.63±2.66	45.25±3.65	4.81±0.22	2.25±0.22	54.99±3.26	45.23±4.25	**4.69**±0.23	2.20±0.14	51.68±4.01	42.25±3.09	4.53±0.31	2.16±0.10
	B4	58.00±4.52	49.25±2.41	4.89±0.13	2.21±0.08	55.01±2.95	45.35±2.74	5.10±0.35	2.21±0.06	52.06±3.73	43.86±4.22	**4.44**±0.29	**2.02**±0.24
	B5	**58.43**±4.53	50.54±4.96	**4.68**±0.14	2.21±0.06	55.91±3.48	47.05±1.61	4.90±0.15	2.27±0.05	53.26±4.53	46.08±3.74	4.65±0.17	2.04±0.20
	B6	55.90±4.46	51.05±4.19	4.72±0.15	**2.06**±0.14	55.58±2.71	47.61±3.54	5.00±0.20	**2.16**±0.09	**53.64**±3.90	45.72±4.07	4.47±0.07	2.09±0.06
	Ours	57.64±4.01	**55.23**±4.06	4.69±0.20	2.16±0.12	**57.32**±3.23	**54.03**±3.52	4.83±0.22	2.25±0.22	52.74±3.54	**48.62**±3.17	4.51±0.12	2.08±0.20
Hanabi	B0	13.05±1.09	N/A	7.00±0.27	N/A	10.89±0.74	N/A	7.78±0.14	N/A	9.18±0.74	N/A	7.08±0.26	N/A
	B1	13.17±0.68	N/A	7.53±0.23	N/A	11.58±1.06	N/A	7.85±0.15	N/A	9.38±0.85	N/A	7.11±0.26	N/A
	B2	13.12±1.04	N/A	7.31±0.17	N/A	11.81±0.60	N/A	7.28±0.26	N/A	8.77±0.38	N/A	6.80±0.18	N/A
	B3	14.42±0.66	N/A	7.19±0.19	N/A	12.53±0.72	N/A	7.36±0.28	N/A	10.34±0.67	N/A	6.72±0.15	N/A
	B4	14.31±0.77	N/A	6.97±0.23	N/A	13.98±0.89	N/A	7.21±0.18	N/A	11.00±0.81	N/A	6.63±0.11	N/A
	B5	15.41±0.81	N/A	6.84±0.25	N/A	13.49±0.41	N/A	7.12±0.20	N/A	11.26±0.46	N/A	**6.35**±0.12	N/A
	B6	14.96±0.69	N/A	6.99±0.20	N/A	13.69±0.72	N/A	**7.04**±0.13	N/A	11.52±0.49	N/A	6.69±0.21	N/A
	Ours	**16.24**±1.17	N/A	**6.77**±0.22	N/A	**14.45**±1.23	N/A	7.11±0.13	N/A	**12.49**±0.77	N/A	6.48±0.13	N/A

Because deployment may shift partner responses, active cells are monitored in fixed recent-instance windows using the same probe, loss, and coverage rules after minimum observed support. Cells that fail these checks or lack valid observations for a fixed period are disabled from the next instance for the remainder of that version. Audits freeze deployed states and all updates, including expiry, and audit or utility data never enter training, qualification, monitoring, or rehearsal.

5 Experiments

5.1 Experimental Design and Evaluation

Tasks and models. We evaluate two-agent collaboration in Collab-Overcooked (Sun et al., 2025), Hanabi (Bard et al., 2020), and CoBlock (Wu et al., 2024) , using fixed coordination families centered on cross-role handoffs, two-player cooperation, and complementary-resource coordination, respectively, with disjoint data splits. For utility evaluation, Overcooked and CoBlock cap messages at $\beta_* = 0.5$ of B0's development median, with a non-inferiority margin $\epsilon = 5$ pp. Hanabi retains its native hint rules. All main comparisons use homogeneous dyads of Claude Sonnet 4.6, GPT-5.5, and Gemini 3.6 Flash, with eight runs per dataset and backbone and fixed inference settings within each backbone.

Baselines. We compare against B0 Frozen, B1 Raw History, B2 Reflection (Shinn et al., 2023), B3 Partner Notes, B4 Structured Retrieval, and adapted B5 Hypothetical Minds (Cross et al., 2025) and B6 A-ToM (Mu et al., 2026), under common external interfaces while preserving method-specific adaptation and histories.

Implementation details. Each adaptive method completes $N = 48$ formation instances under standard communication, with checkpoints at $\mathcal{N} = \{0, 2, 4, 8, 12, 16, 24, 32, 48\}$. Evaluation freezes deployed states rather than the latest training weights. Tasks are paired across conditions, API draws are independent, and evaluation never updates persistent state.

Relational-state controls. At $n = N$, we compare frozen states on complete held-out utility tasks in Overcooked and CoBlock at β_*. *Real* retains both agents' deployed states. *One-sided Cross* replaces one directional state and its version-matched qualification using preassigned reciprocal donors matched by dataset, backbone, role, and formation scale. The opposite state remains fixed,

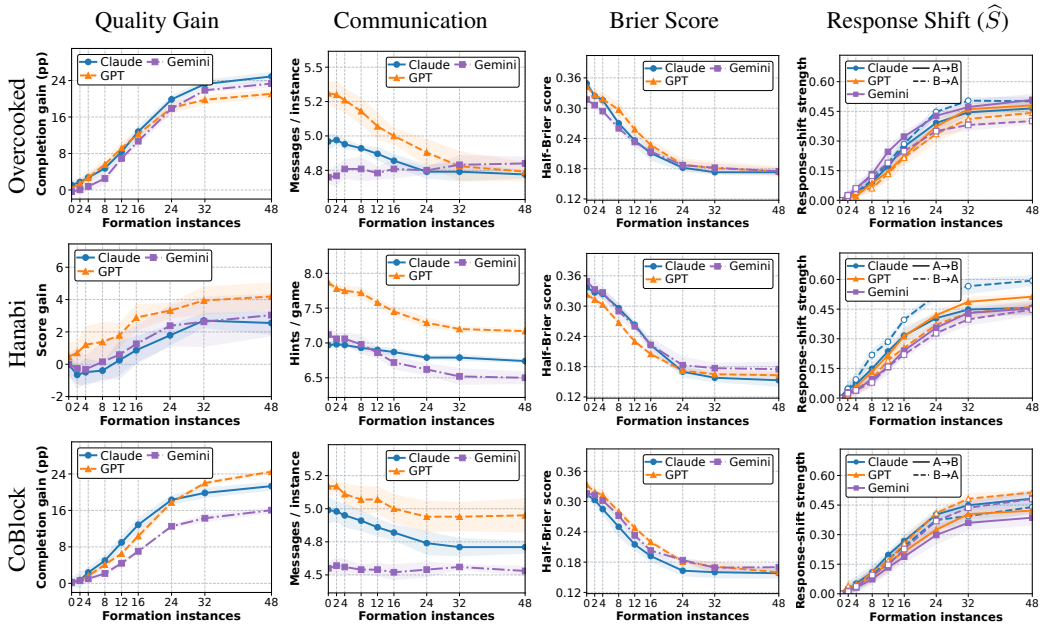


Figure 2: Formation over shared experience. Quality gain and communication use dataset-specific units: Hanabi reports score out of 25 and hints per game, whereas Overcooked and CoBlock report completion percentage points and messages per instance.

and directions are tested separately. *Generic Guidance* retains each agent's frozen real-state readout rule but replaces nonempty guidance with task-specification text using the same prompt slot and length budget. All runs, including inactive ones, are retained. Quality, actual communication, guidance eligibility, and injection frequency are reported.

Metrics. J denotes order completion in Overcooked, score (/25) in Hanabi, and structure completion in CoBlock. L denotes transmitted messages per instance, or hints per Hanabi game. With z_0 denoting B0's empty state, common-budget gain is $G(z, \beta_*) = J(z, \beta_*) - J(z_0, \beta_*)$, and the non-inferiority gap is $D_{\mathrm{NI}}(z) = J(z, \beta_*) - J(z_0, \beta_{\mathrm{full}})$ against full-budget B0. We test $H_0 : D_{\mathrm{NI}} \leq -\epsilon$, with communication substitution additionally requiring lower measured communication. Retrospective tasks-to-benefit is the earliest checkpoint from which estimated gain remains positive through $N = 48$. State-control effects $\Delta J = J(\mathrm{cond}, \beta_*) - J(\mathrm{Real}, \beta_*)$ are reported as mean $\pm$ SD of paired run-level differences. Prospective prediction uses usage half-Brier loss on observed future probes. Using full Coordination Responses on matched audit snapshots, the response shift $\widehat{S}^{(n)}_{r,d} = \sum_x \mu_d(x) D_{\mathrm{KL}}(\widehat{p}^{(n)}_{r,d}(\cdot \mid x) \| \widehat{p}_{0,d}(\cdot \mid x))$ measures departure from the history-neutral law. Suppressing shared indices and conditioning on x, the correspondence discrepancy $\widehat{\Delta}^{\mathrm{corr}}_{r,d} = \frac{1}{K} \sum_{\ell=1}^{K} \sum_x \mu_d(x) D_{\mathrm{KL}}(\widehat{p}^{\mathrm{real}} \| \widehat{p}^{\mathrm{neut},\ell})$ averages Real–Neutralized divergences over K histories. The cross-dyad discrepancy $\widehat{\Delta}^{\mathrm{pair}}_{r,d} = \sum_x \mu_d(x) D_{\mathrm{KL}}(\widehat{p}^{\mathrm{real}} \| \widehat{p}^{\mathrm{cross}})$ measures source-state sensitivity under a one-sided cross-dyad replacement. All three contrasts are repeated on the predeclared Hard unseen-instance bank.

