COORDLEDGER: THE COORDINATION TAX OF MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent LLM systems are usually judged by whether they outperform a single agent. Yet a multi-agent win can come from three different sources—more inference calls, access to more information, or better coordination—and standard comparisons cannot tell them apart. We introduce COORDLEDGER, a benchmark that compares each communicating system against the single-agent references implied by its deployment access contract. When all task-critical information could be given to one model, the deployment is *centralizable*, and the reference is a compute-matched single model. When no single role may see all of it, the deployment is *structurally distributed*, and we use two references: a single agent under the same access restriction, and a compute-matched single agent that sees everything. COORDLEDGER contains 1,957 instances across five tracks, with 58,248 frozen core runs over six model configurations from five model families. The two regimes behave very differently. On centralizable tasks, the compute-matched single model matches or beats both the Star and Chain topologies in five of six configurations. On structurally distributed tasks, communication clearly helps: Chain raises mean accuracy from 0.188 for the access-restricted agent to 0.335. But the all-access agent reaches 0.495, so communication recovers less than half (48%) of the value that pooling information makes available. We call the remaining gap the *coordination tax*. Stage-level traces show where it is paid: as information becomes distributed, the dominant failure shifts from delivering evidence to integrating it at the final decision. A multi-agent gain, in short, is only interpretable against the reference its deployment permits.

1 INTRODUCTION

Large language models are increasingly deployed as teams: several model instances take on roles, use tools, observe different parts of a task, and exchange intermediate results Li et al. (2023); Wu et al. (2024); Chen et al. (2024); Hong et al. (2024). The appeal is intuitive. Agents that each see part of a problem should, by talking, be able to solve what none of them could solve alone Du et al. (2024); Shen et al. (2025). Evaluations usually test this intuition by placing a communicating system next to a single agent and comparing scores. When the team wins, however, it is rarely clear *what* won.

A multi-agent gain can come from three places. The team may simply receive more inference calls than the single agent it is compared against. It may collectively see information that the single agent never received. Or communication itself may help—or hurt—in moving that information to where the decision is made. End-task accuracy mixes all three. Compute-aware baselines already show that much of the apparent advantage shrinks once inference opportunity is controlled Wang et al. (2024b;a); Smit et al. (2024); Kim et al. (2025), and recent benchmarks show that distributed information and communication constraints change performance substantially Zhang et al. (2026); Park et al. (2026); Yang et al. (2026); Li et al. (2026b). What these studies leave open is which single-agent reference a given comparison should use, and what each reference actually measures.

We argue that the appropriate reference depends on the deployment access contract (Figure 1): which information each permitted execution may read. Consider the same supplier-selection task under two contracts. One role holds the compliance rules, another holds the delivery times, and the right supplier depends on both. If a single model may read both views, the deployment is *centralizable*, and the team should be compared with a strong single model under the same inference budget. If no

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.

permitted role may jointly read both, the deployment is *structurally distributed*: communication now creates access value, but the relevant question is how much of that available value the team actually realizes. The same task thus calls for a different reference under a different contract.

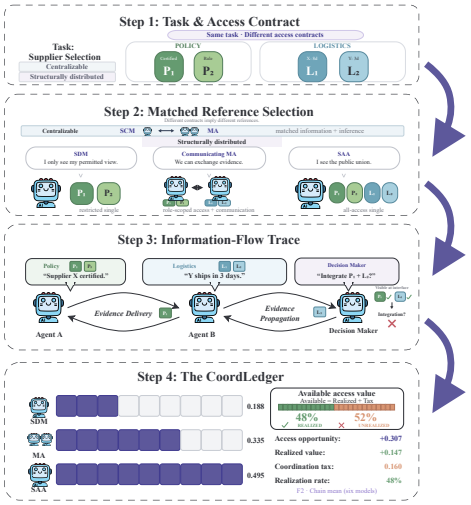

Figure 1: **COORDLEDGER: from deployment access contracts to coordination accounting.** Holding the task fixed, the access contract determines the matched reference: S_{CM} for centralizable deployments, and S_{DM} plus the counterfactual all-access S_{AA} for structurally distributed deployments. The ledger separates value realized through communication from the remaining coordination tax; on F2/Chain, communication realizes 48% of the available access value.

COORDLEDGER turns this principle into a benchmark. It contains 1,957 instances across five tracks—Controlled Communication, Natural Transfer, Access-Constrained Integration, Distributed-State Observability, and Stateful Tool Interaction—with 58,248 frozen core runs over six model configurations from five model families. On centralizable tracks, we compare two standard communication topologies, Star and Chain, with a Compute-Matched Single (S_{CM}). On structurally distributed tracks, we add two references: a Deployment-Matched Single (S_{DM}) that keeps the access restriction without communication, and an All-Access Compute-Matched Single (S_{AA}) that reads the union of public role views under the same inference budget. Writing the deployment value as Δ_{deploy}, the access-associated opportunity as Δ_{access}, and the coordination residual as Δ_{coord}, the three references form a ledger:

$$\Delta_{\mathrm{deploy}} = \Delta_{\mathrm{access}} + \Delta_{\mathrm{coord}}. \tag{1}$$

When $\Delta_{\mathrm{coord}} < 0$, the team realizes less than the all-access reference; we call this shortfall, $-\Delta_{\mathrm{coord}}$, the *coordination tax*.

The two regimes look very different. On centralizable Controlled Communication tasks, S_{CM} matches or exceeds both Star and Chain in five of six configurations: once one model has the same information and inference budget, communication does not help. On the structurally distributed Access-Constrained Integration track, communication is clearly valuable: Chain raises mean accuracy from 0.188 under S_{DM} to 0.335. The all-access reference, however, reaches 0.495. Chain therefore recovers 0.147 of an available 0.307—less than half—and falls short of S_{AA} in all six configurations.

Stage-level traces show where the tax is paid. We follow the required evidence through four stages—delivery, propagation, visibility at the decision interface, and integration—and assign a stage only when the recorded trace supports it. On Controlled Communication, most recorded failures occur early: evidence is never delivered or is lost in transit. On the distributed tracks, integration dominates, accounting for 0.528 of all scheduled runs on Access-Constrained Integration and 0.737 on Distributed-State Observability. Having the information is increasingly not the same as using it. Targeted interventions confirm the distinction: preserving evidence in Star reduces delivery failures, and carrying evidence forward in Chain nearly eliminates propagation loss.

In summary, our contributions are threefold. (1) We make the deployment access contract an explicit part of multi-agent evaluation and release COORDLEDGER, a 1,957-instance benchmark covering both centralizable and structurally distributed settings. (2) We introduce matched references that control inference opportunity and decompose a multi-agent gain into realized deployment value, access-associated opportunity, and the coordination tax. (3) We pair these contrasts with stage-level information-flow diagnostics and with topology and communication-budget stress tests, showing under which contracts communication creates value, how much of the available value it realizes, and where the rest is lost.

2 Related Work

Most comparisons in prior work can be read as one of the three references in our ledger (Section 3): a centralized single model with the same information (S_{CM}), a single agent under the deployment's access restriction (S_{DM}), or a single agent that sees everything (S_{AA}).

Agent evaluation. Recent benchmarks increasingly study interactive, stateful, and long-horizon agent behavior (Liu et al., 2024; Yao et al., 2025; Barres et al., 2025; Li et al., 2026a; Xi et al., 2026). They typically fix an environment and interface and measure performance within it; CoordLedger instead holds the task fixed and asks which single-agent reference makes a multi-agent score interpretable.

