Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Naming the channel increases evidence integration in language model teams

Christian Jörg Rustenbach

christian.rustenbach@med.uni-tuebingen.de

University of Tübingen

Research Article

Keywords: Multi-agent systems, Agent communication, Large language models, Information integration, Team coordination

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11676035/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Agents can collectively hold every fact needed for a decision while no individual has the evidence to justify it. We tested whether naming the location of new activity changes how often language model teams assemble distributed evidence after the primary communication channel closes. Of 163 newly started five-agent teams, 158 qualified and 112 spontaneously wrote to a backup channel. At each first write, we copied the team state into four isolated continuations and randomly assigned the conditions to execution slots. The four other agents received no notice, neutral metadata, a generic activity notice or the same notice naming the channel. The prespecified endpoint was the presence of all three authenticated decision rules in a single actual model input within 60 turns of channel closure. Complete integration occurred in 2, 2, 17 and 112 of the same 112 teams, respectively. Naming the channel increased integration by 84.8 percentage points over generic activity (simultaneous 95% confidence interval 69.1–92.5). All 95 discordant pairs favored naming. Under generic activity, 97 teams read another agent's task message and 90 received a core rule from another agent. Sixteen submitted a valid solution, compared with 112 under named activity. In this system, linking an event to a channel already visible in the tools changed whether distributed evidence reached one agent and supported a valid decision.

Introduction

One agent knows the general rule, a second knows the exception, and a third knows whether a directive overrides them both. A team can collectively possess the answer while no individual can justify it. The distinction has a long experimental history in human groups [1]. It becomes directly measurable in language model teams, where the information supplied to each agent can be inspected turn by turn.

Agent communication and distributed evidence

Using HiddenBench, a set of 65 decision tasks, Li et al. [2] found that teams given different pieces of information reached the correct decision less often than individual agents given all the facts. Other experiments varied communication mode or the distribution of information [3]. Analyses of failure traces identified problems with alignment and task verification [4]. In shared workspaces, knowledge of an action and knowledge of its location can support different aspects of coordination [5]. In agent teams, an agreement detector can change how a debate proceeds [6], and an acknowledgment protocol can shape cooperation between reinforcement-learning agents [7]. Here we vary the information in a notice about an event that all experimental conditions share.

T226 compared enabling the entire notification policy with pull-only communication, using the qualified team before a write as the starting point [8]. T227 recruited new teams and randomized after a spontaneous write, holding the event and all five histories fixed. No earlier trajectory entered the analysis. The key prediction was that identifying where activity occurred would increase integration

beyond a generic alert. Of 112 matched team states, 95 reached the endpoint only in the named member of this comparison and none only in the generic member.

Experimental design

Each team consisted of five instances of the same language model with separate interaction histories, identical instructions and equal tool access. A team first had to solve two distinct tasks using another agent’s verified information from the primary channel, board, within 40 agent turns, without prior use of the alternate channel. The host then disabled board. The read and write tools listed both channel names. The backup, annotations, remained available to every agent. Until the first successful write there was no activity notice. The study therefore distinguishes discovering a place to write from recruiting teammates to use it. Figure 1 shows recruitment and assignment.

Methods

Task and information environment

Every task required a decision about a synthetic record. Each agent held one private source. Three agents held the base rule, an exception and a directive, respectively. Two held distractors. For example, the base rule could determine the default disposition while the date and resource specified in an exception changed which rule applied. Even an inapplicable rule had to be consulted and cited to support the answer. A valid submission specified the disposition, governing rule, active rules, required obligations and the evidence linking each rule to its author. No notice contained a record, a solution or an instruction to read or write.

Before qualification, the system drew one of four decision classes with equal probability (allow, block, hold or review), then sampled a POST task with replacement within that class from a bank of 80. The order of turns and distribution of sources were shared across continuations of the same team. All agents ran gpt-6.1-sol with medium reasoning and Flex service. The tools permitted reading and writing messages, submitting a solution or waiting. The scheduler gave each agent one turn per five-turn round. No agent had a scripted coordination role.

Randomization and interventions

Randomization and interventions
A first successful write to annotations qualified regardless of content or task identifier (Table 1). The host committed the write and saved the signed shared state, including all five private histories. A concealed, uniformly drawn permutation assigned each of the four notices to one of four serial execution slots (24 possible orders). Model calls after the split were fresh. Continuations had isolated histories and could not read one another’s messages. Assignment, timing and recipient identity were verified from the recorded inputs.