Group-level identification. GTU requires all four components simultaneously, and we combine them by intersection–union, $p_{\mathrm{ID},d} = \max\{p_{\mathrm{Form},d}, p_{\mathrm{Corr},d}, p_{\mathrm{Pair},d}, p_{\mathrm{Gen},d}\}$, where $p_{\mathrm{Gen},d}$ aggregates the Hard-bank tests. Dataset-level bidirectional claims are Holm-corrected at $\alpha = 0.05$, with utility hypotheses forming a separate predeclared family. Run-level effects characterise heterogeneity rather than individual GTU decisions.

5.2 Results and Analysis

Sustaining quality with less communication. Table 2 shows that our method attains the highest mean $J(\beta_*)$ in all six Overcooked and CoBlock backbone settings. Relative to full-budget B0,

Table 3: Interaction specificity, held-out generalization, and directional attribution. Main and Hard audits report correspondence- and partner-specific response-law discrepancies, while directional interventions isolate the effect of enabling each learned state.

			Main audit		Hard audit			Dir. attribution	
	Method	Dir.	$\widehat{\Delta}^{\mathrm{corr}} \uparrow$	$\widehat{\Delta}^{\mathrm{pair}} \uparrow$	$\widehat{S}_{\mathrm{H}} \uparrow$	$\widehat{\Delta}^{\mathrm{corr}}_{\mathrm{H}} \uparrow$	$\widehat{\Delta}^{\mathrm{pair}}_{\mathrm{H}} \uparrow$	$\widehat{D}^{\mathrm{dir}}_{\mathrm{Cold}} \uparrow$	$\widehat{D}^{\mathrm{dir}}_{\mathrm{Exp}} \uparrow$
Overcooked	B4	$A \to B$	0.0985 $\pm$0.0040	0.0728 $\pm$0.0092	0.1183 $\pm$0.0139	0.0711 $\pm$0.0080	0.0568 $\pm$0.0063	0.0442 $\pm$0.0094	0.0607 $\pm$0.0181
		$B \to A$	0.0881 $\pm$0.0148	0.0736 $\pm$0.0109	0.1275 $\pm$0.0149	0.0558 $\pm$0.0098	0.0401 $\pm$0.0054	0.0450 $\pm$0.0121	0.0638 $\pm$0.0172
	Ours	$A \to B$	0.1966 $\pm$0.0088	0.1876 $\pm$0.0418	0.2619 $\pm$0.0112	0.1623 $\pm$0.0315	0.1541 $\pm$0.0234	0.1112 $\pm$0.0439	0.1525 $\pm$0.0117
		$B \to A$	0.2018 $\pm$0.0150	0.1724 $\pm$0.0101	0.2604 $\pm$0.0158	0.1431 $\pm$0.0265	0.1613 $\pm$0.0179	0.1093 $\pm$0.0333	0.1362 $\pm$0.0322
Hanabi	B4	$A \to B$	0.0741 $\pm$0.0067	0.0614 $\pm$0.0086	0.0977 $\pm$0.0126	0.0365 $\pm$0.0057	0.0407 $\pm$0.0113	0.0347 $\pm$0.0060	0.0538 $\pm$0.0104
		$B \to A$	0.0648 $\pm$0.0036	0.0623 $\pm$0.0107	0.0801 $\pm$0.0238	0.0337 $\pm$0.0057	0.0303 $\pm$0.0054	0.0367 $\pm$0.0088	0.0447 $\pm$0.0073
	Ours	$A \to B$	0.1886 $\pm$0.0100	0.1731 $\pm$0.0412	0.2164 $\pm$0.0128	0.1431 $\pm$0.0117	0.1273 $\pm$0.0062	0.0803 $\pm$0.0128	0.1138 $\pm$0.0201
		$B \to A$	0.1576 $\pm$0.0115	0.1691 $\pm$0.0151	0.1927 $\pm$0.0146	0.1089 $\pm$0.0151	0.1083 $\pm$0.0207	0.0767 $\pm$0.0276	0.0984 $\pm$0.0079
CoBlock	B4	$A \to B$	0.1042 $\pm$0.0062	0.0940 $\pm$0.0141	0.1575 $\pm$0.0197	0.0622 $\pm$0.0084	0.0679 $\pm$0.0197	0.0490 $\pm$0.0065	0.0580 $\pm$0.0052
		$B \to A$	0.0963 $\pm$0.0058	0.0798 $\pm$0.0176	0.0991 $\pm$0.0082	0.0517 $\pm$0.0091	0.0519 $\pm$0.0069	0.0477 $\pm$0.0110	0.0710 $\pm$0.0082
	Ours	$A \to B$	0.2294 $\pm$0.0188	0.1781 $\pm$0.0204	0.2806 $\pm$0.0258	0.1675 $\pm$0.0103	0.1359 $\pm$0.0161	0.0995 $\pm$0.0293	0.1558 $\pm$0.0320
		$B \to A$	0.2089 $\pm$0.0083	0.1982 $\pm$0.0234	0.2591 $\pm$0.0212	0.1689 $\pm$0.0137	0.1449 $\pm$0.0246	0.0997 $\pm$0.0391	0.1360 $\pm$0.0231

Table 4: Ablations on GPT-5.5. Behavioral identification effects are reported separately.

	Overcooked			Hanabi			CoBlock		
Variant	$J(\beta_*) \uparrow$	$L \downarrow$	**Brier**$\downarrow$	$J(\beta_*) \uparrow$	$L \downarrow$	**Brier**$\downarrow$	$J(\beta_*) \uparrow$	$L \downarrow$	**Brier**$\downarrow$
Full	**54.38** $\pm$3.48	**2.32** $\pm$0.14	**0.109** $\pm$0.022	**14.45** $\pm$1.23	**7.11** $\pm$0.13	0.144 $\pm$0.018	**54.03** $\pm$3.52	2.25 $\pm$0.22	**0.116** $\pm$0.021
No qualification	51.99 $\pm$2.93	2.41 $\pm$0.15	0.145 $\pm$0.024	12.73 $\pm$1.02	7.39 $\pm$0.33	0.200 $\pm$0.038	51.14 $\pm$2.86	2.22 $\pm$0.16	0.194 $\pm$0.021
No monitoring	50.60 $\pm$5.18	2.38 $\pm$0.16	0.151 $\pm$0.024	13.80 $\pm$1.18	7.16 $\pm$0.23	0.165 $\pm$0.025	53.18 $\pm$5.09	**2.19** $\pm$0.18	0.168 $\pm$0.023
No usage	51.70 $\pm$4.16	2.47 $\pm$0.10	0.189 $\pm$0.021	12.58 $\pm$1.11	7.74 $\pm$0.34	0.264 $\pm$0.058	47.51 $\pm$1.93	2.28 $\pm$0.10	0.203 $\pm$0.067
Argmax	51.30 $\pm$3.60	2.42 $\pm$0.13	0.117 $\pm$0.017	13.44 $\pm$0.57	7.28 $\pm$0.18	**0.139** $\pm$0.006	52.59 $\pm$4.25	2.20 $\pm$0.12	0.135 $\pm$0.020
Frozen predictor	43.61 $\pm$6.97	2.49 $\pm$0.09	0.255 $\pm$0.058	10.62 $\pm$1.14	7.82 $\pm$0.39	0.265 $\pm$0.040	41.99 $\pm$2.96	2.30 $\pm$0.09	0.245 $\pm$0.047
No rehearsal	50.46 $\pm$4.12	2.43 $\pm$0.12	0.161 $\pm$0.012	13.00 $\pm$0.92	7.54 $\pm$0.15	0.169 $\pm$0.023	46.56 $\pm$5.63	2.24 $\pm$0.14	0.167 $\pm$0.028
Shared learner	52.72 $\pm$5.59	2.46 $\pm$0.10	0.182 $\pm$0.018	12.52 $\pm$1.86	7.51 $\pm$0.61	0.201 $\pm$0.011	44.14 $\pm$5.20	2.26 $\pm$0.12	0.203 $\pm$0.018
Generic guidance	38.58 $\pm$5.66	2.50 $\pm$0.08	0.121 $\pm$0.032	9.75 $\pm$1.67	7.98 $\pm$0.42	0.162 $\pm$0.015	38.53 $\pm$5.61	2.29 $\pm$0.09	0.128 $\pm$0.009
Count	48.38 $\pm$6.12	2.48 $\pm$0.09	0.180 $\pm$0.025	11.97 $\pm$1.25	7.83 $\pm$0.66	0.221 $\pm$0.025	51.68 $\pm$4.32	2.27 $\pm$0.11	0.214 $\pm$0.043
Retrieval	50.76 $\pm$4.14	2.45 $\pm$0.11	—	11.82 $\pm$0.64	7.83 $\pm$0.36	—	47.40 $\pm$2.52	2.28 $\pm$0.10	—

constrained-budget quality improves by 15.56–24.41 percentage points while using 51.16–56.63% fewer messages, establishing NI under the prespecified 5-point margin in all six settings. Our method's own standard-to-constrained quality drop is only 1.77–4.12 points, compared with 12.84–20.62 for B0. In Hanabi, our method achieves the highest mean score on all three backbones and uses fewer hints than B0, although B6 and B5 use slightly fewer hints on GPT-5.5 and Gemini 3.6 Flash, respectively. These results indicate greater robustness to the tested communication restriction rather than superiority under arbitrary budgets.

Figure 2 shows quality gains emerging with shared experience while standard-rule communication decreases. In Overcooked and CoBlock, One-sided Cross and Generic Guidance both yield lower mean quality than Real at β_*, suggesting that both state identity and learned guidance content matter for utility under the tested budget.

Evidence for dyadic tacit understanding. Figure 2 tracks prospective prediction separately from behavioral formation. Successive candidate versions improve future-probe half-Brier scores relative to the context-key count predictor, while directional response shifts emerge in both directions. Because candidate forecasts are evaluated before deployment, this ordering is consistent with the proposed learning-and-use mechanism, but does not itself establish GTU.