Centralized references. Apparent multi-agent gains often shrink when a single model receives matched information and inference opportunity (Wang et al., 2024b;a; Smit et al., 2024; Choi et al., 2025; Kim et al., 2025; Tran & Kiela, 2026). These comparisons correspond to S_{CM} and presuppose that a permitted centralized execution can observe everything the team sees; test-time-compute work further shows why matching inference opportunity matters (Wang et al., 2023; Snell et al., 2025).

Distributed information. Multi-agent benchmarks increasingly vary topology and context management or make fragmented views and privacy constraints part of the task (Zhu et al., 2025; Wang et al., 2025; Sun et al., 2025; Zhang et al., 2026; Park et al., 2026; Yang et al., 2026). Comparisons with complete-information references reveal substantial shortfalls in hidden-profile, expert-team, and multi-turn settings (Stasser & Titus, 1985; Li et al., 2026b; Pappu et al., 2026; Laban et al., 2026). These results motivate S_{AA} but do not measure value relative to what the deployment otherwise allows. CoordLedger adds S_{DM} and a realization rate, paralleling the gap-recovered logic of information-asymmetric debate (Khan et al., 2024; Burns et al., 2024).

Failure attribution. Trace-based work attributes multi-agent failures to agents, steps, propagation paths, or failure modes (Cemri et al., 2025; Zhang et al., 2025b; Chen et al., 2026; Shen et al., 2025). CoordLedger instead localizes loss along a fixed information path and links stage attribution to the outcome ledger. Appendix Table A1 gives the dimension-by-dimension comparison.

3 CoordLedger

CoordLedger evaluates a communicating system against the references its deployment permits. We first define access contracts and the references they admit (Section 3.1), then the ledger that compares them (Section 3.2), the diagnostics that localize its losses (Section 3.3), and finally the five tracks that instantiate it (Section 3.4).

3.1 Access Contracts and Admissible References

Instances. The unit of evaluation is a task–deployment pair rather than a question in isolation. Each instance specifies (i) the task, (ii) the public information available to the deployment, (iii) the role-scoped views into that information, (iv) the access contract governing which views may be observed together, and (v) the target outcome. Where the representation permits, instances also annotate the task-critical semantic units used for diagnosis. Gold answers, gold derivations, and diagnostic annotations are evaluator-only and never enter a model's input.

Access contracts. Let $\mathcal{U}$ be the public information units of an instance, $\mathcal{U}^\star \subseteq \mathcal{U}$ the task-critical units, and $V_r \subseteq \mathcal{U}$ the view held by role $r \in \mathcal{R}$. An access contract specifies the set $\mathcal{E}$ of permitted single executions, each observing a view $V(e) \subseteq \mathcal{U}$ formed from role views that the contract allows to be read together. The deployment is *centralizable* if some permitted execution observes all task-critical information,

$$\exists\, e \in \mathcal{E} : \mathcal{U}^\star \subseteq V(e), \tag{2}$$

and *structurally distributed* otherwise. The regime is fixed by the contract before any model is run: it is a property of the deployment, not of task difficulty or protocol success. For tracks without unit-level annotation, $\mathcal{U}^\star$ is the full public information and the regime follows from construction.

Admissible references. The contract also determines which single-agent references are meaningful. In a centralizable deployment, the *Compute-Matched Single* (S_{CM}) observes the complete permitted centralized view. In a structurally distributed deployment, the *Deployment-Matched Single* (S_{DM}) respects the restriction: it processes role views one at a time, never observing more than a permitted view $V(e)$, and exchanges no messages with other agents. The *All-Access Compute-Matched Single* (S_{AA}) observes the union $\bigcup_{r\in\mathcal{R}} V_r$. S_{AA} is not deployable under the contract; it answers what the same model would achieve if the restriction were lifted. All references follow the matched inference contract of Section 3.2.

3.2 The Collaboration Ledger

Matched inference. All contrasts are computed within a fixed model configuration, protocol, metric, and frozen instance set. Compared arms share the same provider-call opportunity and generation-budget contract; realized tokens and latency are logged separately (Appendix H) but are not equalized. Protocol and access contracts for every arm are listed in Appendix C.

Centralizable deployments. Let p_M and p_C denote the performance of a communicating system and of S_{CM} on the same instances. We report

$$p_M - p_C, \tag{3}$$

which asks whether communication adds value beyond a centralized model with the same information and inference opportunity.

Structurally distributed deployments. Let p_M, p_D, and p_A denote the performance of the communicating system, S_{DM}, and S_{AA}. The ledger consists of three contrasts,

$$\begin{aligned} \Delta_{\mathrm{deploy}} &= p_M - p_D, \quad \Delta_{\mathrm{access}} = p_A - p_D, \\ \Delta_{\mathrm{coord}} &= p_M - p_A, \quad \Delta_{\mathrm{deploy}} = \Delta_{\mathrm{access}} + \Delta_{\mathrm{coord}}. \end{aligned} \tag{4}$$

Δ_{deploy} is the value communication delivers over what the deployment otherwise allows; Δ_{access} is the value that lifting the restriction makes available; and Δ_{coord} is the remaining difference between the team and the all-access reference. When $\Delta_{\mathrm{coord}} < 0$, we call $-\Delta_{\mathrm{coord}}$ the *coordination tax*.

Realization rate. When $\Delta_{\mathrm{access}} > 0$, the fraction of the available value that communication realizes is

$$\rho \;=\; \frac{\Delta_{\mathrm{deploy}}}{\Delta_{\mathrm{access}}} \;=\; 1 + \frac{\Delta_{\mathrm{coord}}}{\Delta_{\mathrm{access}}}, \tag{5}$$

and $1-\rho$ is the unrealized share. ρ plays the role that performance gap recovered plays in weak-to-strong and debate evaluation (Burns et al., 2024; Khan et al., 2024). We report ρ per configuration with paired-bootstrap intervals and summarize across configurations by the ratio of mean contrasts.

An instance-level view. Because outcomes are binary and paired by instance, the tax separates into two directions. Writing $y_M, y_D, y_A \in \{0,1\}$ for per-instance correctness,

$$-\Delta_{\mathrm{coord}} \;=\; \underbrace{\Pr[\,y_A = 1,\; y_M = 0\,]}_{\text{loss } L} \;-\; \underbrace{\Pr[\,y_A = 0,\; y_M = 1\,]}_{\text{gain } G}. \tag{6}$$

L counts instances where the pooled information sufficed for the all-access model but the team failed; G counts instances where the team succeeded and the all-access model did not, for example because each role reasons over a shorter context. A positive tax means that losses outweigh gains, not that gains are absent. Restricting the diagnostics of Section 3.3 to L localizes the tax itself.

Interpretation. The identity in Equation 4 is an accounting identity over matched outcome contrasts. The coordination tax therefore measures the shortfall of a given protocol relative to S_{AA} under the benchmark contract, wherever between role-scoped access and the final decision that shortfall arises; it is not a general causal effect of coordination.

