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.

Event-triggered activity notifications improve collective information recovery after communication disruption in large language model teams

Christian Jörg Rustenbach

christian.rustenbach@med.uni-tuebingen.de

Eberhard Karls Universität Tübingen

Research Article

Keywords: Multi-agent systems, Large language models, Communication disruption, Information pooling, Paired experiments

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11640109/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

Background.

A team can retain an alternative communication channel after its primary channel fails, yet evidence posted there may never reach a teammate. This study tested whether notifying agents of activity on the alternative improves collective information recovery.

Methods.

Eighty qualified five-agent teams each supplied four isolated branches. In the randomized paired comparison, both branches lost the primary channel and retained identical rights to read and write on the alternative. One branch relied on agent-initiated checks (B0). In the other (B+), the first successful write also triggered an activity notification carrying no task evidence for each teammate. The primary endpoint required all three authenticated task-critical sources together in at least one submitted model input by cycle 12.

Results.

The endpoint occurred in 2/80 teams (2.50%) under B0 and 61/80 (76.25%) under B+. All 59 discordant pairs favored B+. The paired risk difference was + 73.75 percentage points (conservative ≥ 95% confidence interval + 55.82 to + 84.05, exact two-sided McNemar $p = 3.46945 \times 10^{-18}$). First writes occurred in 60 and 61 teams, respectively. A teammate's authenticated core source reached a model input in 20 and 61 teams.

Conclusions.

Enabling the conditional notification rule increased complete information recovery in this task family and model configuration. The arms diverged mainly after the first write. The observed stages leave the contributions of attention, channel discovery and timing unresolved.

Introduction

Communication can fail even when messages arrive intact. A team may have an alternative route, a member may post the missing evidence there, and the team may still fail because nobody who needs that evidence reads it in time. The distinction is familiar from hidden-profile experiments, in which a group can hold enough information collectively yet fail to pool it [1]. Cooperative-work systems have long used awareness of colleagues' activity to help people decide where and when to coordinate [2]. An activity notification is especially informative experimentally because it may change access to evidence without supplying the evidence itself.

In large language model (LLM) teams, HiddenBench showed that a structured protocol can improve reasoning when information is distributed [3]. Toxtli and Delaflor found that agents could share private information yet still reach the wrong group decision [4]. The Multi-Agent System Failure Taxonomy (MAST) classified failures in multi-agent traces [5], and MAS-FIRE measured how injected faults affected reliability (Jia et al., arXiv:2602.19843v1, preprint). Zhou et al. showed how an unaffected path can preserve evidence after a communication fault (arXiv:2610.15459v1, preprint). Multi-Order Communication (MOC) changes the evidence passed between agents across multiple hops [6]. These findings leave a narrower question open. When an alternative channel already exists, is a signal that someone has used it sufficient to improve recovery of the evidence posted there?

The hypothesis arose from a sequence of method-development studies (Online Resource 1, Table S1). In an early, technically censored pilot with a simpler evidence-assembly task, all four observed fallback branches recovered. In two subsequent completed four-team pilots with semantic decisions, no fallback branch solved its task by the specified deadline. Some teams never used the fallback; others posted or transferred only a subset of what was needed. Extending observation to 12 cycles in the second pilot exposed further reading without further publication of core sources. These observations suggested notifying teammates after someone had already written. An exploratory comparison yielded 3/6 successes with the notification and 0/6 without it. A subsequent extension, authorized after those results were known, yielded 2/2 and 0/2. The combined historical finding motivated an independent prospective test. These developmental studies were documented in internal project reports.

I tested this question in 80 independent, qualified teams. From the same team state, I compared two branches with identical channel access and tool permissions. Only one enabled a notification after the first fallback write. I authenticated the sources present in each agent's submitted input, distinguishing a write from receipt of another agent's evidence, and joint access to all three sources from a correct solution. The sample size, primary contrast, cycle-12 endpoint, test and replication criteria were fixed before T226. By cycle 12, 2/80 teams had all three sources in one model input under agent-initiated checks, compared with 61/80 under the activity rule.

Methods

Independent teams and qualification

The confirmatory T226 experiment was conducted on 3–4 November 2026 using the custom-developed ING-01 experimental runtime. The host kept the task oracle and validation records separate from the five private agent contexts, enforced tool permissions, and authenticated events used to validate the endpoints. A parent was one independently initiated five-agent team. All agents used the same model alias (gpt-6.1-sol), medium reasoning effort, default service tier and cooperative instructions. Each had one turn per group cycle and could write or read a named task channel, submit a decision with evidence or wait. Source ownership varied by task. No participant was assigned a leadership or routing role.