Table 1
Notices as delivered. The strings below are the frozen UTF-8 JSON payloads. Each was inserted as a user message once per eligible agent. The original writer received none

N · None	**No additional message**
S · Neutral	{"session_format":{"encoding":"utf8_text","format":"standard"}}
G · Generic	{"channel_status":{"activity":"new_activity","channel":"unspecified"}}
K · Named	{"channel_status":{"activity":"new_activity","channel":"annotations"}}

Endpoints and observation time

Complete integration (Q3) required all three core rules for the current task to appear together in the input to a single successful, settled model call. The agent's own rule was authenticated against its private record. Each rule obtained from another agent required a prior read, a matching authored write and a receipt issued to that reader. References in generated text alone could not satisfy the endpoint. The separate task endpoint was a submission accepted by the host's deterministic validator. W6 and W12 denote 30 and 60 individual turns after primary-channel closure. All continuations ran to W12, including those that succeeded earlier.

ChatGPT (OpenAI) assisted with study planning, implement and debug software and assist with statistical interpretation. Matplotlib rendered the figures from the protocol and recorded outcomes. Only the designated study model produced the experimental agents' actions. The author made the scientific decisions and takes responsibility for the work..

Analysis and study conduct

Estimand and statistical plan

The unit of inference was the newly recruited team at its first spontaneous write. For each adjacent pair of conditions we estimated the difference in the probability of Q3 by W12 among qualified writing teams. Each team contributed a binary outcome to every condition and one paired difference to each contrast. The estimand averages over sampled tasks, actual write times and fresh model responses. The denominator for each comparison was 112 writing teams.

The privately archived analysis plan specified neutral versus none (S − N), generic versus neutral (G − S), and named versus generic (K − G). We used two-sided exact McNemar tests, with $p = 1$ for zero discordance, and Holm adjustment over all three contrasts at familywise $\alpha = .05$ [9]. Risk differences equal the excess of favorable over reverse discordant pairs divided by 112. Conservative intervals combine exact binomial bounds for the two discordant cell probabilities, with multiplicity protection across six cells for simultaneous 95% coverage under independent, identically distributed parents. Under

the sharp null, labels for the two compared slots are exchangeable within each isolated team. The prespecified K − N contrast was descriptive.

The target of 112 teams gave 82.5% marginal power for a 20-percentage-point difference at 40% discordance and the most stringent Holm threshold (.05/3). Power under the same assumptions was 19.6% for a 10-point difference. The plan restricted confirmatory analyses to a full cohort of 112. Additional prespecified analyses considered task clustering, equal weighting of decision classes, a common 20-turn window after the write and conservative Hoeffding bounds allowing heterogeneous independent parent differences. Intermediate endpoints and their timing were descriptive.

Recruitment, delivery and data integrity

From 6 to 7 November 2026, 158 of 163 fresh teams qualified. Of those, 112 wrote to annotations within 60 turns (70.9%, exact 95% interval 63.1–77.8%). The other 46 completed the observation period without a write. All qualifying writes occurred by turn 25 (range 11–25, median 17). The 112 teams spanned 57 task identifiers, 25 distinct operative rule configurations and all four decision classes. Some first writes concerned an earlier task (33/112), and 49/112 carried no current-task core rule.

All 448 continuations reached turn 60. In each active message condition, all four eligible recipients per team received the assigned notice. At matched first-recipient positions, both neutral and generic notices cost one more input token than the named notice. A transport amendment (A01) during recruitment allowed a fourth identical dispatch after three documented, cost-free capacity refusals and a wait of at least 15 minutes. Five calls used this rule and all succeeded. Follow-up, outcomes, message bytes and analysis rules were preserved. All 27,782 dispatches were reconciled, including 65 cost-free capacity refusals. Endpoint extraction followed technical closure and verification of the sealed raw-data copy. No interim arm-effect analysis was performed.

Results

The location field accounted for the largest paired difference

By turn 60, complete integration occurred under no notice in 2/112 teams (1.8%), neutral metadata in 2/112 (1.8%), generic activity in 17/112 (15.2%) and named activity in 112/112 (100%). Figure 2 shows the actual combinations across teams. The same two teams succeeded under all four conditions. Fifteen more succeeded under generic and named activity, and the remaining 95 succeeded only when the channel was named. The exact 95% interval for the named-arm proportion was 96.8% to 100%.