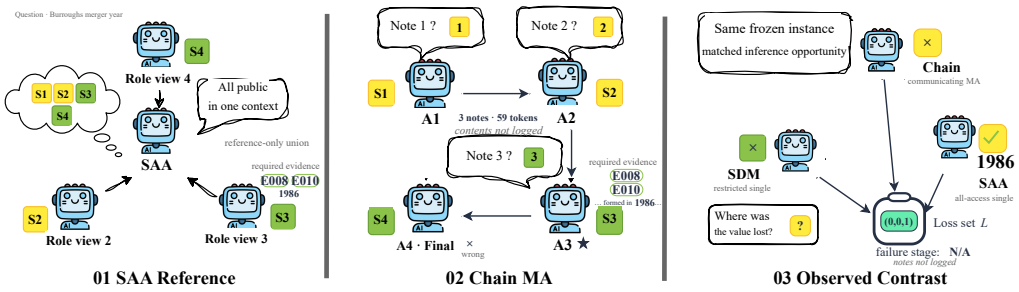

Figure 2: **A real F2 loss instance under matched references. (01) SAA Reference:** the all-access single agent observes the union of the public role views; the required evidence E008/E010 in Role View 3 supports the correct answer, 1986. **(02) Chain MA:** the same evidence is available to Agent A3, but the chain passes three notes totaling 59 tokens and produces an incorrect final answer; the note contents were not logged. **(03) Observed Contrast:** on the same frozen instance under matched inference opportunity, S_{DM} and Chain are incorrect whereas S_{AA} is correct, yielding $(y_D, y_M, y_A) = (0, 0, 1)$, a loss-set instance L. Because the note contents are unavailable, the outcome contrast is observed but the failure stage cannot be uniquely localized and remains N/A.

A worked instance. Figure 2 illustrates one loss-set instance from Access-Constrained Integration (F2). The frozen question asks for the year of the Burroughs merger, and the required evidence E008/E010 is contained in Role View 3. The all-access S_{AA} reference observes the union of the public shards and correctly answers 1986. Under Chain, Agent A3 has access to the required evidence, but after three inter-agent notes totaling 59 tokens, the final answer is incorrect; S_{DM} is also incorrect. The instance therefore contributes $(y_D, y_M, y_A) = (0, 0, 1)$ and belongs to the loss set L: pooled access suffices for S_{AA}, while the communicating system fails. Because the Chain note contents were not logged, the trace identifies the outcome contrast but not the stage at which the evidence was lost; the failure stage therefore remains N/A.

3.3 Information-Flow Diagnostics

The ledger measures how much value is lost; the diagnostics ask where. For a failed run of a communicating system, we follow the annotated task-critical units from the roles that hold them to the component that produces the final decision, checking four stages in order: *delivery*, whether the required units enter the communication path; *propagation*, whether delivered units survive subsequent messages; *decision-interface observability*, whether they are visible to the component that produces the final decision; and *integration*, whether that decision is correct once they are visible. A failure is attributed to the first stage whose check fails. Each check is evaluated on logged fields—message contents, routing, and the inputs of the decision component. When a check cannot be evaluated from the record, for example because a message was sent but its contents were not logged, the run is assigned only to the coarser category the record supports, or to N/A. Parser, format, timeout, and other runtime failures are recorded separately from these semantic stages. Decision-interface observability is additionally tested with controlled visibility interventions (Appendix I).

3.4 Benchmark Composition

CoordLedger contains 1,957 core instances across five tracks (Table 1; Figure 4). Controlled Communication and Natural Transfer are centralizable and use S_{CM}; F2 and F3 distribute task-critical evidence or state and use the full three-reference ledger. Stateful Tool Interaction contains 77 state-changing airline and retail tasks and is retained as a floor-limited headroom evaluation. Dataset-level composition and post-core stress-test slices are detailed in Appendix B.

4 Experimental Setup

Models. We evaluate six open-weight configurations from five model families: Qwen3-8B-AWQ, Qwen3-32B-AWQ, Llama-3-8B-Instruct (GGUF Q4_0), Gemma-3-12B-IT, Ministral-3-14B-Instruct-

Table 1: **Core COORDLEDGER benchmark composition.** The five core tracks contain 1,957 unique instances. Topology and communication-budget variants are post-core stress tests reported in the appendix.

Track	Construction	N	Evaluation role
Controlled Communication	Control-A/B/C-V2/D	480	mechanism isolation
Natural Transfer	five public dataset families	1,000	external transfer
Access-Constrained Integration (F2)	distributed evidence/access	200	primary identification
Distributed-State Observability (F3)	distributed state observations	200	hard stress
Stateful Tool Interaction	retail and airline tasks	77	state/tool headroom

2512 (FP8), and Phi-4. Open weights let us freeze the serving configuration of every arm and release it with the benchmark; the panel spans 8B to 32B parameters. Because serving formats and quantization differ across families, we compare configurations only through within-configuration contrasts, never through absolute scores.

Protocols and references. Each configuration is evaluated with Full-Information Single, No Communication, Star, Chain, and the matched single reference of Section 3.2 (S_{CM} or S_{DM}); structurally distributed tracks additionally include S_{AA}. Star aggregates role outputs centrally, whereas Chain passes messages sequentially. The frozen core contains 58,248 scheduled runs over 1,957 instances. Full arm semantics and release manifests are in Appendices C and N.

Execution and matching. Runs use frozen prompts, parsers, access contracts, greedy decoding ($T = 0$), and no outcome-dependent retries. Compared arms share provider-call opportunity and generation-budget contracts; realized token use, communication, latency, and failures are logged separately. Core communication is capped at $B_{\mathrm{comm}} = 512$ transmitted tokens. Prompt and accounting details are in Appendices D and H.

Statistics. Task performance uses the frozen correctness criterion of each track; Stateful Tool Interaction uses exact final-state success. Primary comparisons are paired by instance within configuration and protocol. We report paired-bootstrap 95% intervals (10,000 resamples) and paired McNemar tests with Holm correction within each prespecified family and configuration; intervals are unadjusted. Intent-to-run is primary and retains parser, format, and runtime failures; the common-evaluable sensitivity analysis is reported in Appendix L.

5 RESULTS

We ask three questions: when communication creates value (Section 5.1), how much of the available value it realizes (Section 5.2), and where the rest is lost (Section 5.3). Section 5.4 tests robustness. Table 2 reports the configuration-level ledger.

5.1 WHEN DOES COMMUNICATION CREATE VALUE?

Against a matched centralized model, communication does not help. On Controlled Communication, communication is far better than no communication and still worse than a single model with the same information. Averaged over configurations, Star raises accuracy from 0.231 without communication to 0.780, yet S_{CM} reaches 0.884 (Appendix E). Star falls short of S_{CM} in five of six configurations (mean -0.105; the exception is Llama-3-8B at $+0.015$), and Chain in all six (mean -0.442; Table 2). A gain over isolated agents thus does not survive a reference that has the same task-critical information and inference opportunity.

On natural multi-hop data, Star reaches parity. On Natural Transfer, Star matches S_{CM} on average (mean $+0.002$, positive in four of six configurations), while Chain remains below it (mean -0.191). Contrasts vary across constituent datasets (Table A8). The centralizable result therefore changes in magnitude across tasks, but communication does not overtake the matched single model.

When access is structurally distributed, communication creates value. On Access-Constrained Integration (F2), Chain improves over S_{DM} in all six configurations (mean $\Delta_{\mathrm{deploy}} = +0.147$);

Table 2: **Matched results on CoordLedger.** Left: Controlled Communication contrasts against S_{CM}. Right: the Chain ledger on Access-Constrained Integration (F2), with realization rate $\rho = \Delta_{\mathrm{deploy}}/\Delta_{\mathrm{access}}$. The mean row averages contrasts across configurations; its ρ is the ratio of mean contrasts. Within each numeric column, the best, second-best, and third-best model-level values are highlighted in indigo, lavender, and gray-purple, respectively; higher is better, so for Δ_{coord}, values closer to zero rank higher. $^{\circ}$Unadjusted paired-bootstrap 95% interval includes zero; all other F2 intervals exclude zero. Absolute accuracies and intervals are in Appendix E and Table A7.