Before intervention (PRE), teams had up to eight cycles to solve two distinct tasks using authenticated information from another agent via the board. Previous access to the alternative channel disqualified the parent. Empirical qualification (E0) additionally required live-model execution, a neutral reviewed manifest and valid trace and semantic checks. I retained all PRE starts in the flow. I stopped recruitment at 80 qualifications or 130 starts, subject to the prespecified technical, time and financial stops. A qualified parent with a later failure would retain its place in the analysis. Supplementary Methods give the task contract, stopping rules and prospective power assumptions.

Task and experimental conditions

The decision task distributed five private records, one per agent. Three carried the base rule, exception and directive. Two were distractors. The public query specified a resource and date. The correct response comprised a disposition (ALLOW, HOLD, REVIEW or BLOCK), the winning and active rules, a set of obligations and the authenticated authorship of all three core sources. Even an inapplicable rule had to be cited as present. Guessing the disposition or reproducing another agent's description could not supply the required provenance.

At E0 the host took a common snapshot of the team state, including private histories. Four isolated copies then performed the same new POST task, with the same actor order and model contract (Fig. 1). A retained both channels. B0 denied access to the board and left the alternative, named annotations, available through explicit reads and writes. B+ used the same permissions and added the activity rule. C denied access to both channels. A valid POST success or complete source set under C would have stopped the experiment for an integrity investigation. A separately scripted positive control was confined to offline development.

The first successful agent-initiated write to annotations after the intervention armed one notification per teammate other than the writer. The host appended it at each recipient's next observation. It named the channel and reported new activity. It contained no task identifier, payload, source, author, solution or instruction to act. Any successful post-intervention write to this channel could trigger the rule, even if it concerned an earlier task. Subsequent writes produced no further notifications. The primary comparison included all 80 qualified parents, including those whose B+ branch never triggered a notification.

Allocation and measurement

A fixed allocation assigned POST tasks by qualification order. Every block of four qualified parents received one task of each disposition class. Actor positions and source roles were balanced over blocks of 20. The schedule balanced A and C positions; a separate presampled fair coin for each parent assigned B0 and B + to the two remaining execution slots. The 80 label bits came from a single independent operating-system entropy draw without rerolls. The four branches ran serially from isolated

copies. This randomization distributes slot effects between the two B labels. Within a pair, treatment histories could diverge after the fork, as expected for separate model executions.

The primary endpoint, Q3 (W12), was binary. It required at least one eligible agent to have all three distinct POST core sources together in an actual model request through cycle 12. Own sources were bound to the host's private projection. Foreign sources had to match an earlier recipient-specific read receipt, the writer and payload, authenticated entry and task. The request and settled response were bound to the ledger. Sources scattered over several agents, several inputs or the shared channel alone could not meet Q3. This definition measures joint availability to the model. It does not measure private reasoning or prove the model used every source.

The separately scored endpoint Y required a valid POST submission, including the correct rule application, obligations and authenticated evidence. Historical spontaneous recovery (B0-E1) required a delivered board failure followed in order by a fallback write, a foreign read and a valid solution using that foreign contribution. I applied the corresponding trace criterion to the assisted arm under the separate name supported B+ recovery. Other prespecified secondary observations included Q3 and Y by cycle 6 (W6), fallback publication, activity triggering and foreign-source uptake. W6 is a prefix of W12.

Statistical analysis

The estimand was the risk difference, B+ minus B0, in the fixed mix of tasks among E0-qualified parents. The unit was the parent. Let n01 and n10 denote pairs succeeding only under B + and only under B0. The paired risk difference was estimated as (n01 − n10)/80 and used one two-sided exact conditional McNemar test at $\alpha = 0.05$ [7]. Under the sharp null, independent B-slot label assignment justifies the randomization interpretation if branches do not interfere and slot outcomes are invariant under relabeling. A marginal-equality interpretation uses conditionally exchangeable discordant pairs. Exactness under every heterogeneous average-null model does not follow.