Table 3 provides the controlled identification evidence. Across the six dataset-direction cells, our method yields $\widehat{\Delta}^{\mathrm{corr}} = 0.158$–$0.229$ and $\widehat{\Delta}^{\mathrm{pair}} = 0.169$–$0.198$ on the main audit bank. On the Hard bank, $\widehat{S}_{\mathrm{H}} = 0.193$–$0.281$, $\widehat{\Delta}^{\mathrm{corr}}_{\mathrm{H}} = 0.109$–$0.169$, and $\widehat{\Delta}^{\mathrm{pair}}_{\mathrm{H}} = 0.108$–$0.161$. These results indicate that the learned behavioral relation remains sensitive to both real context–response correspondence and directional-state source on unseen same-family instances. Directional state interventions further produce direct response-law effects of 0.077–0.156 with the opposite state held Cold or Experienced, providing method-level attribution separately from the GTU criteria.

Table 5: Parameter sensitivity on Overcooked and CoBlock. Entries report mean±SD over eight independent runs. Gray cells indicate the default setting, and blue cells indicate the best value in each metric column within a parameter block.

	Overcooked			CoBlock		
N	$J\uparrow$	$\widehat{S}\uparrow$	Brier↓	$J\uparrow$	$\widehat{S}\uparrow$	Brier↓
8	43.51±2.46	0.031±0.009	0.245±0.018	44.80±3.83	0.019±0.012	0.252±0.018
16	47.79±3.20	0.067±0.010	0.188±0.014	47.77±1.82	0.070±0.007	0.195±0.012
24	49.91±2.75	0.114±0.010	0.140±0.010	50.19±3.09	0.100±0.018	0.147±0.008
32	53.35±2.48	0.166±0.012	0.118±0.014	53.65±2.94	0.154±0.008	0.125±0.008
48	**54.38**±3.48	**0.212**±0.008	**0.109**±0.009	**54.03**±3.52	**0.194**±0.011	**0.116**±0.006
β	$G\uparrow$	$D_{\mathrm{NI}}\uparrow$	$L\downarrow$	$G\uparrow$	$D_{\mathrm{NI}}\uparrow$	$L\downarrow$
1.0	0.109±0.031	**0.103**±0.021	5.36±0.13	0.081±0.039	**0.074**±0.022	5.07±0.24
0.7	0.158±0.034	0.100±0.016	3.50±0.19	0.181±0.031	0.058±0.021	3.37±0.13
0.5	**0.209**±0.027	**0.080**±0.029	**2.32**±0.13	**0.244**±0.040	**0.038**±0.033	**2.25**±0.16
0.3	**0.235**±0.034	0.005±0.023	**1.45**±0.22	**0.269**±0.028	−0.009±0.031	**1.40**±0.25
L_{val}	Brier↓	Cov.↑	$J\uparrow$	Brier↓	Cov.↑	$J\uparrow$
4	0.296±0.010	**0.892**±0.011	51.88±2.86	0.311±0.014	**0.890**±0.008	51.99±2.97
8	0.255±0.009	0.782±0.018	53.23±1.82	0.258±0.008	0.782±0.019	53.57±1.26
16	**0.187**±0.010	**0.716**±0.013	**53.55**±2.94	**0.204**±0.009	**0.714**±0.012	**54.85**±1.78
32	**0.169**±0.008	0.565±0.014	**53.72**±2.23	**0.167**±0.010	0.564±0.012	52.37±2.46

Ablation Study All ablations use GPT-5.5 with eight independent runs per dataset and matched task sequences. Formal GTU inference is restricted to the full method.

Table 4 shows that the full method achieves the best quality and communication results across all datasets, together with the best or near-best Brier results (Argmax is best on Hanabi). Argmax guidance remains competitive, indicating a modest benefit from retaining the full response distribution. Removing qualification, monitoring, rehearsal, usage semantics, or predictor learning consistently degrades performance. Among the component-removal variants, freezing the predictor causes the largest quality degradation, while removing rehearsal or usage semantics also leads to substantial losses. Notably, Generic Guidance preserves near-full Brier accuracy but sharply reduces quality and increases communication, separating relational prediction from its downstream use. Shared-learner and simpler representation controls are also consistently weaker. These ablations are descriptive rather than additional GTU claims.

Parameter Sensitivity Analysis We vary N, β, and L_{val} independently on Overcooked and CoBlock using GPT-5.5 as the backbone; because only one backbone is used, absolute metric levels are not directly comparable to the main three-backbone results. The N sweep uses deployed checkpoints from the same formation runs; at $N = 8$, no candidate has yet completed the $L_{\mathrm{val}} = 16$ qualification window, so guidance injection is empty and the positive $\widehat{S}$ reflects finite-sample KL bias rather than deployed guidance effects. Each L_{val} setting restarts formation on matched task sequences, while the β sweep reuses frozen deployed states. All evaluation updates remain disabled. Table 5 summarizes the eight-run results, examining experience dependence, communication sensitivity, and the prediction–deployment tradeoff.

6 CONCLUSION

This work formalizes Dyadic Tacit Understanding (DTU) as a directional, history-induced change in coordination behavior and defines Genuine Dyadic Tacit Understanding (GTU) through the joint requirements of Formation, Interaction Specificity, and Cross-Instance Generalization. Building on this theory, we introduce TACT, a decentralized framework that learns partner-conditioned coordination-response distributions, prospectively qualifies their deployment, and converts them into private distributional guidance without modifying the base LLMs. Experiments on Collab-Overcooked, Hanabi, and CoBlock across three LLM backbones show that persistent dyads can develop reusable partner-specific coordination regularities, with bidirectional GTU supported across all three evaluated benchmark families while maintaining comparable task quality under constrained communication. The current study is limited to fixed dyads and pre-specified coordination families, motivating future work on changing partners, heterogeneous roles, and larger multi-agent settings.

AI USE STATEMENT

We used generative AI tools solely for language polishing and proofreading of the manuscript. We did not use generative AI for experiments and data analysis. The authors take full responsibility for the final content, including all text, analyses, and conclusions.

REPRODUCIBILITY STATEMENT

All benchmarks used in our experiments are publicly available. We provide implementation details, hyperparameters, source code, configurations, and reproduction scripts in an anonymous repository to support independent replication.

REFERENCES

Ido Aharon, Emanuele La Malfa, Michael Wooldridge, and Sarit Kraus. Tacit coordination of large language models. *arXiv preprint arXiv:2601.22184*, 2026.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368, 2025.

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

Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482, 1996.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel Yamins, and Nick Haber. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 6507–6546, 2025.

Jianxin Gao, Tianyi Yu, Linna Deng, Runze Li, and Zining Wang. Testing interchangeability in llm agent teams. *arXiv preprint arXiv:2609.05279*, 2026.

Robert D Hawkins, Michael C Frank, and Noah D Goodman. Characterizing the dynamics of learning in repeated reference games. *Cognitive science*, 44(6):e12845, 2020.

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

Florian Kandra, Vera Demberg, and Alexander Koller. Llms syntactically adapt their language use to their conversational partner. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 873–886, 2025.

Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. *arXiv preprint arXiv:1903.05168*, 2019.

Charles Metzing and Susan E Brennan. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213, 2003.

Chunjiang Mu, Ya Zeng, Qiaosheng Zhang, Kun Shao, Chen Chu, Hao Guo, Danyang Jia, Zhen Wang, and Shuyue Hu. Adaptive theory of mind for llm-based multi-agent coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29608–29616, 2026.

Hadi Nekoei, Xutong Zhao, Janarthanan Rajendran, Miao Liu, and Sarath Chandar. Towards few-shot coordination: Revisiting ad-hoc teamplay challenge in the game of hanabi. In *Conference on Lifelong Learning Agents*, pp. 861–877. PMLR, 2023.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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

Chen Qian, Yufan Dang, Jiahao Li, Wei Liu, Zihao Xie, Yifei Wang, Weize Chen, Cheng Yang, Xin Cong, Xiaoyin Che, et al. Experiential co-learning of software-developing agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5628–5640, 2024.

Ramón Rico, Miriam Sánchez-Manzanares, Francisco Gil, and Cristina Gibson. Team implicit coordination processes: A team knowledge–based approach. *Academy of management review*, 33 (1):163–184, 2008.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6279–6296, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

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

Saujas Vaduguru, Yilun Hua, Yoav Artzi, and Daniel Fried. Success and cost elicit convention formation for efficient communication. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42033–42050, 2026.

Po-Ya Angela Wang, Chinmaya Mishra, Aslı Özyürek, Paula Rubio-Fernández, and Esam Ghaleb. Aligned but not partner-specific: Distinguishing how multimodal llm agents succeed in reference games without human-like conventions. *arXiv preprint arXiv:2606.08081*, 2026.

Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, and Min Zhang. Agentdropout: Dynamic agent elimination for token-efficient and high-performance llm-based multi-agent collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24013–24035, 2025.

Guande Wu, Chen Zhao, Claudio Silva, and He He. Your co-workers matter: Evaluating collaborative capabilities of language models in blocks world. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4941–4957, 2024.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. 2024a.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. In *International Conference on Learning Representations*, volume 2025, pp. 75389–75428, 2025.

Guibin Zhang, Muxin Fu, Kun Wang, Frank Wan, Miao Yu, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *Advances in Neural Information Processing Systems*, 38:12988–13018, 2026.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. In *International Conference on Learning Representations*, volume 2024, pp. 19373–19401, 2024b.

A Theoretical Analysis and Proofs

The results below establish the main theoretical properties of the proposed formulation and derive explicit consequences of the response and usage definitions. They concern the specified probability laws and evaluation protocols. They do not guarantee that a finite online learner acquires GTU, that empirical qualification controls deployment risk, or that a detected relation improves task utility.