	Centralizable Controlled Communication		**Structurally Distributed** F2 · Chain			
Model	**Star**$-S_{\mathrm{CM}}$	**Chain**$-S_{\mathrm{CM}}$	Δ_{deploy}	Δ_{access}	Δ_{coord}	ρ
Qwen3-8B	-0.148	-0.375	$+0.105$	$+0.285$	-0.180	0.37
Qwen3-32B	-0.208	-0.521	$+0.160$	$+0.320$	-0.160	0.50
Llama-3-8B†	$\mathbf{+0.015}$	$\mathbf{-0.331}$	$\mathbf{+0.290}$	$+0.330$	$\mathbf{-0.040}^{\circ}$	**0.88**
Gemma-3-12B	-0.102	-0.450	$+0.135$	$+0.280$	-0.145	0.48
Ministral-3-14B	-0.183	-0.479	$+0.120$	$\mathbf{+0.350}$	-0.230	0.34
Phi-4	-0.002	-0.498	$+0.075$	$+0.280$	-0.205	0.27
Mean	$\mathbf{-0.105}$	$\mathbf{-0.442}$	$\mathbf{+0.147}$	$\mathbf{+0.307}$	$\mathbf{-0.160}$	**0.48**

†The Llama-3-8B S_{DM} arm fails on output formatting, which inflates its Δ_{deploy}, Δ_{access}, and ρ but does not enter Δ_{coord}. Excluding it, the means are $\Delta_{\mathrm{deploy}} = +0.119$, $\Delta_{\mathrm{access}} = +0.303$, $\Delta_{\mathrm{coord}} = -0.184$, and $\rho = 0.39$.

unadjusted intervals exclude zero in all six, and five of six tests remain significant after Holm correction. Distributed-State Observability (F3) points the same way with weaker evidence: all six estimates are positive (mean $\Delta_{\mathrm{deploy}} = +0.055$), intervals exclude zero in three configurations, and two tests survive correction.

The Llama-3-8B configuration makes these estimates conservative. Its S_{DM} arm fails on output formatting, which inflates Δ_{deploy} and Δ_{access} but does not enter Δ_{coord}. Without it, mean Δ_{deploy} falls to $+0.119$ and remains positive in every configuration, while the unrealized share rises from 52% to 61% (Section 5.2). We keep the configuration under the prespecified intent-to-run analysis; the common-evaluable sensitivity analysis is in Appendix L.

5.2 How Much of the Available Value Is Realized?

Chain realizes less than half of the available value. Figure 3 summarizes the matched-reference picture across regimes and highlights the F2/Chain realization gap. Mean accuracy rises from 0.188 under S_{DM} to 0.335 under Chain, while S_{AA} reaches 0.495. Of the 0.307 that pooling the information makes available, Chain realizes 0.147 ($\rho = 0.48$), leaving 52% unrealized. Excluding Llama-3-8B, ρ ranges from 0.27 (Phi-4) to 0.50 (Qwen3-32B). Chain falls short of S_{AA} in all six configurations, and the unadjusted intervals for Δ_{coord} exclude zero in five of them; the exception is Llama-3-8B (Table A7).

The size of the tax depends on the protocol. Star's coordination residual on the same track is much smaller. It is close to zero in five of six configurations ($+0.015$, -0.025, -0.010, 0.000, and $+0.030$), with Ministral-3-14B the exception (-0.160); the mean is -0.025, against -0.160 for Chain, and both positive estimates have intervals that include zero (Table A7). In Star, every role reports directly to the aggregator; in Chain, evidence must survive every intermediate message before it reaches the decision. The contrast suggests that the tax is set largely by how a protocol routes information, not only by how the information is distributed.

5.3 Where Is the Value Lost?

The bottleneck moves from delivery to integration. Applying the diagnostics of Section 3.3 to all scheduled runs, and reporting failure mass as a fraction of all runs rather than renormalizing within failures, we find that the bottleneck moves from delivery to integration. On Controlled Communication, most of the recorded failure mass (0.349) arises early: required evidence never

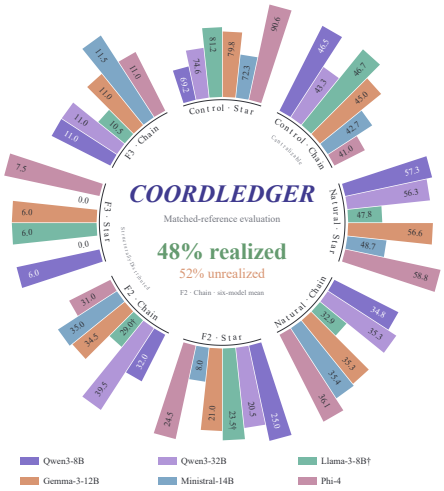

Figure 3: **Matched-reference evaluation across regimes, with the F2/Chain realization gap highlighted at the center.** Each spoke reports one model configuration for the protocol–track pair on the inner ring: Star and Chain against S_{CM} on the centralizable tracks, and against S_{DM} on F2 and F3. The center summarizes F2/Chain: communication realizes 0.147 of the available 0.307 ($\rho = 0.48$), leaving 52% unrealized as coordination tax.

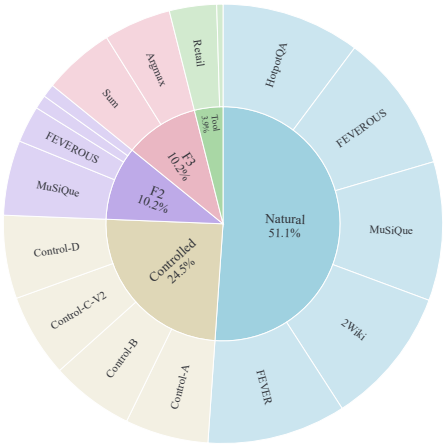

Figure 4: **Benchmark composition of CoordLedger.** The inner ring shows the five core tracks and their shares of the 1,957 instances; the outer ring shows the constituent datasets or task families. Natural Transfer contributes 1,000 instances (51.1%), Controlled Communication 480 (24.5%), F2 and F3 200 each (10.2%), and Stateful Tool Interaction 77 (3.9%).

enters the communication path (0.187) or is lost in transit (0.102). Natural Transfer (0.534) shifts toward integration. On the structurally distributed tracks, integration dominates, reaching 0.528 on F2 and 0.737 on F3. Decision-interface observability has no bar because it is tested with controlled visibility probes rather than inferred from traces.

Targeted interventions move the targeted stage. Each intervention keeps the instances and scoring rules and changes only how information is routed or shown. Evidence-preserving Star reduces delivery failures; carry-forward Chain reduces propagation loss to near zero; and evidence-visible aggregation shows that information held by a worker is not the same as information visible to the final decision. These interventions validate the stages and are not part of the core comparison.

5.4 Robustness

Post-core tests on the two Qwen anchor models show that no communication topology dominates across tasks. On Controlled Communication, Round-Table is the strongest evaluated multi-agent topology, reaching 0.794 and 0.910 for Qwen3-8B and Qwen3-32B, respectively, but remains below the matched S_{CM} references (0.840 and 0.954). On the Natural Transfer stress slice, Star is strongest at 0.590 for both models and is close to, or slightly above, S_{CM} (0.570 and 0.585). Communication capacity also shows diminishing returns: increasing B_{comm} from 128 to the core setting of 512 tokens improves Star substantially, whereas increasing it further to 2048 yields no reliable additional gain; paired 512-versus-2048 intervals include zero. These tests support task-dependent topology effects rather than a universal communication architecture or monotonic communication-scaling law. Full stress-test results are in Appendix J; the stateful tool track remains floor-limited (Appendix K).