Table 2
Prespecified paired comparisons. For A − B, the two counts are teams succeeding only under A or only under B. Differences and simultaneous intervals are in percentage points (pp)

Comparison	Only A	Only B	Difference [95%]	Exact p	Holm p
S − N	0	0	0.0 [− 4.8, 4.8]	1	1
G − S	15	0	13.4 [1.5, 23.9]	6.10×10^{-5}	1.22×10^{-4}
K − G	95	0	84.8 [69.1, 92.5]	5.05×10^{-29}	1.51×10^{-28}

Table 2 gives all three prespecified tests. For named versus generic activity, all 17 generic successes also succeeded with the channel name. For generic versus neutral, all 15 discordances favored generic activity. The descriptive overall difference between named and absent notices was 98.2 points (marginal 95% interval 89.1–99.9). The neutral comparison estimates the effect of this particular neutral message. Identical observed outcomes give no general equivalence result.

Reading, integration and task completion

Under generic activity, 97 teams read another agent's message for the current task and 90 received at least one core rule from another agent in a model input. Only 17 assembled all three rules in one input, and 16 submitted a valid answer. Under named activity, all 112 teams reached each of these endpoints (Fig. 3). Under no notice and neutral metadata, respectively, 43 and 40 teams read another agent's current-task message, 40 and 35 received a core rule from another agent, and two in each condition submitted a valid answer.

Time and task sensitivity

By turn 30, complete integration had occurred in only 13/112 named and 4/112 generic continuations (one team each under neutral and absent notices). Among named continuations, the median time from the write to the first read of another agent's message was one turn, and the median time to complete integration was 17 turns. The corresponding medians under generic activity were seven turns among 97 readers and 20 among 17 integrators. These medians describe different sets of event-positive teams.

All 112 teams had at least 20 turns remaining when they wrote. Within that common post-write window, the named advantage over generic activity was 73.2 points and the generic advantage over neutral was 6.25 points. For the full horizon, task-cluster bootstrap intervals (10,000 replicates, seed 227, stratified by decision class) were 78.0–91.0 and 7.4–20.2 points, respectively. Equal class weighting gave differences of 85.8 and 12.6 points. The conservative simultaneous Hoeffding bounds remained positive for the named contrast (55.6–100 points) and spanned zero for the smaller generic contrast (− 15.8 to 42.6 points).

Discussion

Once one member of a five-agent team had written to a backup channel, a one-time notice naming that channel changed which teams brought all three decision rules together. In 95 of 112 starting states, integration succeeded in the named continuation and failed in its generic counterpart. The notice contained no rule, granted no permission and supplied no extra turn. Every recipient already had the channel name in the tool definitions. What the named notice supplied was the link between that familiar address and the new activity. Valid, source-supported decisions followed in all 112 named continuations.

What did the generic message already achieve?

A simple explanation would be that the generic message left most agents unable to find the channel. The traces place a limit on that account at the team level. Of the 95 teams that integrated only under named versus generic activity, 73 had already received another agent's core rule somewhere in the generic continuation. Seven more read another agent's task message without receiving a core rule, and 15 recorded no such read. Thus, for most of these paired failures, at least one exchange of relevant evidence had occurred. The remaining problem was getting the full set of records together in one agent's context.

Several explanations fit. Naming the same address for four recipients could lead more source holders to contribute, prompt repeated reads after further writes, or coordinate publication and retrieval on the current task. A recognized literal could also make a correctly formed tool call easier. The experiment manipulated the actual message field. It did not independently vary these later actions, the salience of the word, or its fit to the tool schema. The 73-team breakdown conditions on realized outcomes and is exploratory. It supports hypotheses about subsequent coordination without identifying a causal mediator or an agent's mental state.