A.1 Response-Law Properties and the DSIR Identity

Fix a direction $d = i \to j$, a realized history h_n, and audit distribution $\mu_d = \mu_{\mathcal{F}}^{i\to j}$. Write $p_h(c \mid x) = p_{i\to j}^{\star(n)}(c \mid x, h_n)$ and $p_0(c \mid x) = p_0^\star(c \mid x)$. The response extractor is measurable and its valid response alphabet is finite at each audited context. Under history neutrality, $P(C = c \mid X = x, H = h, \eta = 0) = p_0(c \mid x)$ for every admissible history. Response-support overlap requires $p_h(\cdot \mid x) \ll p_0(\cdot \mid x)$ for μ_d-almost every x.

The local log ratio is defined on the reference support. If $p_h(c \mid x) = 0$ and $p_0(c \mid x) > 0$, use $\rho(c; x, h_n) = -\infty$ and the convention $0 \log(0/a) = 0$ in expectations. Overlap rules out a positive numerator with a zero denominator but does not supply a uniform positive lower bound on p_0. Hence S is well defined in $[0, +\infty]$; statements requiring a finite value additionally assume integrability of the local KL over μ_d.

Proposition 1 (DSIR normalization and shift characterization)**.** *Under these conditions, for almost every audited context,*

$$\begin{cases} p_h(c \mid x) = p_0(c \mid x) e^{\rho(c;x,h_n)}, \\ \displaystyle\sum_{c: p_0(c|x)>0} p_0(c \mid x) e^{\rho(c;x,h_n)} = 1, \\ \mathbb{E}_{C\sim p_h(\cdot|x)}[\rho(C; x, h_n)] = D_{\mathrm{KL}}(p_h(\cdot \mid x) \| p_0(\cdot \mid x)). \end{cases} \tag{11}$$

Moreover $S_d^{(n)} \geq 0$, with equality if and only if $p_h(\cdot \mid x) = p_0(\cdot \mid x)$ for μ_d-almost every x. If the cold-start behavioral law equals the history-neutral law, then $\rho_d^{\star(0)} = 0$ on their common support and $S_d^{(0)} = 0$.

Proof. **Step 1: Recover the normalized response law.** Exponentiating the defining log ratio gives the first identity wherever $p_h > 0$. The zero-mass convention extends it to the rest of the reference support. Summing then gives one because p_h is a probability law and overlap places all its mass within that support.

Step 2: Average the response-specific residual. Multiplying the log ratio by $p_h(c \mid x)$ and summing over c gives the third identity. Averaging over the fixed audit distribution yields

$$S_d^{(n)} = \mathbb{E}_{X\sim\mu_d} \mathbb{E}_{C\sim p_h(\cdot|X)}[\rho(C; X, h_n)]. \tag{12}$$

The equality is interpreted through the KL integral when the result is extended-valued; finite integrability gives the ordinary finite expectation.

Step 3: Establish nonnegativity and its equality case. For each context, the log-sum inequality gives $D_{\mathrm{KL}}(p_h \| p_0) \geq 0$, with equality exactly when the two normalized laws coincide. The integral of this nonnegative function is zero exactly when it is zero almost everywhere under μ_d. This proves both statements about S. Substituting the assumed cold-start equality gives a zero log ratio on the common support and hence a zero aggregate. □

Empty history alone is not a substitute for cold-start behavioral equivalence: changing the role prompt, inference settings, or other current inputs can change a response law even when neither condition has cross-instance memory. Likewise, a finite smoothed estimate is not evidence that population support overlap has been verified.

A.2 Why the Identification Conditions Are Distinct

For the same direction, history, and declared support, write

$$\begin{cases} \Delta_d^{\mathrm{corr}}(h_n) = \mathbb{E}_{X\sim\mu_d}\mathbb{E}_{\widetilde{h}\sim\mathcal{N}(h_n)} D_{\mathrm{KL}}\big(p_h(\cdot \mid X)\|p_{\widetilde{h}}(\cdot \mid X)\big)\,, \\ \Delta_d^{\mathrm{pair}}(h_n) = \mathbb{E}_{X\sim\mu_d} D_{\mathrm{KL}}(p_h(\cdot \mid X)\|p_{\mathrm{cross}}(\cdot \mid X))\,. \end{cases} \tag{13}$$

Support conditions for these contrasts concern their own reference laws and do not follow automatically from overlap with the Cold law. Correspondence specificity and partner specificity require the respective behavioral distinctions under the declared controls. Generalization repeats Formation and both specificity contrasts on the fixed unseen-instance support; in the empirical protocol this is the predeclared Hard bank.

The following finite constructions illustrate why the identification conditions cannot be collapsed. They are mathematical examples, not additional benchmark results. Take $C \in \{0,1\}$ and a Cold law with $P(C=1) = 1/2$.

A positive shift without specificity. Suppose every history of the same length induces $P(C = 1) = 4/5$, including its neutralizations and matched donor histories. Then

$$S = \frac{4}{5}\log\frac{8}{5} + \frac{1}{5}\log\frac{2}{5} > 0, \qquad \Delta^{\mathrm{corr}} = \Delta^{\mathrm{pair}} = 0. \tag{14}$$

Generic exposure can therefore produce a response shift without the required relational distinctions.

Correspondence and partner specificity need not coincide. With two equally weighted contexts, let Real predict probability $4/5$ at one context and $1/5$ at the other. A neutralization that swaps these associations produces positive correspondence discrepancy. If all matched dyads learn the same associations, however, donor replacement has zero effect. Conversely, let different dyads acquire distinct context-independent response frequencies. Permuting correspondence preserves each frequency and has no effect, although replacement by a dyad with a different frequency changes the response law. Thus neither Corr nor Pair implies the other.

These constructions motivate the joint requirement

$$\mathrm{GTU}_d^{(n)} = \mathrm{Form}_d^{(n)} \wedge \big(\mathrm{Corr}_d^{(n)} \wedge \mathrm{Pair}_d^{(n)}\big) \wedge \mathrm{Gen}_d^{(n)}. \tag{15}$$

The claim remains conditional on the declared coordination family, observation interface, nuisance matching, and intervention scope.

A.3 Relation to Conditional Mutual Information

The reference in $S_d^{(n)}$ is history-neutral rather than the history-marginalized law used in conditional mutual information. Introduce an auxiliary history distribution $\nu(dh \mid x)$ (equal to $\nu(dh)$ only when histories and contexts are independent), the joint law $\mu_d(dx)\nu(dh \mid x)p_h(dc \mid x)$, and the mixture $\overline{p}(c \mid x) = \int p_h(c \mid x)\nu(dh \mid x)$.

Proposition 2 (Reference-anchored KL decomposition)**.** *Assume the relevant KL integrals are finite and, for μ_d-almost every x, $p_h(\cdot \mid x) \ll p_0(\cdot \mid x)$ for $\nu(\cdot \mid x)$-almost every h. Then*

$$\mathbb{E}_{X\sim\mu_d,\, H\sim\nu(\cdot|X)} D_{\mathrm{KL}}(p_H(\cdot \mid X)\|p_0(\cdot \mid X)) = I(C;H \mid X) + \mathbb{E}_{X\sim\mu_d} D_{\mathrm{KL}}(\overline{p}(\cdot \mid X)\|p_0(\cdot \mid X))\,. \tag{16}$$

The conditional mutual information is evaluated under the auxiliary joint law.

Proof sketch. Insert $\overline{p}(\cdot \mid x)$ into the log ratio and take expectation under the auxiliary joint law; the two resulting terms are respectively $I(C;H \mid X)$ and the reference-mismatch KL.

A.4 Predictive Loss, Usage Aggregation, and Qualification

The training objective and usage projection concern the response predictor; neither by itself identifies the behavioral response law induced after the predictor is deployed as guidance.

Proposition 3 (Conditional log loss and usage projection)**.** *Fix the context representation and valid response sets. Let $p_{\mathrm{Bal}}(C \mid \kappa)$ denote the conditional response law under the instance–key–record sampling distribution used for training. For any predictor q with the required support,*

$$\mathbb{E}_{\mathrm{Bal}}[-\log q(C \mid \kappa)] = H_{\mathrm{Bal}}(C \mid \kappa) + \mathbb{E}_{\kappa\sim\mathrm{Bal}} D_{\mathrm{KL}}(p_{\mathrm{Bal}}(\cdot \mid \kappa) \| q(\cdot \mid \kappa)) \,. \tag{17}$$

Moreover, for a fixed deterministic usage map and any complete-response laws p, q, their pushforwards p^U, q^U are normalized and satisfy

$$D_{\mathrm{KL}}\big(p^U(\cdot \mid \kappa) \| q^U(\cdot \mid \kappa)\big) \le D_{\mathrm{KL}}(p(\cdot \mid \kappa) \| q(\cdot \mid \kappa)) \,. \tag{18}$$

Proposition 4 (Half-Brier interpretation)**.** *For a usage outcome Y with law p on a finite alphabet and a normalized forecast z,*

$$\mathbb{E}[\ell_B(z, Y)] = \frac{1}{2}(1 - \|p\|_2^2) + \frac{1}{2}\|z - p\|_2^2, \qquad 0 \le \ell_B(z, Y) \le 1. \tag{19}$$

If the valid usage set contains a labels, the uniform forecast has loss $\frac{1}{2}(1 - 1/a)$ for every valid realized label.

Proof sketch. Expanding $\ell_B(z, Y) = \frac{1}{2}(\sum_u z_u^2 - 2z_Y + 1)$ and completing the square gives the identity; the simplex diameter gives the upper bound, and $z_u = 1/a$ gives the uniform-reference loss.

The qualification reference is uniform over each probe's valid usage set, distinct from $p_0^\star$, with half-Brier loss determined by the set size. Qualification allows a reference-loss tolerance $\epsilon_{\mathrm{ref}} \ge 0$ on observed probes, while deployment may alter the behavioral response law. The gate is therefore an empirical eligibility rule rather than a finite-sample risk certificate or a GTU test.