6 Discussion

For evaluators: report the contract with the score. Multi-agent performance is interpretable only relative to the reference permitted or implied by the deployment access contract. Evaluations should therefore report the contract, the corresponding matched contrast, and traces sufficient for stage attribution; incomplete traces can identify an outcome gap without localizing its source (Figure 2).

For system builders: target the tax, not only the gain. Under distributed access, communication can create value while leaving much of the available opportunity unrealized. Our diagnostics suggest optimizing not only evidence sharing but also its preservation, routing, and visibility at the final decision interface; ρ can serve as an objective for protocol and topology search (Zhuge et al., 2024; Zhang et al., 2025a).

What the ledger identifies. The ledger separates two questions that are otherwise conflated. The S_{DM}-to-MA contrast measures value realized by communication under the deployment restriction, whereas the MA-to-S_{AA} contrast measures how much of the all-access opportunity remains unrealized. Neither contrast by itself is a causal estimate of a particular communication mechanism: the tax is defined relative to the matched S_{AA} reference and can contain both losses and gains at the instance level. Its role is therefore diagnostic and comparative—to determine whether an apparent multi-agent improvement survives the appropriate reference and, when it does, how much of the access opportunity remains unused. Mechanistic claims require the trace-based analyses and controlled interventions of Section 5.3.

The coordination tax has a human counterpart. Human groups with distributed information often fail to integrate uniquely held evidence (Stasser & Titus, 1985), and recent LLM studies report similar shortfalls (Li et al., 2026b). Our traces sharpen this observation: on distributed tracks, the dominant recorded bottleneck is integration after evidence becomes available.

For benchmark design: make the reference part of the specification. A benchmark that changes what information agents may observe without changing its single-agent reference risks attributing access advantages to coordination. COORDLEDGER instead treats the task, access contract, and matched reference as a single evaluation specification. This makes extensions easier to interpret: new topologies or communication policies may change how information moves, but they should not change the public task information, deployment restriction, scoring rule, frozen instance set, or matched inference opportunity. Under these invariants, a change in the reported ledger can be attributed to the evaluated communication design rather than to a shifting reference. The same principle can therefore support future multi-agent benchmarks with different architectures, access structures, and interaction protocols.

Limitations. Our conclusions are bounded by the evaluated open-weight model panel and the matching contract. The six configurations span several model families and serving formats but do not establish whether the coordination tax shrinks for substantially larger or frontier models. Moreover, matched references control provider-call opportunity and generation budget rather than exact realized token counts or FLOPs; a stricter token- or compute-capped reference would further sharpen the centralizable comparison. Stage attribution also depends on available trace annotations, and the stateful tool-use track remains floor-limited. The decomposition is operational rather than universally causal. Appendix M discusses these boundaries and additional scope limitations.

7 CONCLUSION

We introduced COORDLEDGER, a benchmark that reads a multi-agent score against the single-agent references its deployment permits. When one model may see all task-critical information, the reference is a compute-matched single model; when no single role may, the reference is a pair: a single agent under the same restriction and a single agent that sees everything.

The two regimes lead to different conclusions. On Controlled Communication, communication does not beat a matched single model with the same information: S_{CM} matches or exceeds Star in five of six configurations and Chain in all six. When access is structurally distributed, communication clearly helps, raising Chain from 0.188 to 0.335, yet it realizes less than half of the 0.307 that pooling the information makes available, and its failures concentrate at integration rather than delivery. A multi-agent win is therefore not self-explanatory. Before asking whether a team beats a single agent, one should ask which single agent its deployment allows. We will release the benchmark, code, and machine-readable results to make that question routine.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and editing, literature search and discovery, and selected aspects of research ideation and technical assistance. All AI-assisted content and research outputs were reviewed and verified by the authors, who take full responsibility for the final manuscript and accompanying artifacts.

REFERENCES

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982. Accepted as an oral presentation at ICML 2026.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4971–5012. PMLR, 2024. URL https://proceedings.mlr.press/v235/burns24b.html.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4082. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b1041e52d3be19f0a9bc491657488e4a-Abstract-Datasets_and_Benchmarks_Track.html.

Mengzhuo Chen, Junjie Wang, Fangwen Mu, Yawen Wang, Zhe Liu, Huanxiang Feng, and Qing Wang. Seeing the whole elephant: A benchmark for failure attribution in LLM-based multi-agent systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19888–19905, 2026. doi: 10.18653/v1/2026.acl-long.912. URL https://aclanthology.org/2026.acl-long.912/.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/578e65cdee35d00c708d4c64bce32971-Abstract-Conference.html.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3405. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Juergen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6507b115562bb0a305f1958ccc87355a-Abstract-Conference.html.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on*

Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 23662–23733. PMLR, 2024. URL https://proceedings.mlr.press/v235/khan24a.html.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems, 2025. URL https://arxiv.org/abs/2512.08296.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59f6421e64707225fdf5b28840679a07-Abstract-Conference.html.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2264. URL https://proceedings.neurips.cc/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Keyu Li, Junhao Shi, Yang Xiao, Mohan Jiang, Jie Sun, Yunze Wu, Dayuan Fu, Shijie Xia, Xiaojie Cai, Tianze Xu, Weiye Si, Wenjie Li, Dequan Wang, and Pengfei Liu. AgencyBench: Benchmarking the frontiers of autonomous agents in 1m-token real-world contexts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7422–7440. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.337. URL https://aclanthology.org/2026.acl-long.337/.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Systematic failures in collective reasoning under distributed information in multi-agent LLMs, 2026b.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

Aneesh Pappu, Batu El, Hancheng Cao, Carmelo di Nolfo, Yanchao Sun, Meng Cao, and James Zou. Multi-agent teams hold experts back, 2026.

Minjun Park, Donghyun Kim, Hyeonjong Ju, Seungwon Lim, Dongwook Choi, Taeyoon Kwon, Minju Kim, and Jinyoung Yeo. PAC-BENCH: Evaluating multi-agent collaboration under privacy constraints. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31030–31056, 2026. doi: 10.18653/v1/2026.findings-acl.1552. URL https://aclanthology.org/2026.findings-acl.1552/.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in LLM-based multi-agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12347–12361, 2025. doi: 10.18653/v1/2025.emnlp-main.623. URL https://aclanthology.org/2025.emnlp-main.623/.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45883–45905. PMLR, 2024. URL https://proceedings.mlr.press/v235/smit24a.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Garold Stasser and William Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6): 1467–1478, 1985. doi: 10.1037/0022-3514.48.6.1467.

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. doi: 10.18653/v1/2025.emnlp-main.249. URL https://aclanthology.org/2025.emnlp-main.249/.

Dat Tran and Douwe Kiela. Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets, 2026. URL https://arxiv.org/abs/2604.02460.

Haochun Wang, Sendong Zhao, Jingbo Wang, Zewen Qiang, Bing Qin, and Ting Liu. Beyond frameworks: Unpacking collaboration strategies in multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21361–21375, 2025. doi: 10.18653/v1/2025.acl-long.1037. URL https://aclanthology.org/2025.acl-long.1037/.