How this extends work on distributed information

In hidden-profile studies, useful evidence can remain private even when a group has ample opportunity to discuss it [1, 2]. Agent-system evaluations have examined communication mode, protocol structure and failure traces [2–4]. Here the task-relevant evidence, model and tool permissions were held fixed across four continuations. Changing what teammates learned about a single event produced two experimentally distinguishable increments. Signaling activity increased integration. Identifying its location yielded a considerably larger increase. Early shared-editor research described how cues to others' actions support coordination [10]. Blackboard architectures likewise placed a common workspace at the center of cooperation among otherwise separate knowledge sources [11]. Here we can follow a location cue from delivery to the arrival of authenticated records in an agent's input, then check the ensuing decision.

A practical implication follows for evaluation. A dashboard that counted any read of a teammate's message would record 97 successes under generic activity. A dashboard that required all necessary sources in one actual input would record 17. Both describe real events. Only the latter directly answers whether an agent had all the evidence available to support its decision. Checking accepted submissions

adds a further distinction, visible in the one generic continuation that assembled the records without producing a valid answer.

Scope and reproducibility

Limits of the claim

Selection preceded randomization. The effect concerns the 112 teams that first qualified and then spontaneously used the backup channel. It gives no estimate of what a notice would do for the 46 qualified teams that never wrote. Likewise, all four continuations saw the same realized write, while fresh model responses introduced stochastic variation after the split. The joint-outcome display describes these realized continuations. It does not classify a team’s permanent ability or guarantee future success under the named notice.

The study used one model configuration, five agents with the same instructions, a finite synthetic rule bank, one practical backup channel and one pair of activity strings. With both channel names visible in the tool definitions, even “unspecified” could point indirectly to the single available alternative. The one-token difference, the literal channel identifier and the rest of the tool interface travel with the implemented contrast. A different channel name, multiple competing addresses, independently trained models or a shorter time budget could change the result. Q3 measures verifiable co-presence of evidence in model input. Functional task validity is a separate secondary measure. Neither outcome directly measures a model’s understanding.

Data, code and study records

The T227 protocol and statistical analysis plan were archived privately before the first main-study request. The live-study manifest records prospective approval for this new design. Two inherited fields retained the values scientific_claims_allowed=false and prospectively_frozen_replication=false. The historical template was for deterministic test runs. The separately timestamped T227 plan authorized live execution and specified the analysis. The study is a new prospective experiment. The earlier T226 dataset is cited separately [8].

The portable analysis starts with extracted team outcomes. On the original host, signature checks and receipt validation linked model inputs to recorded team states and the sealed, 404,672,512-byte event database. Offline reanalysis of the exported endpoints reproduces the prespecified contrasts and task-cluster bootstrap. Reconstructing every model input from the event stream requires the protected original environment.

Conclusion

For a team that had already opened a route for communication, knowing that someone had acted and knowing where they had acted led to markedly different outcomes. In this experiment, replacing “unspecified” with “annotations” in a single notice to each teammate raised complete integration from

17 to 112 of the same 112 teams. The experimentally established effect concerns that change in the notice. The process by which it organizes later contributions and reads remains open.

Declarations

Competing interests

The author declares no competing interests.

Ethics and consent

All tasks and records were synthetic. The study involved software agents and included no human participants, animals or identifiable personal data. Participant consent and consent for publication of participant data do not apply.

Funding

The author personally funded all study costs. The research received no external funding.

Author Contribution

C.J.R. conceived the study, approved the protocol and statistical analysis plan, supervised the conduct of the study, and reviewed the interpretation of the results. As the sole author, C.J.R. takes responsibility for the manuscript.

Data Availability

The T227 recruitment data, arm-level endpoints, first-event times, figure source data and study-design extract are available from Zenodo at https://doi.org/10.5281/zenodo.24079472. The dataset covers 163 new starts and four matched continuations for each of 112 writer teams. The corresponding author holds the underlying event database and private interaction histories separately.

Code availability

The T227 Zenodo record [12] includes the offline Python scripts and reference outputs for the primary paired comparisons and the stratified task-cluster bootstrap. To reproduce the published analysis, extract the archive and run python3 verify.py.