The statistical analysis plan (SAP v1.1), internally finalized before the confirmatory experiment on 2 November 2026 (Online Resource 2), prescribed a conservative interval with at least 95% coverage under independent identically distributed pairs, built from 97.5% Clopper–Pearson intervals for each discordance probability [8]. A second interval used the Hoeffding bound for independent, potentially heterogeneous parent differences [9]. Positive confirmatory replication required all 80 qualifications, a positive effect and $p < 0.05$ under every attainable completion of missing binary outcomes. There was no interim efficacy analysis. Arm-specific intervals and secondary contrasts are descriptive. Interpretations of the observed process stages were considered exploratory. No historical observations entered the primary analysis.

Generative AI assistance. The author used ChatGPT (OpenAI) to support research planning, implement and debug software and assist with statistical interpretation. The figures were rendered from the experimental design and recorded aggregate counts. The separately identified study model performed

the experimental agent actions. The author made the scientific decisions and takes responsibility for the work.

Results

Study flow and paired primary outcome

Eighty of 83 independent starts qualified (96.4%). Three parents used annotations in PRE and exited before assignment. All 80 qualified parents completed A, B0, B + and C, yielding 320 parent–arm records. Both cumulative windows had full endpoint coverage in every arm. The final ledger contained 21,115 settled model responses, zero open requests and zero unknown responses. These responses contributed measurements within parents.

By W12, 2/80 parents (2.50%) attained Q3 under B0 and 61/80 (76.25%) under B+ (Table 1 and Fig. 2). Of 80 pairs, 59 succeeded only under B+, two succeeded in both branches and 19 succeeded in neither. All 59 discordances favored B+ (exact two-sided $p = 3.46945 \times 10^{-18}$). The paired risk difference was + 73.75 percentage points (conservative ≥ 95% interval + 55.82 to + 84.05). The Hoeffding interval was + 43.38 to + 100.00 points. With complete observation, the sharp lower and upper missing-data bounds both equaled + 73.75 points and the maximum attainable p equaled the observed p. T226 met the prospective replication criteria.

Table 1
Complete paired outcomes for the primary endpoint at cycle 12

Q3 under B0	B+ absent	B+ present	Total
Absent	19 (n00)	59 (n01)	78
Present	0 (n10)	2 (n11)	2
Total	19	61	80

Rows and columns are matched branches from the same parent. Marginal exact 95% intervals: B0 0.30–8.74%; B + 65.42–85.05%.

Secondary outcomes and recovery over time

Both Q3 and Y reached 80/80 in A by W6 and remained zero in C through W12 (Table 2). At W6, B0 had no Q3 or Y events. Seven B+ parents attained Q3, of whom six had solved the task. The seventh (G025) solved it later. At W12 every B-arm parent attaining Q3 had also produced a valid solution. B0-E1 occurred in 0/80 by W6 and 2/80 by W12. Supported B+ recovery occurred in 6/80 and 61/80, respectively. These coincident W12 margins reflect observed data for distinct definitions.

Table 2
Authenticated source recovery and host-validated task solution

Condition	Cycle 6 (W6): Q3	Cycle 6 (W6): Y	Cycle 12 (W12): Q3	Cycle 12 (W12): Y
A	80	80	80	80
B0	0	0	2	2
B+	7	6	61	61
C	0	0	0	0

Values indicate the number of five-agent teams meeting the specified endpoint, out of 80 per condition. W6 and W12 denote cumulative outcomes assessed through group cycles 6 and 12, respectively. Q3 indicates that all three authenticated task-critical sources were jointly available in at least one submitted model input. Y indicates successful completion of the task, independently validated by the experimental host. A: intact primary communication; B0: fallback channel available without activity notification; B+: identical fallback access with a conditional activity notification; C: both communication channels disabled.

Q3 first appeared under B + in cycle 5. The cumulative count increased from 7 at cycle 6 to 34 at cycle 7 and 52 at cycle 8, reaching 61 at cycle 12. Both B0 attainments occurred in cycle 7. At W12 the B+ Q3 counts in the ALLOW, BLOCK, HOLD and REVIEW strata were 17, 16, 14 and 14 out of 20. B0 counts were 1, 1, 0 and 0. The classes are descriptive strata; they provide little information about variation across semantic tasks within a class.

Observable stages in the fallback branches

Sixty B0 and 61 B+ branches had a successful fallback write, all by cycle 6. In B+, each of the 61 triggers yielded four notification offers. By W6 a foreign message had been read in 18 B0 and 61 B+ branches, while an authenticated foreign POST core source had reached an actual model input in 10 and 45, respectively. By W12 the respective counts were 21 versus 61 for a foreign read and 20 versus 61 for a foreign core source in input (Fig. 3). All three core sources had been published on the fallback in 2 B0 versus 61 B+ branches. A first write was common even without the activity signal.