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. Reasoning in token economies: Budget-aware evaluation of LLM reasoning strategies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19916–19939, 2024a. doi: 10.18653/v1/2024.emnlp-main.1112. URL https://aclanthology.org/2024.emnlp-main.1112/.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6106–6131, 2024b. doi: 10.18653/v1/2024.acl-long.331. URL https://aclanthology.org/2024.acl-long.331/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://research.google/pubs/self-consistency-improves-chain-of-thought-reasoning-in-language-models/.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Zhiheng Xi, Dingwen Yang, Jiaqi Liu, Jixuan Huang, Honglin Guo, Baodai Huang, Tinggang Chen, Qi Zhang, Zhonghang Lu, Chenyu Liu, Jiajun Sun, Jiazheng Zhang, Dingwei Zhu, Xin Guo, Junzhe Wang, Zhihao Zhang, Yuming Yang, Junjie Ye, Minghe Gao, Dongrui Liu, Jiaming Ji, Guohao Li, Tao Gui, Qi Zhang, and Xuanjing Huang. AgentGym2: Benchmarking large language model agents in de-idealized real-world environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44451–44479. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2058. URL https://aclanthology.org/2026.acl-long.2058/.

Xin Yang, Junhao Wang, Bintao Tang, Xuxin Cheng, Cao Liu, Ke Zeng, and Wenyuan Jiang. When 20 agents fail to sort: The distributed sorting benchmark for scalable multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34002–34021, 2026. doi: 10.18653/v1/2026.findings-acl.1698. URL https://aclanthology.org/2026.findings-acl.1698/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5492ecbce4439401798dcd2c90be94cd-Abstract-Conference.html.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems, 2025b. URL https://proceedings.mlr.press/v267/zhang25cq.html.

Yuzhe Zhang, Feiran Liu, Yi Shan, Xinyi Huang, Xin Yang, Yueqi Zhu, Xuxin Cheng, Cao Liu, Ke Zeng, Terry Jingchen Zhang, and Wenyuan Jiang. SILO-BENCH: A scalable environment for evaluating distributed coordination in multi-agent LLM systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29379–29398, 2026. doi: 10.18653/v1/2026.acl-long.1354. URL https://aclanthology.org/2026.acl-long.1354/.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, 2025. doi: 10.18653/v1/2025.acl-long.421. URL https://aclanthology.org/2025.acl-long.421/.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62743–62767. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhuge24a.html.

Appendix

A Extended Comparison with Related Evaluation Paradigms

Table A1 provides the dimension-by-dimension comparison referenced in Section 2. The table is intended as an evaluation-design comparison rather than a ranking of prior benchmarks: a checkmark denotes an explicit primary design element, a filled circle denotes a related or partial component, and a cross denotes that the dimension is not a primary evaluation axis.

Table A1: **Comparison with closely related evaluation paradigms.** ✓ denotes an explicit primary design element, ● a related or partial component, and × that the dimension is not a primary evaluation axis. "MA" denotes multi-agent, and "All-access CF" denotes an explicit centralized counterfactual that removes the deployment access restriction under matched inference opportunity.

Work	MA	Dist. access	Comp. match	All-access CF	Failure diag.	Cost/topol.
AgentBench	×	×	×	×	●	×
τ-bench / τ^2-bench	●	●	×	×	●	×
Token Economies	●	×	✓	×	×	✓
MultiAgentBench	✓	●	×	×	●	✓
Collab-Overcooked	✓	●	×	×	✓	●
SILO-BENCH	✓	✓	×	×	●	✓
PAC-BENCH	✓	✓	×	×	●	●
TraceElephant	✓	●	×	×	✓	×
This work	✓	✓	✓	✓	✓	✓

B Benchmark Datasheet and Extended Composition

Main Table 1 summarizes the five core tracks. Table A2 extends that view with the two post-core stress-test slices and their run counts. Controlled instances are synthetic and structurally audited. Natural Transfer uses deterministic, pre-outcome selections from the cited public datasets. F2 and F3 are structurally filtered regimes in which the deployment contract distributes task-critical information or state. Stateful Tool Interaction is retained as a headroom evaluation. The topology and communication-budget studies are post-core stress tests and are excluded from the core scorecard.

Table A2: **Extended frozen benchmark composition.** The first five rows contain 1,957 unique core instances; stress-test sizes are reported as runs.

Track	Composition	Size	Evaluation role
Controlled Communication	Control-A/B/C-V2/D	480 instances	mechanism isolation
Natural Transfer	HotpotQA, FEVEROUS, MuSiQue, 2Wiki, FEVER	1,000 instances	external transfer
F2: Access-Constrained Integration	distributed evidence/access	200 instances	primary identification
F3: Distributed-State Observability	distributed state observations	200 instances	hard stress
Stateful Tool Interaction	retail and airline tasks	77 instances	state/tool headroom
Topology stress	P6/P7 on Control480 and Natural200	3,520 runs	post-core
Budget stress	$B_{\text{comm}} \in \{128, 512, 2048\}$ on Natural100	1,200 runs	post-core

The natural-data adapters preserve the original questions, answers, and source annotations. The frozen Natural1000 split contains 200 examples from each dataset family. Under the adopted support-mapping metric, source-grounded evidence coverage is undefined for FEVER and is therefore recorded as `N/A_NOT_DEFINED`, never as zero.

C Protocol Specifications and Information Boundaries

P1 is the Full-Information Single and receives the complete centralized public view permitted by the deployment contract. P2 uses shard-local agents without inter-agent messages and combines their outputs with a plurality aggregator. P3 and P4 are Star and Chain, respectively. P5 is the matched single-model reference: S_{CM} is used on centralizable suites, whereas S_{DM} preserves the deployment restriction on F2/F3. P6 (Hierarchy) and P7 (Round-Table) are post-core stress protocols. The all-access reference S_{AA} is a separate identification arm rather than a core protocol.

Table A3 summarizes the information, communication, and inference contracts for the primary arms. "Role scoped" means that no model call directly receives the union of information assigned to distinct roles. Matched references preserve the frozen provider-call opportunity and generation-budget contract; they do not assert equality of realized tokens or FLOPs.

Table A3: **Protocol and access contracts in CoordLedger.** The matched-reference rows implement the comparison contract defined in Section 3.2; protocol rows retain their native execution structure.

Arm	Direct information view	Comm.	Inference contract	Evaluation role
Full-Info Single	complete permitted centralized view	no	one-call	centralized sanity/reference arm
No Communication	role-scoped views	no	protocol-native	isolates value of message exchange
Star / Chain	role-scoped views	yes	protocol-native	communicating MA systems
S_{CM}	complete centralizable public view	no	matched	centralizable matched reference
S_{DM}	deployment-permitted role-scoped view	no	matched	distributed deployment reference
S_{AA}	union of public role views	no	matched	all-access centralized counterfactual

Every run separates evaluator-only fields from `PublicTaskView` and `AgentView`. Model prompts cannot access private gold answers, gold derivations, or required semantic-unit identifiers. Communication tokens count only inter-agent text that is actually transmitted. Provider input and output tokens, communication tokens, call count, latency, truncation, and parse status are logged separately.

D Frozen Prompt, Parser, and Access Contracts

The benchmark implementation freezes prompts, parsers, seeds, access contracts, and one-shot retry rules before outcome analysis. Across all protocols, evaluator-only fields remain outside the model-visible task view. The principal prompt-side distinction is therefore the information and message interface implied by the deployment contract rather than access to private labels. Table A4 summarizes the invariants shared by the primary comparison arms.

Table A4: **Prompt and access invariants across primary arms.** A checkmark means the property is part of the frozen execution contract.