A.5 Validity and Scope of Joint Behavioral Inference

The confirmatory procedure uses conditional response-law equality tests. Their validity requires a fixed audit bank, independently reset response draws, and exchangeability under the particular null being tested. Missingness, shared samples, and donor relationships are part of that sampling design; they cannot be removed by treating every logged row as independent.

Proposition 5 (Intersection–union and outer multiplicity control)**.** *Suppose each component p-value is super-uniform under its null, conditional on the fixed evaluated state panel and audit design. For a conjunction of necessary alternatives with nulls $H_{01}, \dots, H_{0q}$, let $p_\cap = \max_{a \le q} p_a$. Then*

$$P\left(p_\cap \le \alpha \mid \bigcup_{a=1}^{q} H_{0a}\right) \le \alpha. \tag{20}$$

This applies to the Hard-bank conjunction, directional identification, and bidirectional dataset-level identification. Applying Holm's procedure to the three resulting dataset-level p-values controls their familywise error at α, without requiring independence between those tests.

Proof. **Step 1: Control one necessary false component.** Under the conjunction's null, at least one component null, say H_{0a}, is true. The event $\{\max_b p_b \le \alpha\}$ is a subset of $\{p_a \le \alpha\}$, whose probability is at most α. Repeating the same reasoning through nested conjunctions proves the first assertion.

Step 2: Control the outer family. Let m_0 of the three dataset hypotheses be true. If Holm rejects any true hypothesis, the first rejected true hypothesis must have a p-value at most α/m_0: at that step at least the m_0 true hypotheses remain. By the union bound, the probability that any true p-value is at most α/m_0 is no more than $m_0(\alpha/m_0) = \alpha$. □

B Algorithmic Details and Implementation

Online formation produces directional states through the main-text train–candidate–active lifecycle, while read-only behavioral audits measure the response laws induced by those states.

Table 6: Version roles and permitted operations during a formation instance.

Copy	Within-instance role	Boundary operation	Evidence source
Train	Not used as an unqualified controller	Updated after instance scoring	Current complete records and bounded rehearsal
Candidate	Read-only shadow forecasts	Qualified only at its fixed block boundary	Later-instance preselected probes
Active	Read-only private guidance	Replaced as a whole or has cells disabled for the next instance	Version-matched qualification and deployed forecasts

B.1 Response Learning and Versioned Deployment

Local checkpoint and deployment package. Active weights, model version, response and usage semantics, qualified and disabled cells, and monitoring records form an atomic deployment package. Frozen evaluation keeps this package read-only and disables learning, promotion, monitoring, and expiry.

Version lifecycle and prospective qualification. Each directional learner maintains a train model, a frozen candidate, and at most one active model (Table 6). The active version and enabled cells remain fixed within each instance while the candidate produces shadow forecasts that never enter prompts.

At each selected probe, context construction, probe selection, and forecast storage precede the response, which is recorded as observed or missing. Candidate and active forecasts are scored before train updates. At every scheduled boundary, any completed candidate is qualified and a fresh candidate is copied from the updated train model, regardless of promotion. Deployment changes take effect only in the next instance. Responses may subsequently train later versions without altering their stored forecasts.

Input and response construction. A family-specific adapter supplies legal predecision information. The context extractor maps role, coordination stage, dependency position, observable demand, and local features to κ. Role-relative fields support reuse across changes in surface objects and positions. Fixed mappings $K_{\mathcal{F}}$ and V_i produce the training key and verification cell without prescribing partner responses.

Each opportunity has a stable identity linking its context, probe-selection status, forecast version, response window, and observation status. Labels use executed, observable behavior and distinguish response onset from completion where observable. A truncated window is not automatically labeled as an absent response. The extraction policy distinguishes observable absence from missing labels. Complete responses from failed tasks remain eligible.

Conditional prediction. Categorical context and response embeddings, together with applicable normalized numeric features, are encoded into a common space. Unit-normalized towers are scored by their inner product, with softmax over all valid responses:

$$\log q_{\theta_i}(C \mid \kappa) = \frac{s_{\theta_i}(\kappa, C)}{\tau} - \operatorname{logsumexp}_{C' \in A_{\mathcal{F}}(\kappa)} \frac{s_{\theta_i}(\kappa, C')}{\tau}. \tag{21}$$

Valid support is determined by predefined semantics and current legal information.

Training uses complete current-instance records and bounded rehearsal. Let $\mathcal{I}$ index represented valid instances, Ω_a the keys in instance a, and $\mathcal{D}_{a,\omega}$ its records for key ω. Then

$$P_{\mathrm{Bal}}(e) = \frac{1}{|\mathcal{I}|} \frac{1}{|\Omega_a|} \frac{1}{|\mathcal{D}_{a,\omega}|}, \qquad e \in \mathcal{D}_{a,\omega}. \tag{22}$$

Bounded optimizer steps minimize complete-response negative log likelihood. Instance identities serve only sampling and support accounting, and bounded rehearsal need not reproduce the complete historical distribution. Train parameters remain unchanged when no complete current or rehearsal records are available.

Prospective qualification. For each candidate and cell, $a_{v,\gamma}$ counts selected probes, $m_{v,\gamma}$ fully observed probes, and $n^{\mathrm{train}}_{v,\gamma}$ distinct training instances encountered by that version. Training support records exposure rather than surviving buffer contents. Selected missing probes remain in the coverage denominator and are not replaced. At the block boundary, compute

$$\begin{cases} \widehat{\mathrm{cov}}_{v,\gamma} = m_{v,\gamma}/a_{v,\gamma}, \\ \widehat{L}_{v,\gamma} = m^{-1}_{v,\gamma} \sum\limits_{e \in \mathcal{P}^{\mathrm{obs}}_{v,\gamma}} \ell_B\left(q^U_{\theta_{i,v}}(\cdot \mid \kappa_e), \mathcal{U}_i(\kappa_e, C_e)\right), \\ \widehat{L}^{\mathrm{unif}}_{v,\gamma} = m^{-1}_{v,\gamma} \sum\limits_{e \in \mathcal{P}^{\mathrm{obs}}_{v,\gamma}} \tfrac{1}{2}\left(1 - |A^U_i(\kappa_e)|^{-1}\right). \end{cases} \tag{23}$$

Zero denominators make a cell ineligible. Otherwise, qualification applies the support, coverage, absolute-loss, and uniform-reference tests of main-text Eq. (913), with matching model and semantic versions.

Algorithm 1: Independent directional formation and prospective deployment

Input: Frozen LLMs; distinct same-family tasks $T_1, \ldots, T_N$; fixed extractors, usage maps, gates, version metadata, W, and L_{val}.

foreach $i \in \{A, B\}$ **do**
  Initialize a separate train model, optimizer, support ledger, and bounded buffer
  Set candidate and active to absent; set enabled cells to $\varnothing$
for $k = 1, \ldots, N$ **do**
  Stage I: act with fixed deployed states and log shadow forecasts
  Clear raw cross-instance dialogue and copy each current active package for T_k
  Initialize instance records, pending windows, and per-cell probe-selection flags
  while T_k *has not terminated* **do**
    foreach *agent decision opportunity* (i, t) **do**
      Build $\kappa_{i,t}, \omega_{i,t}, \gamma_{i,t}$ from legal predecision information
      if *a candidate exists and this is the first eligible probe in its cell* **then**
        Store candidate version, valid support, and $q^U_{\theta_{i,v}}$ before the response
        Mark the cell selected for this instance, even if its later response is missing
      if *an active model exists and this is its first eligible enabled-cell probe* **then**
        Store its forecast for version-matched monitoring before the response
      Read $g^{\mathrm{tac}}_{i,t}$ from the fixed active package, or return empty guidance
      The frozen LLM chooses its own action and permitted message
    Execute the ordinary environment interaction; append only legal response observations
    Close completed response windows and retain complete or missing outcome status
  Close remaining windows according to the fixed observation policy
  Stage II: evaluate before learning
  foreach $i \in \{A, B\}$ **do**
    Score selected candidate probes and active monitoring probes from stored forecasts
    Stage monitoring failures or expiry as cell disabling for the next instance
    Form the complete current-plus-rehearsal set $\mathcal{D}_i$
    if $\mathcal{D}_i \neq \varnothing$ **then**
      Sample by Eq. (22) and update train with complete-response NLL
      Update the distinct-instance training-support ledger
    Insert eligible structured records into the bounded rehearsal buffer
    Stage III: change deployment only at an instance boundary
    Apply staged disabling only to the monitored incumbent version
    if $k \geq W$ *and* $(k - W) \bmod L_{\mathrm{val}} = 0$ **then**
      if *a completed candidate exists and has at least one qualified cell* **then**
        Replace active by this entire candidate and its own qualification package
      Copy current train into a new frozen candidate with fresh verification records
    Commit the next-instance active package and enabled-cell set
  Save scheduled checkpoints with the actually deployed states and complete local training state

Guidance and monitoring. For an enabled cell, full-response probabilities are aggregated into q^U and projected through Π_i^{dist}. Guidance conveys participation alternatives and uncertainty. The frozen LLM selects actions and messages under the common communication rule.

Monitoring separately scores the active version's stored forecasts over fixed recent-instance windows, using prespecified minimum support, coverage, loss limits, and expiry. Failure or expiry disables a cell from the next instance for the remainder of that version. Replacement requires a newly qualified candidate. Audit and utility evaluation never update these monitoring windows.

Algorithm 2: Matched read-only response-law audit

Input: Checkpoint panel; fixed Main and Hard banks; intervention masks and reciprocal donors; neutralization constraints; draws M; smoothing and randomization settings.