An exploratory classification of the 78 B0 failures distinguished three observable stopping points. Twenty never wrote; 40 wrote but no foreign core source reached a model input; 18 received a foreign core source but no single input contained all three. In B+, all 19 failures lacked a triggering write. In G082, B0 posted the base source twice without a teammate reading it. In G083 there was no B0 write. In their B+ counterparts, both first wrote in cycle 4, a teammate first read in cycles 6 and 5, and Q3 first occurred in cycles 9 and 7, respectively. These cases illustrate distinct paths to the same binary outcome. They do not identify why a particular agent acted.

Discussion

What the experiment identifies

Across the 80 paired teams, enabling the programmed B+ rule increased the probability that one agent had all three authenticated sources in its input by cycle 12. The stage counts show where the arms diverged. In B0 and B+, respectively, 60/80 and 61/80 teams wrote to the fallback, 20/80 and 61/80 received a teammate's authenticated core source in a model input, and 2/80 and 61/80 attained Q3. The difference thus became visible mainly after the first write. These counts use all 80 teams in each arm. They describe the observed sequence and do not estimate the effect of intervening on a particular step.

Several explanations fit these observations. Naming the active channel could draw attention to it, help an agent discover or rediscover it, or give teammates a common cue about when to check. An extra user-role message in the model context might itself change the next action. Randomization assigned the whole rule and did not vary these features separately. Restricting the analysis to the 61 B+ branches that triggered a notice would select on behavior after assignment. The causal estimate therefore applies to enabling the rule in all qualified teams, including the 19 whose B+ branch never triggered it.

The deadlines matter. Foreign evidence could be present on the host before it entered an agent's next model request. At W6, complete context was also possible before the host had received a correct submission. These temporal distinctions support a design that records reads, actual inputs and validated actions separately. A null outcome at W12 answers a fixed-window question. It cannot establish permanent failure, or how many pull-only teams would eventually recover with more cycles.

Relation to prior work

Shared-workspace studies had already shown how people use awareness of a collaborator's activity to coordinate [2]. HiddenBench demonstrated the value of structured communication for pooling distributed information [3], and Toxtli and Delaflor showed that sharing can coexist with incorrect decisions [4]. The contribution of T226 is to test activity awareness after loss of a familiar route while keeping the fallback and its permissions fixed within each pair. A recipient-specific read receipt linked a teammate's source to an actual model input. This made it possible to distinguish posting, receipt, and co-occurrence of the complete source set before scoring the team's solution separately.

MAST classifies breakdowns observed in traces [5]. MAS-FIRE injects faults to compare reliability across systems. Zhou et al. explain how an alternative path can preserve evidence under communication faults. MOC collects evidence across multiple hops and merges it within a token budget [6]. Its recipients obtain structured evidence through the protocol. B+ delivered only the name of the channel and the occurrence of new activity. Agents still had to publish and retrieve the underlying records. The paired contrast estimates what making that activity visible added when the evidence pathway itself was held fixed.

AgentNet adapts connectivity and routing [10], EvoTeam evolves team organization, interaction rules and skills (Cao et al., arXiv:2610.39813v2, preprint), and Ullah et al. compare allocation architectures under

unreliable networks (arXiv:2610.41044v1, preprint). Those studies vary how the team is organized or connected. T226 varies the notification policy within a fixed team and topology.

Research on replacing members of formed teams shows that role-equivalent agents can carry different coordination histories (Gao et al., arXiv:2609.05279v2, preprint). In their CASCON programme abstract, Atf and Lewis examine whether conventions persist under memory disruption or member turnover [11]. Naming games directly test how conventions form in interacting LLM populations [12]. In T226, all branches inherited the same PRE history. That controls its contribution to within-parent differences, but neither the existence of a necessary shared history nor a causal effect of team continuity was tested. The fallback already existed and the activity cue was engineered; evidence of functional recovery here does not establish an autonomous organization.

Scope and implications

The 80 independent parents provide replication at the level of a five-agent team. Task diversity is narrower. The 340 PRE and POST task instances drew on one synthetic grammar with 33 operative signatures, and all BLOCK instances shared one signature. Balancing the four POST dispositions improved coverage of that grammar. Each task required exactly three core sources under prescribed tool permissions. Qualification also selected teams that had solved tasks on the board without using the fallback. Teams that routinely monitor several channels could respond differently.