References

1. Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*, 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467
2. Li, Y., Naito, A., & Shirado, H. (2026). Systematic failures in collective reasoning under distributed information in multi-agent LLMs. *Proceedings of the 43rd International Conference on Machine Learning*, PMLR 306, 69250–69268. https://proceedings.mlr.press/v306/li26ej.html
3. Toxtli, C., & Delaflor, M. (2026). Hidden profile decision making in multi-agent LLM groups. *Companion of the Computer-Supported Cooperative Work and Social Computing*. ACM. https://doi.org/10.1145/3785651.3831525
4. Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., et al. (2025). Why do multi-agent LLM systems fail? *Advances in Neural Information Processing Systems*, 38. https://doi.org/10.52202/085713-4082
5. Gutwin, C., & Greenberg, S. (2002). A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work*, *11*, 411–446. https://doi.org/10.1023/A:1021271517844
6. Heller, S., Ibrahim, M., Selby, D. A., & Vollmer, S. (2026). Finding common ground: Using large language models to detect agreement in multi-agent decision conferences. *Autonomous Agents and Multi-Agent Systems*, 40, 48. https://doi.org/10.1007/s10458-026-09772-6
7. Phan, T., Sommer, F., Ritz, F., Altmann, P., Nüßlein, J., Kölle, M., Belzner, L., & Linnhoff-Popien, C. (2024). Emergent cooperation from mutual acknowledgment exchange in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, *38*, 34. https://doi.org/10.1007/s10458-024-09666-5
8. Rustenbach, C. J. (2026). ING-01 T226: Paired agent-team outcomes after communication disruption (version 1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.23986375.
9. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70. https://www.jstor.org/stable/4615733
10. Dourish, P., & Bellotti, V. (1992). Awareness and coordination in shared workspaces. *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work*, 107–114. https://doi.org/10.1145/143457.143468
11. Erman, L. D., Hayes-Roth, F., Lesser, V. R., & Reddy, D. R. (1980). The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys*, *12*, 213–253. https://doi.org/10.1145/356810.356816
12. Rustenbach, C. J. (2026). Naming the channel increases evidence integration in language model teams (ING-01 T227): paired endpoint data and analysis code (Version 1.0.0) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.24079472