Stage I: construct frozen conditions
foreach *source run r and evaluated checkpoint n* **do**
  Load Real's actually deployed packages and the history-neutral reference
  for $\ell = 1, \ldots, K$ **do**
    Transform the declared histories without crossing training/verification boundaries
    Rebuild train, candidate, qualification, and active lifecycle from initialization
    Retain the reconstructed deployment package and neutralization diagnostics
  Construct one-sided Cross using preassigned donor packages with matching semantics
  Construct directional enable/disable conditions with the opposite state held fixed
Stage II: collect independent reset-and-sample continuations
foreach *bank, direction, snapshot x, source run, and condition g* **do**
  for $a = 1, \ldots, M$ **do**
    Restore snapshot and current-task interface; verify observation and legal-action hashes
    Load the condition's read-only states; disable all updates, promotion, and expiry
    Execute a bounded continuation with the fixed communication and inference settings
    Record observed full C or missing status, sample identity, versions, and API costs
    Verify unchanged persistent-state hashes and discard episode-local state
Stage III: estimate, compare, and combine
Estimate each response law on its declared valid full-response alphabet
Mark zero-observation cells unestimable; retain coverage and planned-draw denominators
Compute KL effect sizes and symmetric comparison statistics with shared-reference identities
Randomize only under the applicable within-snapshot null-exchangeability scheme
Combine Hard, directional, and bidirectional tests by intersection–union
Apply Holm correction across the three dataset-level bidirectional identification claims; report utility in a separate family
return *Effect sizes, coverage, supported test decisions, and traceable audit records.*

B.2 Frozen Behavioral Audits and Controlled Reconstruction

An audit snapshot stores the simulator state, local observations, current-task context, communication state, and legal action information separately from the learned state. Restoration checks compare legal observations, prompt base, and action sets. Simulator restoration alone is insufficient if an unrecovered inference session or hidden memory can affect responses.

Each planned draw restores the snapshot, loads a read-only deployment package, and executes a bounded continuation under the common audit communication rule, recording complete C or missing status. Fresh API samples are drawn independently, and any sample reuse is explicitly recorded. Shared tasks or simulator seeds do not imply common-random-number coupling of API outputs. Persistent-state hashes are checked before and after each continuation.

Controlled reconstruction of history interventions. The history transformation, directional mask, temporal partitions, and validity constraints are fixed before reconstruction. One-sided and two-sided transformations remain distinct interventions. Each transformed stream is replayed from the same initialization, recomputing training, candidate construction, qualification, and promotion without inheriting Real's deployment decisions. Responses remain on their original side of every training–validation boundary.

Reconstructed deployment packages are evaluated through fresh frozen rollouts. Their monitoring records derive from source-trajectory evidence, not on-policy observations of an independently deployed neutralized system. The contrast therefore tests sensitivity to the specified history transformation rather than reproducing an unobserved alternative formation trajectory.

B.3 Caching, State Integrity, and Resource Accounting

Let P be the predictor parameter count per direction, d_e the encoder output width, and $A = |A_{\mathcal{F}}(\kappa)|$. Cached response embeddings reduce frozen-version readout to one context-encoder evaluation, $O(Ad_e)$ scoring, and $O(A)$ normalization and usage aggregation; without caching, response-encoder evaluations are also required. Caches are keyed by model and semantic version. Train, candidate, and active weights each require $O(P)$ storage.

Resource accounting separates inter-agent transmission from private inference and adaptation. Records include transmitted and blocked messages and their tokens, all LLM input/output calls, separately metered reasoning tokens when available, guidance length, and predictor readout/update latency. The same rules include baseline reflection and partner-modeling calls. Generated but blocked messages still incur inference cost. Audit continuations are reported separately from formation and full utility episodes. Lower transmitted-message counts alone do not imply lower total inference cost or wall-clock time.

C Experimental Configuration

This section specifies the benchmark families, relational learner, baseline interfaces, semantic instantiations, and statistical quantities used in the experiments. Formation updates local adaptive states, whereas behavioral audits and complete utility tasks evaluate frozen deployments.

C.1 Benchmark Configurations

The protocol uses Collab-Overcooked, two-player Hanabi, and CoBlock (Sun et al., 2025; Bard et al., 2020; Wu et al., 2024). For Collab-Overcooked, the selected family follows a Level-4-style two-ingredient workflow. Generated instances vary legal ingredient, workstation, and initial-state configurations while preserving role capabilities, processing dependencies, and handoff semantics. A common prompt adapter excludes unexecuted partner plans and mandatory request-first instructions without removing asymmetric recipe access.

Hanabi uses the standard two-player full game with five colors, five ranks, five cards per hand, eight information tokens, and three life tokens. Instances differ in deck order and starting-player assignment. CoBlock uses goal-dependent construction with a fixed role-level dependency and capability structure. Instances vary geometry, subgoal partitions, inventories, and initial structures subject to these invariants. Both agents retain the native construction and communication interface, with one command per agent per round in the selected wrapper. These task-family choices are study-specific, not official benchmark splits.

Table 7: Experimental sample sizes and audit allocation.

Setting	Overcooked	Hanabi	CoBlock
Shared formation settings			
Formation instances per run, N	48	48	48
Warmup instances, W	8	8	8
Prospective block, L_{val}	16	16	16
Main runs per backbone	8	8	8
Planned sampling allocation			
Development instances	32	64	32
Main audit source instances	24	32	24
Hard audit source instances	24	32	24
Snapshots per source instance and direction	1	1	1
Draws per snapshot and condition, M	32	32	32
Full utility instances	96	192	96
Repetitions per utility instance and condition	1	1	1

C.2 Relational Learner Configuration

The two towers have hidden widths $[128, 128]$ and a common output width of 64, with unit-normalized outputs and inner-product scoring. Training minimizes masked complete-response NLL under instance–key–record balancing (Section B.1). Input widths and response catalogs are benchmark-specific.

Table 8: Relational predictor and optimization configuration.

Parameter	Setting
Architecture	
Hidden layers per tower	$[128, 128]$
Common output dimension	64
Scoring / temperature	Normalized inner product / $\tau = 0.2$
Training objective	Masked full-C NLL
Optimization	
Optimizer	AdamW
Learning rate / weight decay	10^{-3} / 10^{-4}
Optimizer steps per instance	4
Gradient-norm clipping	1.0
Replay and sampling	
Replay capacity	512 records
Records per instance–key	At most 4
Historical instances per update	At most 8
Sampling balance	Instance, then key, then record

Each update uses the current instance and bounded historical rehearsal. The historical-instance limit is not a record-level batch size. Predictor temperature τ is separate from LLM decoding temperature.

C.3 Baseline Conditions

B0–B6 and Ours share legal current-task observations, action semantics, and communication rules. All adaptive conditions allow at most 600 characters of additional cross-instance content in a decision-time prompt, excluding common role/task instructions, current observations, and action specifications. Persistent-state updates occur only at formation-instance boundaries, and audit and utility evaluation are read-only.

Table 9: Persistent adaptation and decision-time use of the compared methods. B5 and B6 are shared-interface adaptations of the cited mechanisms.

Condition	Persistent information	Update and decision-time use
B0 Frozen	None	Ordinary current-instance reasoning without cross-instance adaptation.
B1 Raw History	Legal interaction transcripts	Appends the completed instance's transcript and injects its most recent suffix within the 600-character budget.
B2 Reflection	Rolling verbal reflection	One LLM update per agent after each instance produces a reflection capped at 600 characters for later use (Shinn et al., 2023).
B3 Partner Notes	Task and partner notes	One LLM update per agent after each instance maintains notes under the same 600-character cap.
B4 Structured Retrieval	Historical (κ, C) records	Stores records after each instance and uses deterministic context-based ranking to pack records into the text budget, without an LLM retrieval call.
B5 Hypothetical Minds	Partner hypotheses	Updates persistent hypotheses after each instance and uses a budgeted hypothesis summary for planning (Cross et al., 2025).
B6 A-ToM	Partner-reasoning state	Updates persistent partner-reasoning state after each instance and uses partner-action predictions to adapt reasoning order (Mu et al., 2026).
Ours	Directional model and qualification evidence	Learns complete-response distributions, prospectively qualifies candidates, and supplies private distributional guidance. Rehearsal is used for optimization rather than runtime retrieval.

C.4 Semantic Interface Instantiation

Table 10 specifies the context extractor, response alphabet, usage map, and qualification cell for each benchmark.

Table 10: Semantic scope of the benchmark interfaces. Field groups and response evidence are distinguished from final categorical encodings.

Interface	Overcooked	Hanabi	CoBlock
Context κ	Own role, legally known processing stage and demand, observable station/handoff state, own held object, and received coordination information.	Information/life tokens, public pile progress and discards, remaining cards, visible partner cards, and own-card knowledge from received hints.	Own subgoal and inventory, public construction state, observable dependencies, executed partner actions, and received messages.
Response C	Observable response to a handoff need, its station/ingredient target, and response timing.	Observable hint/play/discard response, its card position or hint referent, and response timing.	Observable response to a coordination need through construction or communication, its block/region target, and response timing.
Usage U_i	Self-side handoff participation and timing, retaining distinctions needed for the current participation decision.	Self-side participation informed by uncertainty over partner responses, without prescribing a particular card action.	Self-side participation in dependency resolution and construction timing, without assuming access to private partner goals.
Qualification cell γ	Predeclared $V_i(\kappa)$ over the legal context fields above.	Predeclared $V_i(\kappa)$ over the legal context fields above.	Predeclared $V_i(\kappa)$ over the legal context fields above.
Response clock	Simulator steps.	Executed player actions.	Executed commands in joint execution order.

CoBlock responses summarize legally observable construction and communication, not an assumed physical block-transfer command. Hanabi observations exclude true own cards and future deck order.