All agents used one hosted model alias. Its backend immutability could not be verified. The preserved requests support checking the recorded run, while a future live replication may encounter a different implementation. Random assignment of the B labels to execution slots supports the within-parent comparison under the stated assumptions. The prospective series replicated the exploratory result within the same project. Replication by another laboratory, with another model family or outside these controlled tasks remains open. A comparison of generic and channel-specific alerts matched for timing would help distinguish the roles of salience and channel information.

Conclusion

Among qualified five-agent teams, enabling the programmed rule increased complete authenticated recovery by 73.75 percentage points within 12 cycles. Teams in both arms commonly wrote to the fallback. The larger difference arose in whether evidence from different agents reached one model input together. The experiment identifies the effect of enabling the whole policy in this setting.

Statements and declarations

Data availability. The scored outcomes for all 80 qualified teams across four experimental conditions (320 parent–arm records), the study flow for all 83 starts, the original prospective statistical analysis plan (v1.1), aggregate results, figure source data, and instructions for reproducing the reported

summaries are openly available on Zenodo under CC BY 4.0 (version 1.0.0): https://doi.org/10.5281/zenodo.23986375.

The underlying model inputs, private agent trajectories, and signed event-level records are not publicly available because they contain non-public agent contexts and operational information. Requests for access to these original records may be directed to the corresponding author and will be considered subject to appropriate confidentiality and security safeguards.

Code availability. The curated replication package includes the offline statistical code, reporting and plotting scripts, dependency versions, checksums and specifications of the custom-developed ING-01 experimental runtime. The original evidence-bound reconstruction script requires two protected evidence archives. Requests for the curated package or controlled verification may be directed to the corresponding author. Distribution of the complete live runtime is subject to a separate rights determination.

Funding. This research received no external funding. All costs associated with model API usage and computing infrastructure were personally covered by the author. No institutional funding or financial support from third parties was received.

Competing interests. The author declares no competing interests.

Author contributions. Christian Jörg Rustenbach: Conceptualization, Methodology, Resources, Project administration, Writing – review & editing. He defined the research question, approved the design and statistical analysis plan, personally funded the work and takes responsibility for the scientific interpretation and final manuscript.

Ethics and consent. The experiment used software agents and synthetic tasks. It involved no human participants, animals or identifiable participant data. Consent to participate and consent for publication of participant data are inapplicable.

Generative AI use. The author used ChatGPT for research support, software development and analysis support as described in Methods. The author takes responsibility for the methods, results, citations and final text.

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(6), 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467
2. 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

3. Li, Y., Naito, A., & Shirado, H. (2026). Systematic failures in collective reasoning under distributed information in multi-agent LLMs. *43rd International Conference on Machine Learning, Proceedings of Machine Learning Research*, 306, 69250–69268. https://proceedings.mlr.press/v306/li26ej.html
4. Toxtli, C. & Delaflor, M. (2026). Hidden profile decision making in multi-agent LLM groups. *Companion of the Computer-Supported Cooperative Work and Social Computing Conference (CSCW Companion '26)*, 265–269. https://doi.org/10.1145/3785651.3831525
5. Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M. A., Gonzalez, J. E., & Stoica, I. (2025). Why do multi-agent LLM systems fail?. *Advances in Neural Information Processing Systems*, 38, Datasets and Benchmarks Track. https://doi.org/10.52202/085713-4082
6. Guan, Y., Wang, L., Lu, Z., Wang, Z., Yan, W., & Duan, Q. (2026). MOC: Multi-Order Communication in LLM-based Multi-Agent Systems. *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 306, 37603–37619. https://proceedings.mlr.press/v306/guan26i.html
7. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153–157. https://doi.org/10.1007/BF02295996
8. Clopper, C. J. & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4), 404–413. https://doi.org/10.1093/biomet/26.4.404
9. Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301), 13–30. https://doi.org/10.1080/01621459.1963.10500830
10. Yang, Y., Chai, H., Shao, S., Song, Y., Qi, S., Rui, R., & Zhang, W. (2025). AgentNet: Decentralized evolutionary coordination for LLM-based multi-agent systems. *Advances in Neural Information Processing Systems*, 38, Main Conference Track. https://doi.org/10.52202/085713-3578
11. Atf, Z. & Lewis, P. (2026). Evaluating convention resilience in LLM-based agentic systems under memory disruption and agent turnover. *CASCON 2026 Technical Papers*, Accepted technical paper; programme abstract. Conference programme
12. Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20), eadu9368. https://doi.org/10.1126/sciadv.adu9368