Figures

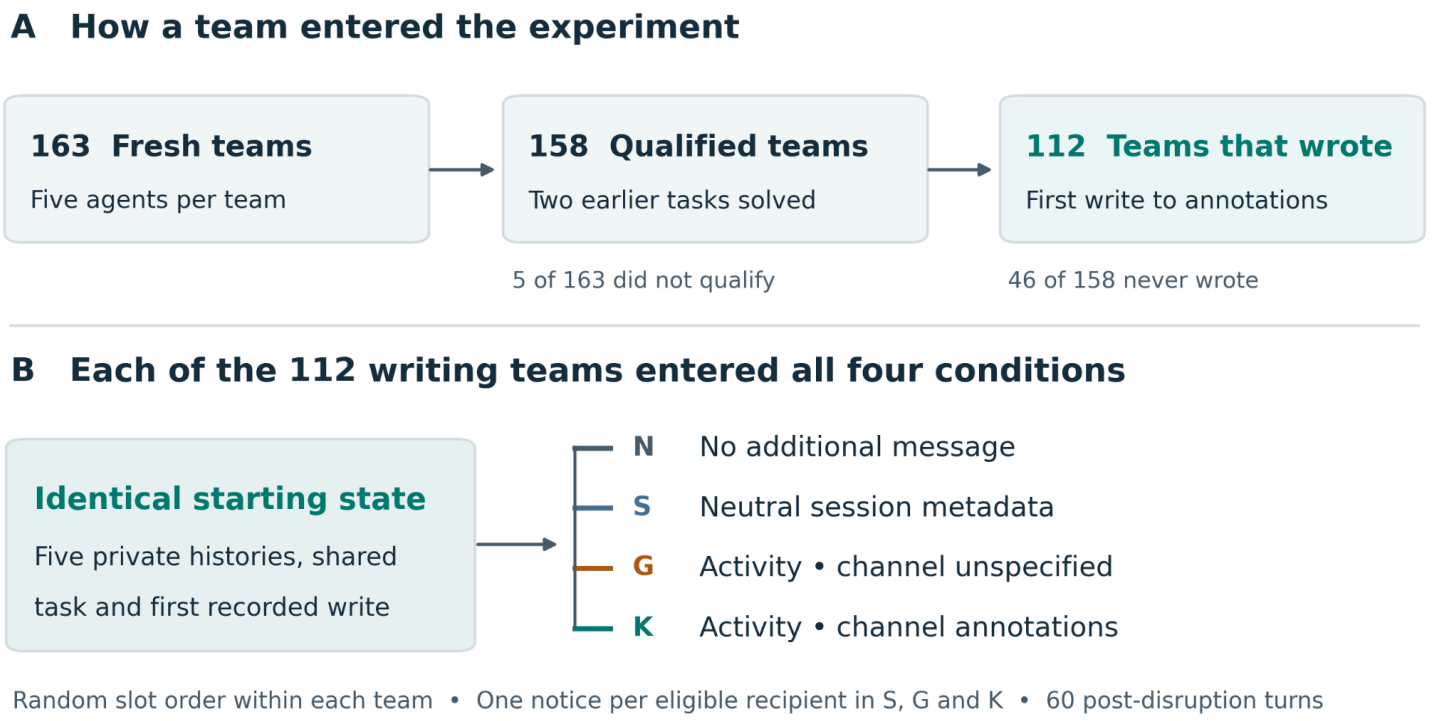


Figure 1

One first write, four possible notices. All conditions began with the same signed team state. In each message condition the four agents other than the writer received the notice on their next scheduled turn. The 60-turn observation window began at primary-channel closure and included the common history before the write

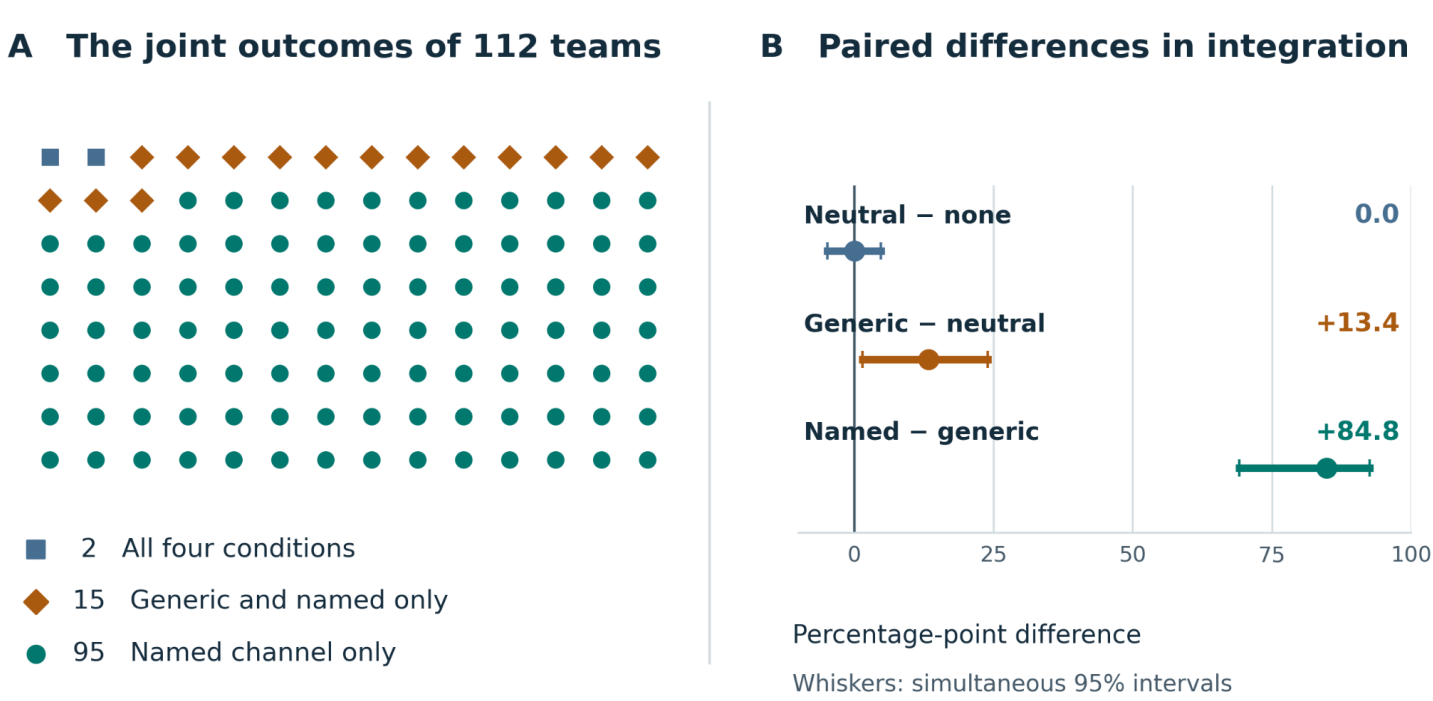


Figure 2

Complete evidence integration by turn 60. A. Each symbol represents one writing team, classified by which of its four continuations reached Q3. Ordering is for display. B. Points show paired risk differences. Intervals protect the family of three prespecified comparisons. Neutral and absent notices produced the same observed outcomes. No equivalence margin was specified