Property	MA	S_{CM}	S_{DM}	S_{AA}
Frozen task instruction / output contract	✓	✓	✓	✓
Evaluator-only fields hidden	✓	✓	✓	✓
Role-scoped direct access	✓	×	✓	×
Union of public role views directly visible	×	✓*	×	✓
Inter-agent messages available	✓	×	×	×
Matched provider-call opportunity / generation budget	protocol	✓	✓	✓

*For centralizable suites, S_{CM} receives the complete centralized public view permitted by the deployment contract.

E Full Core Scorecards and Centralizable Protocols

This section provides absolute performance across the four non-tool core tracks and protocol-level results for the two centralizable regimes. These tables are audit-level complements to the matched contrasts emphasized in the main paper.

Table A5: **Six-model core scorecard (macro accuracy).** Stateful Tool Interaction is omitted here because it is floor-limited and is documented separately in Appendix K. The highest value within each track column is highlighted in indigo. Absolute cross-model values are descriptive; the paper's primary claims rely on within-configuration contrasts.

Model	Control	Natural	F2	F3
Qwen3-8B	0.613	0.472	0.296	0.091
Qwen3-32B	0.668	**0.490**	**0.328**	0.085
Llama-3-8B	0.630	0.406	0.112	0.033
Gemma-3-12B	0.665	0.462	0.277	0.097
Ministral-3-14B	0.648	0.474	0.300	0.090
Phi-4	**0.670**	0.475	0.309	**0.106**

Table A6: **Protocol accuracy on the two centralizable regimes.** S_{CM} is the Compute-Matched Single reference. For visual guidance, only the highest model-level value in each primary Star, Chain, and S_{CM} column is highlighted in indigo; Full and No Communication are reported without rank highlighting.

Model	Full	NoComm	Star	Chain	S_{CM}
Controlled Communication					
Qwen3-8B	0.833	0.235	0.692	0.465	0.840
Qwen3-32B	0.963	0.244	0.746	0.433	**0.954**
Llama-3-8B	0.850	0.221	0.812	**0.467**	0.798
Gemma-3-12B	0.929	0.246	0.798	0.450	0.900
Ministral-3-14B	0.925	0.260	0.723	0.427	0.906
Phi-4	0.946	0.181	**0.906**	0.410	0.908
Natural Transfer					
Qwen3-8B	0.557	0.326	0.573	0.348	0.556
Qwen3-32B	0.558	0.361	0.563	0.353	**0.613**
Llama-3-8B	0.441	0.346	0.478	0.329	0.435
Gemma-3-12B	0.521	0.351	0.566	0.353	0.520
Ministral-3-14B	0.584	0.364	0.487	0.354	0.581
Phi-4	0.532	0.356	**0.588**	**0.361**	0.539

F F2 Identification Details

Table A7: **Matched decomposition for Chain on Access-Constrained Integration (F2).** Intervals are unadjusted paired-bootstrap 95% intervals with $n = 200$ per model, and the identity $\Delta_{\text{deploy}} = \Delta_{\text{access}} + \Delta_{\text{coord}}$ holds for every row. Model-level values are shown without rank coloring; the Mean row highlights the aggregate realized value and coordination residual.

	Observed performance			Matched decomposition		
Model	$\mathbf{S}_{\text{DM}}$	**Chain**	$\mathbf{S}_{\text{AA}}$	Δ_{access}	Δ_{coord}	Δ_{deploy}
Qwen3-8B	0.215	0.320	0.500	$+0.285$ $[0.215, 0.360]$	-0.180 $[-0.250, -0.110]$	$+0.105$ $[0.045, 0.170]$
Qwen3-32B	0.235	0.395	0.555	$+0.320$ $[0.245, 0.390]$	-0.160 $[-0.240, -0.080]$	$+0.160$ $[0.100, 0.225]$
Llama-3-8B	0.000	0.290	0.330	$+0.330$ $[0.265, 0.395]$	-0.040 $[-0.105, 0.025]$	$+0.290$ $[0.230, 0.355]$
Gemma-3-12B	0.210	0.345	0.490	$+0.280$ $[0.210, 0.350]$	-0.145 $[-0.225, -0.065]$	$+0.135$ $[0.070, 0.200]$
Ministral-3-14B	0.230	0.350	0.580	$+0.350$ $[0.275, 0.425]$	-0.230 $[-0.305, -0.160]$	$+0.120$ $[0.045, 0.195]$
Phi-4	0.235	0.310	0.515	$+0.280$ $[0.200, 0.360]$	-0.205 $[-0.280, -0.130]$	$+0.075$ $[0.015, 0.135]$
Mean	–	–	–	$\mathbf{+0.307}$	$\mathbf{-0.160}$	$\mathbf{+0.147}$

The Llama-3-8B S_{DM} arm exhibits a severe format/runtime failure mode. The row is retained in the primary intent-to-run analysis. The common-evaluable analysis is treated only as a sensitivity check, consistent with Section 4 and Appendix L. For Star, the coordination-residual point estimates are +0.015 (Qwen3-8B), -0.025 (Qwen3-32B), -0.010 (Llama-3-8B), 0.000 (Gemma-3-12B), -0.160 (Ministral-3-14B), and +0.030 (Phi-4). The unadjusted 95% intervals for both positive point estimates include zero.

G Natural Transfer Breakdown

This section breaks the Natural Transfer comparison down by constituent dataset, reporting $\text{Star} - S_{\text{CM}}$ contrasts for each model configuration.

Table A8: **Dataset-level Natural Transfer contrasts.** Each entry reports $\text{Star} - S_{\text{CM}}$; positive values favor Star. Header abbreviations are Q3 (Qwen3), L3 (Llama-3), G3 (Gemma-3), and M3 (Ministral-3). For each model configuration, only the largest dataset-level contrast is highlighted in indigo.

Dataset	**Q3-8B**	**Q3-32B**	**L3-8B**	**G3-12B**	**M3-14B**	**Phi-4**
HotpotQA	+0.020	-0.070	+0.080	+0.040	-0.095	+0.075
FEVEROUS	+0.005	-0.025	-0.010	+0.005	-0.070	-0.015
MuSiQue	**+0.045**	-0.040	**+0.105**	**+0.130**	-0.190	**+0.140**
2WikiMultiHopQA	+0.005	-0.125	+0.045	+0.070	-0.125	+0.035
FEVER	+0.010	**+0.010**	-0.005	-0.015	**+0.010**	+0.010

H Cost and Communication Accounting

This section reports per-run provider-call, provider-token, and communication-token accounting for the Qwen3-8B exemplar. These quantities are reported separately because the matching contract controls provider-call opportunity and generation budget rather than exact realized tokens or FLOPs.

Table A9: **Per-run accounting for the Qwen3-8B exemplar.** Provider tokens are input/output totals rather than FLOPs and are not directly comparable across tokenizers. Communication tokens include only transmitted inter-agent text. These are descriptive costs, so no rank highlighting is applied; nonzero communication totals are bold and zeros are muted.

Protocol	Provider calls	Provider tokens	Communication tokens
Controlled Communication			
P1 Full	1	373	0.0
P2 NoComm	5	1356	0.0
P3 Star	5	1665	**102.2**
P4 Chain	4	1339	**90.8**
$\mathbf{S}_{\mathrm{CM}}$	5	2118	0.0
Natural Transfer			
P1 Full	1	1644	0.0
P2 NoComm	5	2412	0.0
P3 Star	5	3261	**371.8**
P4 Chain	4	2749	**352.8**
$\mathbf{S}_{\mathrm{CM}}$	5	8446	0.0

I Information-Flow Diagnostics and Mechanism Interventions

The main text summarizes the diagnostic findings in Section 5.3; we do not repeat a full visual summary here. Instead, this section records the interpretation contract used for the diagnostic interventions.