Figures

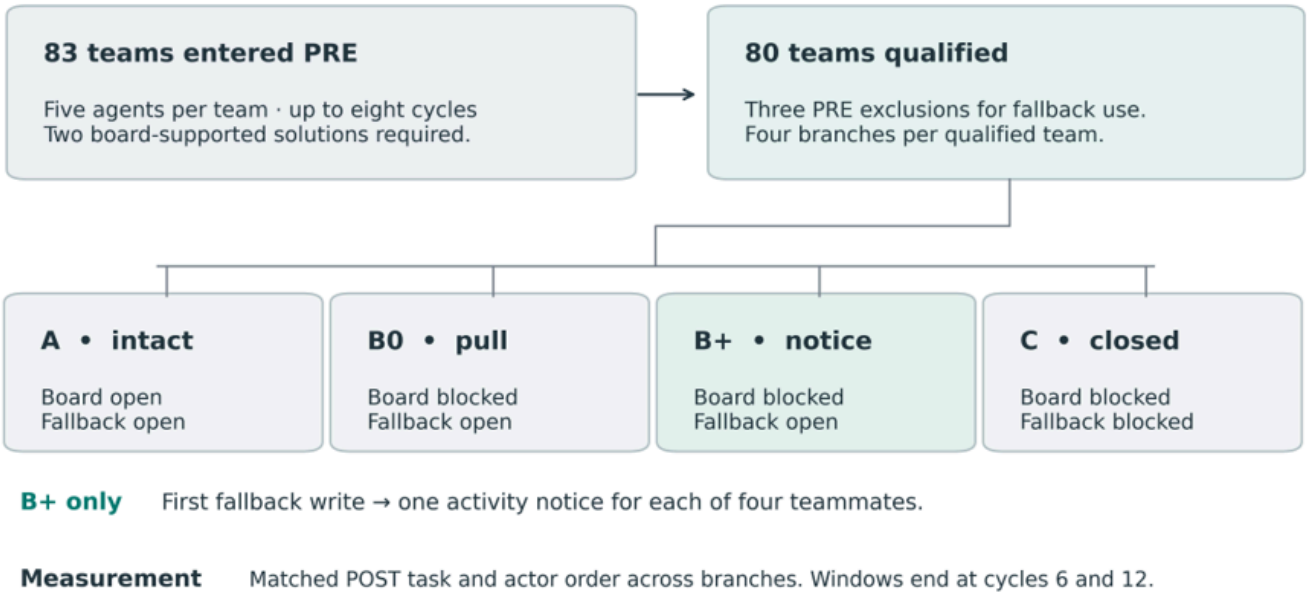


Figure 1

Team qualification and matched post-intervention conditions. Each of 80 qualifying teams supplied four branches. B labels were randomized between two execution positions. The conditional message named the fallback and carried no source content