C.5 Metrics and Statistical Analysis

Task quality and utility contrasts. J is complete-order success in Overcooked, terminal score out of 25 in Hanabi, and complete target-structure success in CoBlock. L counts transmitted messages in Overcooked and CoBlock, or executed legal hints in Hanabi. With z_0 denoting B0's empty state,

$$\begin{cases} G(z,\beta_*) = J(z,\beta_*) - J(z_0,\beta_*), \\ D_{\mathrm{NI}}(z) = J(z,\beta_*) - J(z_0,\beta_{\mathrm{full}}), \\ \Delta J_{\mathrm{cond}} = J(\mathrm{cond},\beta_*) - J(\mathrm{Real},\beta_*). \end{cases} \tag{24}$$

Non-inferiority tests $H_0 : D_{\mathrm{NI}} \le -\epsilon$ with $\epsilon = 5$ percentage points on the success-rate scale. Communication substitution additionally requires lower measured communication. Hanabi has no restricted-budget non-inferiority claim.

Behavioral response laws. Let $N^g_{r,d}(c \mid x)$ count complete responses of type c and $N^g_{r,d,\mathrm{obs}}(x)$ all complete responses for run r, direction d, condition g, and snapshot x. The smoothed estimator is

$$\widehat{p}^g_{r,d}(c \mid x) = \frac{N^g_{r,d}(c \mid x) + \alpha_{\mathrm{audit}}}{N^g_{r,d,\mathrm{obs}}(x) + |A_d(x)|\alpha_{\mathrm{audit}}}, \qquad c \in A_d(x). \tag{25}$$

The alphabet, positive smoothing parameter, and context weights are shared within a comparison. Zero-observation snapshots are unestimable; coverage and planned-draw denominators remain reported.

The primary effect sizes, with conditioning on x suppressed, are

$$\begin{cases} \widehat{S}^{(n)}_{r,d} = \displaystyle\sum_x \mu_d(x) D_{\mathrm{KL}}\left(\widehat{p}^{(n)}_{r,d} \| \widehat{p}_{0,d}\right), \\ \widehat{\Delta}^{\mathrm{corr}}_{r,d} = \dfrac{1}{K}\displaystyle\sum_{\ell=1}^{K}\sum_x \mu_d(x) D_{\mathrm{KL}}\left(\widehat{p}^{\mathrm{real}}_{r,d} \| \widehat{p}^{\mathrm{neut},\ell}_{r,d}\right), \\ \widehat{\Delta}^{\mathrm{pair}}_{r,d} = \displaystyle\sum_x \mu_d(x) D_{\mathrm{KL}}\left(\widehat{p}^{\mathrm{real}}_{r,d} \| \widehat{p}^{\mathrm{cross}}_{r,d}\right). \end{cases} \tag{26}$$

Effects use natural logarithms and are reported in nats. Hard repeats the same contrasts on its own fixed unseen-instance support.

Two independent, draw-matched Cold samples define the finite-sample floor

$$\widehat{S}^{\mathrm{cold}}_d = \sum_x \mu_d(x) D_{\mathrm{KL}}\left(\widehat{p}^a_{0,d}(\cdot \mid x) \| \widehat{p}^b_{0,d}(\cdot \mid x)\right). \tag{27}$$

This is a finite-sample diagnostic rather than a significance threshold.

Component tests and joint identification. Behavioral equality tests use the symmetric JS statistic

$$T_{r,d}^{g,h} = \sum_x \mu_d(x) D_{\mathrm{JS}}\Big(\widehat{p}_{r,d}^{g}(\cdot \mid x), \widehat{p}_{r,d}^{h}(\cdot \mid x)\Big),$$
$$T_{s,d}^{g,h} = \frac{1}{3}\sum_{b=1}^{3}\frac{1}{8}\sum_{r=1}^{8} T_{s,b,r,d}^{g,h}. \tag{28}$$

Here

$$D_{\mathrm{JS}}(p,q) = \frac{1}{2}D_{\mathrm{KL}}(p\|m) + \frac{1}{2}D_{\mathrm{KL}}(q\|m), \qquad m = \frac{p+q}{2}.$$

At $n = N$, necessary components are combined by intersection–union:

$$\begin{cases} p_{\mathrm{Gen},d} = \max\{p_{\mathrm{Form,H},d}, p_{\mathrm{Corr,H},d}, p_{\mathrm{Pair,H},d}\}, \\ p_{\mathrm{ID},d} = \max\{p_{\mathrm{Form},d}, p_{\mathrm{Corr},d}, p_{\mathrm{Pair},d}, p_{\mathrm{Gen},d}\}, \\ p_{\mathrm{ID,both}} = \max\{p_{\mathrm{ID},A\to B}, p_{\mathrm{ID},B\to A}\}. \end{cases} \tag{29}$$

Holm correction at $\alpha = 0.05$ applies to the three resulting dataset-level bidirectional claims. Utility hypotheses form a separate predeclared family.

D Additional Experimental Results

This section complements the main text with full backbone-level quality results and ablations, communication-substitution analyses, and the complete GTU identification panel. The main text reports GPT-5.5 ablations and the primary sensitivity results; here we provide the corresponding backbone-level breakdowns and supporting behavioral analyses.

D.1 Complete Performance Tables

Table 11 reports B0, B4, and Ours on the two message-based benchmarks under restricted and standard communication.

Table 11: Task quality (mean ± SD over 8 runs) for B0, B4, and Ours on message-based benchmarks. $J(\beta_*)$ uses the restricted budget and $J(\beta_{\mathrm{full}})$ uses standard communication. Higher is better.

		B0 Frozen		B4 Struct.		Ours	
Dataset	**Backbone**	β_*	β_{full}	β_*	β_{full}	β_*	β_{full}
Collab-Overcooked							
	Claude Sonnet 4.6	33.78±4.12	48.93±3.51	47.40±2.57	56.63±3.45	57.93±3.94	59.70±3.99
	GPT-5.5	33.53±2.49	46.37±3.67	47.79±4.56	57.22±4.43	54.38±3.48	56.87±3.06
	Gemini 3.6 Flash	32.04±4.38	46.31±2.76	44.92±3.90	52.73±3.44	55.99±4.51	57.91±4.70
CoBlock							
	Claude Sonnet 4.6	34.07±3.39	50.61±2.99	49.25±2.41	58.00±4.52	55.23±4.06	57.64±4.01
	GPT-5.5	29.62±2.12	50.24±2.84	45.35±2.74	55.01±2.95	54.03±3.52	57.32±3.23
	Gemini 3.6 Flash	33.06±4.56	47.44±2.54	43.86±4.22	52.06±3.73	48.62±3.17	52.74±3.54

Table 12 reports all eight conditions under Hanabi's native hint mechanism.

Table 13 extends the main-text GPT-5.5 ablations to Claude Sonnet 4.6 and Gemini 3.6 Flash.

D.1.1 Formation Dynamics Across Benchmarks

Table 14 extends the main-text formation summary to all three benchmarks and all four tracked metrics across nine instance checkpoints. Predictive loss $\widehat{L}_B$ improves before the first promotion checkpoint; behavioral response shift $\widehat{S}$ follows after promotion, consistent with the versioned deployment mechanism. The shaded column ($n = 24$) marks the first qualification boundary under $W = 8$ and $L_{\mathrm{val}} = 16$; a successfully promoted state is first usable from instance 25.

Table 12: Task quality for Hanabi and hints per game under native communication rules. No artificial budget restriction applies.

	Claude 4.6		GPT-5.5		Gemini 3.6 F	
Condition	Score	Hints	Score	Hints	Score	Hints
B0 Frozen	13.05±1.09	7.00±0.27	10.89±0.74	7.78±0.14	9.18±0.74	7.08±0.26
B1 Raw Hist.	13.17±0.68	7.53±0.23	11.58±1.06	7.85±0.15	9.38±0.85	7.11±0.26
B2 Reflection	13.12±1.04	7.31±0.17	11.81±0.60	7.28±0.26	8.77±0.38	6.80±0.18
B3 Part. Notes	14.42±0.66	7.19±0.19	12.53±0.72	7.36±0.28	10.34±0.67	6.72±0.15
B4 Struct. Ret.	14.31±0.77	6.97±0.23	13.98±0.89	7.21±0.18	11.00±0.81	6.63±0.11
B5 Hyp. Minds	15.41±0.81	6.84±0.25	13.49±0.41	7.12±0.20	11.26±0.46	6.35±0.12
B6 A-ToM	14.96±0.69	6.99±0.20	13.69±0.72	7.04±0.13	11.52±0.49	6.69±0.21
Ours	16.24±1.17	6.77±0.22	14.45±1.23	7.11±0.13	12.49±0.77	6.48±0.13

Values are means ± SD over 8 runs. Highlighted cells indicate the highest Score in each backbone column. Hints per game are reported as communication load.

Table 13: Ablation task quality ($J(\beta_*)$, mean ± SD over 8 runs) across all three LLM backbones. Higher is better.

		Overcooked			CoBlock		
Ablation	**Removed component**	Claude	GPT	Gemini	Claude	GPT	Gemini
Full (Ours)	—	**57.93**±3.86	**54.38**±2.44	**55.99**±4.19	**55.23**±4.33	**54.03**±4.74	**48.62**±3.62
No Qualification	Prospective gate	55.02±3.12	51.99±2.93	52.44±3.88	51.66±3.41	51.14±2.86	45.68±4.16
No Monitoring	Post-deploy cell disable	55.38±4.47	50.60±5.18	53.29±4.62	53.20±4.05	53.18±5.09	47.94±3.55
No Usage Sem.	$\mathcal{U}_i = C$	54.13±3.55	51.70±4.16	51.87±3.24	49.05±2.67	47.51±1.93	43.10±3.98
Argmax Guidance	Replace sum by argmax	55.71±2.98	51.30±3.60	53.71±3.07	53.95±3.79	52.59±4.25	47.11±2.84
No Rehearsal	Remove historical buffer	53.60±4.21	50.46±4.12	51.36±4.35	47.83±4.94	46.56±5.63	43.94±4.77
Frozen Predictor	No learning	46.88±5.63	43.61±6.97	44.72±5.91	43.09±3.28	41.99±2.96	37.55±3.11
Shared Learner	Tied $\theta_A = \theta_B$	55.94±4.88	52.72±5.59	53.02±4.13	46.15±4.51	44.14±5.20	41.29±4.42
Generic Guidance	Task text replaces g^{tac}	42.35±4.72	38.58±5.66	39.94±5.28	39.72±5.17	38.53±5.61	34.02±4.89

D.2 Communication Substitution Analysis

The NI claim requires both adequate quality under $\beta_* = 0.5$ and lower measured communication than the corresponding standard-communication reference. Blocked messages are therefore recorded separately from transmitted messages because blocked generations still incur inference cost and reveal budget pressure.