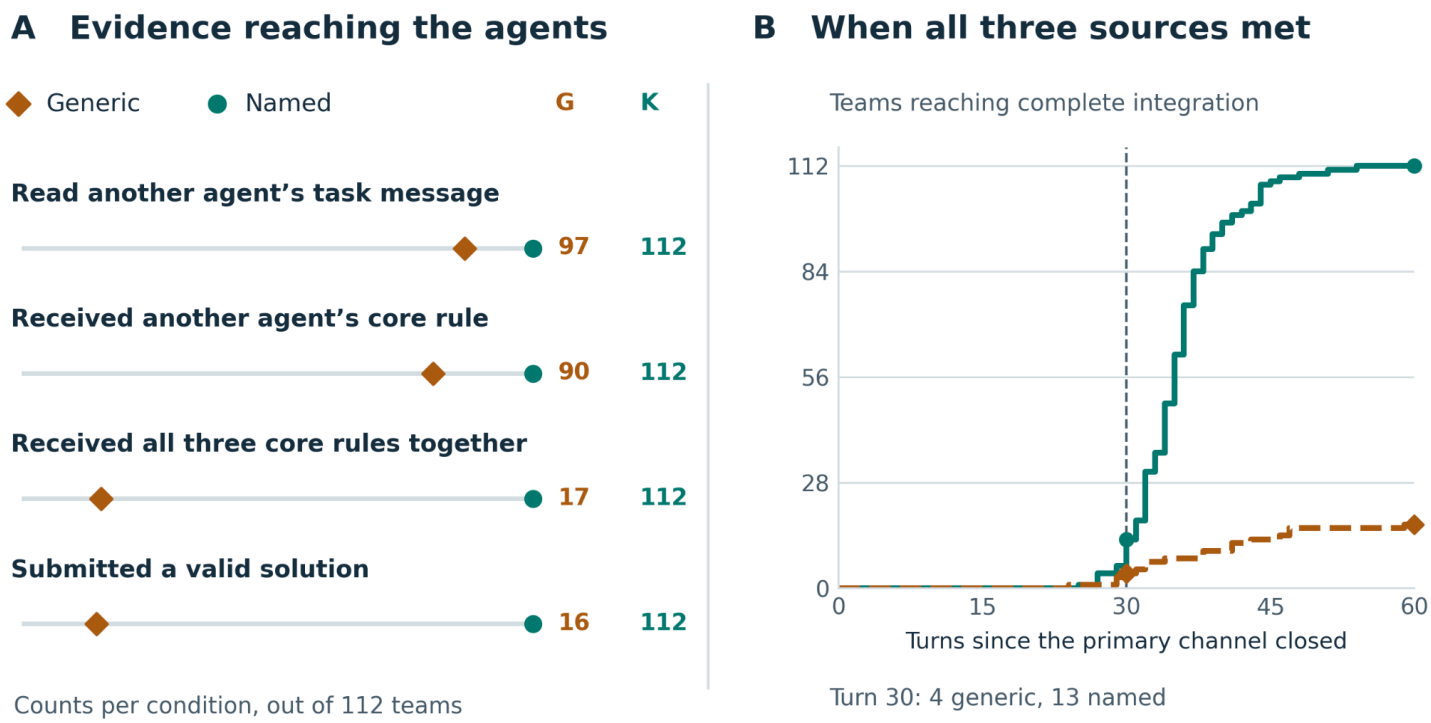


Figure 3

Reading a teammate's message and assembling the evidence were distinct events. A. Symbols show team counts under generic and named activity. The paired numbers give generic and named counts, respectively. A source could reach any agent in the team. B. Cumulative number of teams with at least one input containing all three current-task rules, reconstructed from recorded first-event turns. The dashed vertical line marks the prespecified midpoint. Both curves retain all 112 teams in the denominator