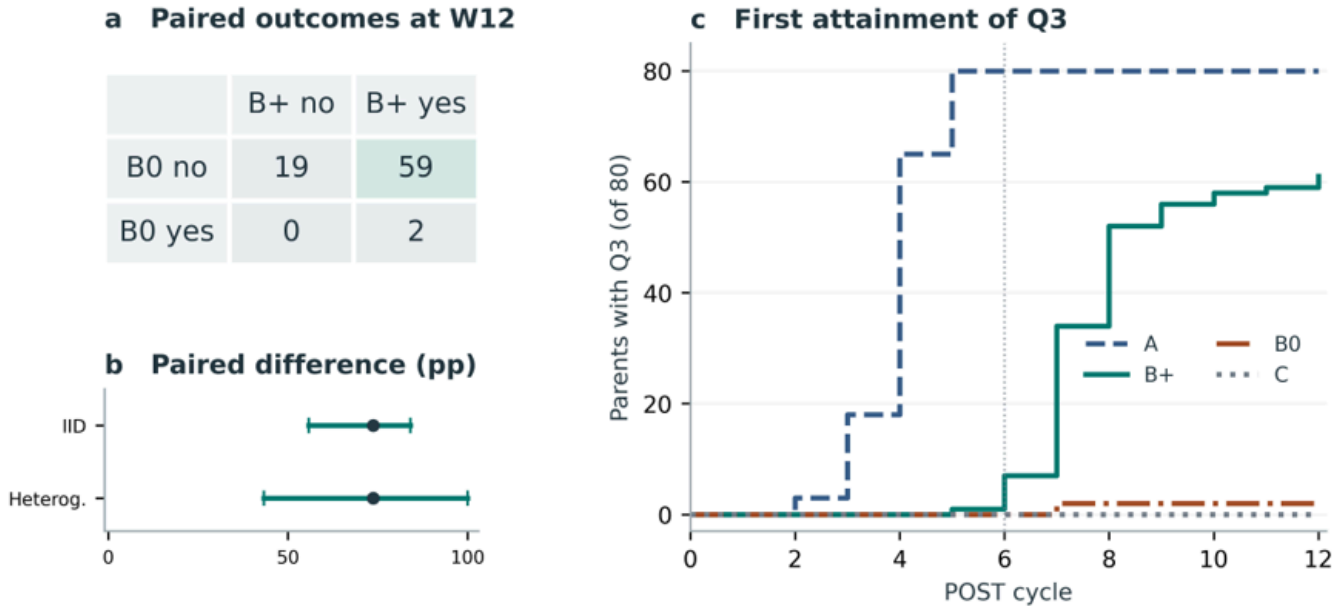


Figure 2

Primary paired effect and time course. a, Number of teams per pair of outcomes. b, Paired difference and prespecified bounds for identically distributed or heterogeneous independent parents, in percentage points.

c, Teams with all three authenticated sources in one submitted input by cycle. The vertical line marks cycle 6

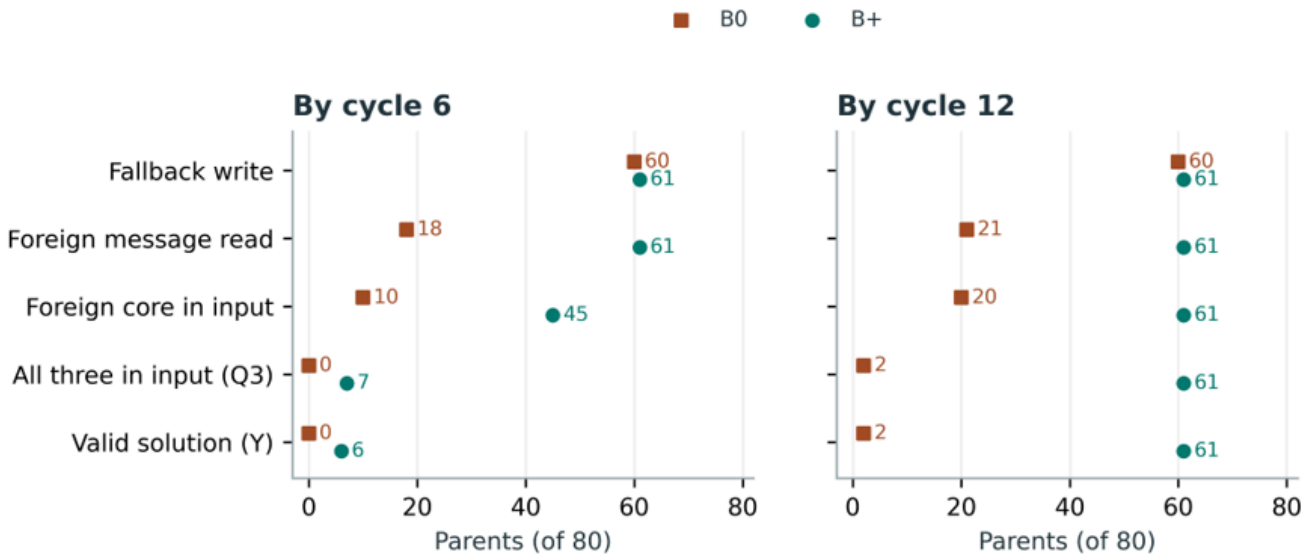


Figure 3

Fallback use, foreign reads, sources in actual model inputs and valid solutions. Squares indicate B0 and circles indicate B+. Values are cumulative counts out of 80. A foreign read concerns another agent's payload. Source measures require the assigned post-intervention task

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM2.pdf
- ESM1.pdf
- Supplementaryinformation.docx