The controlled interventions preserve the benchmark instances and scoring rules while changing only information routing or visibility. Evidence-preserving Star increases final evidence coverage, carry-forward Chain removes most propagation loss, and evidence-visible aggregation shows that worker-held information is not equivalent to information available to the final decision component. These interventions support the delivery, propagation, decision-interface observability, and integration taxonomy but are excluded from the confirmatory core leaderboard.

A stage is reported only when the frozen trace supports that attribution. When message contents or other required trace fields are absent, the diagnosis remains *N/A* or is assigned only to a coarser supported category. Parser, format, timeout, and other runtime failures remain separate from semantic information-flow failures.

J Topology and Communication-Budget Stress Tests

This section reports the two post-core stress tests used in the main text: topology variation and communication-budget variation on the frozen Qwen anchor-model slices. These experiments are robustness analyses rather than part of the primary matched identification design.

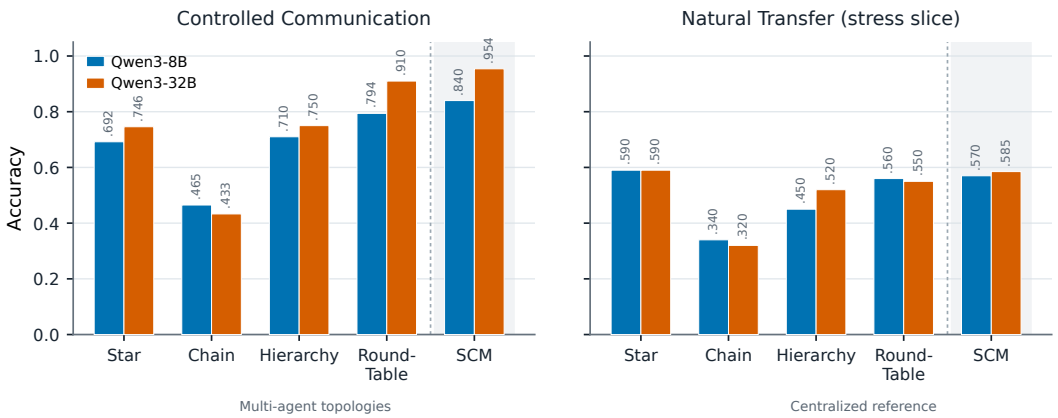

Figure A1: **Post-core topology stress on two Qwen anchor models.** Round-Table is the strongest evaluated multi-agent protocol on Controlled Communication, whereas Star is strongest on the frozen Natural Transfer slice. These results do not imply a universal topology ranking.

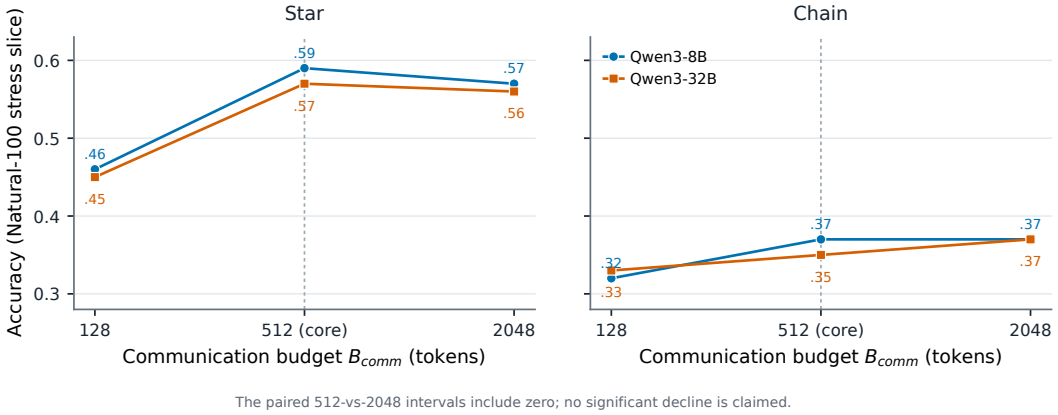

Figure A2: **Post-core communication-budget stress.** Star improves from 128 to 512 transmitted tokens and then saturates or declines slightly in point estimate; Chain is closer to a plateau or a small increase. The prespecified monotonic-improvement hypothesis is not supported, but the paired intervals for 512 versus 2048 tokens include zero.

K Stateful Tool Interaction

The stateful track contains 77 structurally eligible airline and retail tasks. Their required gold actions span at least two exclusive role domains and include a write operation. The environment executes real tool functions against resettable state and scores the objective final database state. All six core model configurations achieve zero exact final-state success across the frozen tool–protocol set. We therefore label this track Floor-Limited/Not-Identifiable-Due-To-Floor; a zero-versus-zero comparison does not establish collaboration neutrality. Unauthorized tool-access attempts are rejected before execution.

L Evaluation Sensitivity and Forensic Corrections

The primary intent-to-run analysis retains every scheduled outcome. The common-evaluable sensitivity analysis excludes an instance only when either arm of a paired contrast has a format/runtime failure; it does not replace the prespecified intent-to-run result. The most consequential caveat concerns Llama-3-8B on F2/F3, whose S_{DM} arm has a severe format-failure mode.

A reporting-layer forensic audit corrected a legacy `Necessity` macro key-collision bug, marked FEVER coverage as not applicable, relabeled stateful zero-versus-zero comparisons as floor-limited,

and resolved semantic-migration provenance to 34 runs across 8 cells. These corrections affect reporting metadata and aggregation only. The raw model outputs, parser semantics, evaluator behavior, protocol executions, and frozen datasets are unchanged.

M Limitations and Scope

Our conclusions are bounded by the evaluated models and the matching contract. The study covers six open-weight configurations from five model families with different serving formats and quantization choices. Absolute cross-family scores should not be interpreted as model rankings; the principal claims rely on paired contrasts within each frozen configuration. The matched references control provider-call opportunity and generation budget, not exact realized token counts or underlying FLOPs.

The decomposition is operational rather than universally causal. The S_{AA} arm is a benchmark-defined counterfactual that removes the deployment access restriction under the matched inference contract; it is neither an oracle nor a model of every possible centralized system. Access-Constrained Integration and Distributed-State Observability cover two forms of distributed information and state, but not the full space of structurally distributed deployments. Natural Transfer uses public datasets that may have appeared in model pretraining corpora and should therefore be interpreted as an external-transfer evaluation rather than an uncontaminated estimate of generalization.

The information-flow diagnostics require structural annotations of task-critical information and its expected path. They are most precise in the controlled suites; for less structured regimes, we report only categories supported by the frozen representation. Stateful Tool Interaction is floor-limited: zero objective final-state success across all core configurations prevents the suite from distinguishing collaboration advantage from neutrality.

Finally, the topology and communication-budget studies use frozen slices and two Qwen anchor models rather than the full panel. They establish task dependence within the evaluated range, not universal topology rankings, communication scaling laws, or conclusions about larger agent populations. Future work should extend the matched design to broader model families, richer stateful environments, additional access structures, heterogeneous agents, and larger-scale systems.

N Reproducibility and Release Manifest

The frozen release records dataset hashes, parser-contract hashes, random seeds, communication budgets, provider-call counts, provider input/output token accounting, one-shot retry rules, and run completeness. The core contains 58,248 scheduled runs with no missing cells or duplicate run identifiers. Post-core identification and stress-test runs are stored under separate manifests and namespaces. All paper-facing quantities are generated from the corrected frozen reporting layer and checked for consistency across tables, figures, and machine-readable artifacts.