For Ours, guidance injection is reported separately from communication: a nonempty injection indicates that a prospectively qualified active state was available at that decision step but does not itself count as inter-agent transmission.

D.3 Complete GTU Condition Verification

The full panel reports direction-level component effects and IUT p-values. Final GTU identification is dataset-level: the two directional IUT values are combined by their maximum, followed by Holm correction across the three dataset-level bidirectional claims.

Table 17 provides a descriptive comparison of behavioral identification effect sizes across baselines; it is not an additional baseline-wise GTU test.

Figure 3 visualizes the two directional state-attribution effects. These interventions support method-level attribution and remain separate from the GTU identification criteria.

D.4 Sensitivity Summary

Sensitivity experiments use GPT-5.5 with eight independent runs on Overcooked and CoBlock, varying $N \in \{8, 16, 24, 32, 48\}$, $\beta \in \{0.3, 0.5, 0.7, 1.0\}$, and $L_{\text{val}} \in \{4, 8, 16, 32\}$ around the defaults $N = 48$, $\beta_* = 0.5$, and $L_{\text{val}} = 16$. For the N sweep, $N = 8$ and $N = 16$ both precede the first qualification boundary; candidate Brier scores may therefore change at these checkpoints, whereas changes in J or $\widehat{S}$ should not be interpreted as effects of deployed guidance.

Table 14: Formation dynamics: mean ± SD over 8 runs at each instance checkpoint, grouped by benchmark and metric. Shaded column ($n = 24$) marks the first qualification boundary.

Benchmark	Metric	$n = 0$	$n = 2$	$n = 4$	$n = 8$	$n = 12$	$n = 16$	$n = 24$	$n = 32$	$n = 48$
Collab-Overcooked	*(promotion at $n = 24$)*									
	$J \uparrow$	33.12±2.10	33.60±2.85	34.18±3.59	35.77±4.14	37.99±4.27	40.83±4.08	47.28±3.49	52.34±2.84	56.10±2.06
	$L \downarrow$	2.42±0.09	2.41±0.13	2.40±0.17	2.39±0.23	2.38±0.26	2.37±0.21	2.36±0.18	2.35±0.16	2.35±0.11
	$\widehat{L}_B \downarrow$	0.282±0.023	0.235±0.031	0.198±0.038	0.144±0.043	0.110±0.045	0.088±0.041	0.064±0.038	0.055±0.031	0.049±0.023
	$\widehat{S} \uparrow$	0.022±0.026	0.027±0.028	0.025±0.024	0.030±0.025	0.028±0.028	0.033±0.028	0.041±0.029	0.220±0.049	0.310±0.029
Hanabi	*(promotion at $n = 24$)*									
	$J \uparrow$	11.04±0.43	11.12±0.68	11.22±0.92	11.49±1.02	11.88±1.08	12.35±1.06	13.34±0.92	13.99±0.74	14.39±0.48
	$L \downarrow$	7.29±0.09	7.24±0.12	7.20±0.13	7.13±0.18	7.06±0.19	7.01±0.17	6.94±0.13	6.89±0.12	6.84±0.09
	$\widehat{L}_B \downarrow$	0.265±0.019	0.213±0.028	0.174±0.029	0.120±0.035	0.089±0.035	0.071±0.032	0.053±0.032	0.047±0.026	0.045±0.019
	$\widehat{S} \uparrow$	0.023±0.021	0.023±0.024	0.024±0.024	0.026±0.027	0.028±0.026	0.031±0.028	0.038±0.029	0.210±0.042	0.280±0.028
CoBlock	*(promotion at $n = 24$)*									
	$J \uparrow$	32.25±1.95	32.63±2.70	33.09±3.33	34.32±3.75	36.02±3.84	38.21±3.70	43.53±3.25	48.30±2.70	52.63±1.95
	$L \downarrow$	2.12±0.08	2.12±0.11	2.13±0.14	2.13±0.15	2.13±0.16	2.14±0.15	2.14±0.11	2.15±0.11	2.15±0.08
	$\widehat{L}_B \downarrow$	0.271±0.021	0.225±0.029	0.188±0.036	0.137±0.029	0.105±0.045	0.085±0.040	0.064±0.036	0.056±0.029	0.052±0.020
	$\widehat{S} \uparrow$	0.023±0.022	0.024±0.022	0.022±0.027	0.026±0.030	0.026±0.027	0.030±0.031	0.040±0.029	0.205±0.037	0.290±0.028

Promotion instance (first checkpoint at which a qualified active state exists): $n = 14$ for Overcooked and CoBlock, $n = 10$ for Hanabi. $\widehat{L}_B$ falls before promotion; $\widehat{S}$ rises sharply only after, consistent with the versioned deployment mechanism (Section 4).

Table 15: Communication statistics at $\beta_* = 0.5$ for Ours and B0. L^{tx}: transmitted; L^{blk}: blocked. η_g: guidance coverage.

		B0 Frozen		Ours		
Dataset	**Backbone**	L^{tx}	L^{blk}	L^{tx}	L^{blk}	η_g
Collab-Overcooked						
	Claude 4.6	2.38	4.14	2.39	0.19	0.74
	GPT-5.5	2.48	3.82	2.32	0.16	0.72
	Gemini 3.6 F	2.41	4.34	2.31	0.21	0.67
CoBlock						
	Claude 4.6	2.09	6.04	2.16	0.07	0.69
	GPT-5.5	2.16	5.52	2.25	0.05	0.64
	Gemini 3.6 F	2.12	5.36	2.08	0.09	0.62

Transmitted-message values use the same $L(\beta_*)$ definition as the main-text table. Blocked generations incur inference cost but are not counted as transmitted communication.

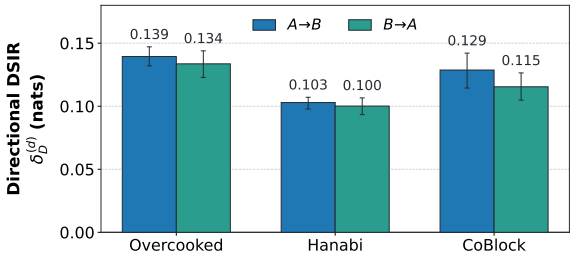


Figure 3: Directional attribution of the learned behavioral state across benchmarks. The two directions are reported separately; these interventions are method-level attribution analyses rather than additional GTU component tests.

Scope and limitations. DTU is studied for persistent pairs of frozen black-box LLMs within prespecified coordination families and legally observable interfaces, so its conclusions remain conditional on those semantics, interventions, and audit supports. GTU identifies partner-specific behavioral structure but does not guarantee task benefit, and response-shift strength need not grow monotonically with formation length. Finite audits support panel-level dataset claims rather than individual-dyad judgments. Cross-family transfer, changing partners, larger teams, stronger partial-observation controls, and explicit accounting of private inference cost remain direct extensions.

Table 16: Complete IUT panel for Ours across all six direction–dataset cells. Main = standard difficulty tier; Hard = held-out hard tier. $\widehat{S}$, $\widehat{\Delta}^{\mathrm{corr}}$, $\widehat{\Delta}^{\mathrm{pair}}$: response-shift strength, correspondence specificity, and partner specificity.

		Main tier			Hard tier			
Dataset	**Direction**	$\widehat{S}$	$\widehat{\Delta}^{\mathrm{corr}}$	$\widehat{\Delta}^{\mathrm{pair}}$	$\widehat{S}$	$\widehat{\Delta}^{\mathrm{corr}}$	$\widehat{\Delta}^{\mathrm{pair}}$	Dir. IUT p
Collab-Overcooked								
	$A \to B$	0.286	0.1966	0.1876	0.2619	0.1623	0.1541	0.0041
	$B \to A$	0.292	0.2018	0.1724	0.2604	0.1431	0.1613	0.0087
Hanabi								
	$A \to B$	0.244	0.1886	0.1731	0.2164	0.1431	0.1273	0.0134
	$B \to A$	0.218	0.1576	0.1691	0.1927	0.1089	0.1083	0.0218
CoBlock								
	$A \to B$	0.304	0.2294	0.1781	0.2806	0.1675	0.1359	0.0009
	$B \to A$	0.278	0.2089	0.1982	0.2591	0.1689	0.1449	0.0063

Table 17: Descriptive identification effect sizes over six direction–dataset cells for GPT-5.5. Formal GTU inference uses the predeclared full-method IUT only.

		Effect size (avg. over 6 cells)		
Condition	**Description**	$\widehat{S}$	$\widehat{\Delta}^{\mathrm{corr}}$	$\widehat{\Delta}^{\mathrm{pair}}$
B0 Frozen	No history	0.0143	0.0085	0.0054
B1 Raw Hist.	Unstructured history prefix	0.0564	0.0367	0.0261
B2 Reflection	Free-form self-reflection	0.0787	0.0481	0.0362
B3 Partner Notes	Unstructured partner notes	0.0993	0.0684	0.0436
B4 Struct. Ret.	Structured partner template	0.1354	0.0925	0.0769
B5 Hyp. Minds	Hypothetical-minds prompt	0.1543	0.1165	0.1045
B6 A-ToM	Analytic ToM prompt	0.1596	0.1363	0.1058
Ours	TACT (full)	**0.2746**	**0.2012**	**0.1849**

Metrics are averaged within each direction–dataset cell and then across the six cells. These values are descriptive and receive no baseline-wise GTU decision or Holm correction.