TRACE2ACT: AUDITING WHEN REFUTED PREMISES DRIVE IRREVERSIBLE MULTI-AGENT DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A final action can look correct while its transcript reveals that the stated justification depends on a premise the available record already refutes. Trace2Act introduces commitment-boundary evaluation for this failure. Given fixed evidence and an irreversible-action rubric, its deterministic aggregation rule identifies a reinforcing path from an evidence-refuted claim to commitment before any challenge or retraction. On a pooled home set of $1{,}000$ matched transcripts, answer-level factuality misses 125 action-coupled episodes and adds 36 positives outside the commitment-boundary endpoint; adjacent evaluators consequently expose different failure sets on identical records. A 600-transcript blinded endpoint audit yields Cohen's $\kappa = 0.813$ and Krippendorff's $\alpha = 0.809$. By retaining the premise-to-action path, Trace2Act supplies an auditable criterion for evaluating how agent workflows authorize irreversible decisions.

1 INTRODUCTION

Multi-agent systems increasingly mediate incident triage, patch approval, procurement, and compliance decisions. In these workflows, the path to commitment is a distinct safety signal: a team can choose a plausible action while its transcript shows that the stated justification depends on a premise contradicted by the available record. Repetition, delegation, and shared memory can turn that premise into decision-grade input; a restart, merge, purchase, or filing then commits the workflow to its consequence. Evaluation therefore needs the chain from available evidence through transcript reuse to action justification.

We call this event *premise-to-action coupling*. It occurs at the commitment boundary where a checkably false claim becomes the stated reason for an irreversible action. Common evaluator interfaces project a trajectory onto a local claim, final answer, task outcome, or aggregate turn quality. That projection omits the justification path. Trace2Act makes this distinction testable by applying several evaluator interfaces to identical transcripts and recording their asymmetric episode-level disagreements.

Trace2Act couples a closed evidence pack with a domain-specific irreversible-action rubric. A deterministic episode rule follows an evidence-refuted claim through reinforcing transcript edges and records a positive event when the path reaches a qualifying operationalize edge before challenge or retraction. The resulting claim-state graph exposes the evidence, edge sequence, and action that determine the decision. On the pooled home comparison, factuality misses 125 Trace2Act-positive episodes and adds 36 positives outside the endpoint; the same pattern appears in independently authored external packs. A blinded home audit independently assesses the endpoint.

This paper contributes:

1. A commitment-boundary scored object and deterministic audit rule that recover the path by which a checkably false premise authorizes irreversible action.
2. A matched-transcript comparison of adjacent evaluator interfaces, with pooled count and denominator reporting that makes asymmetric disagreements directly auditable.
3. An independently annotated endpoint audit and external-pack comparison that test whether the scored object remains recoverable beyond one authoring surface.

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.

Together, the definition and comparisons establish an evaluation principle for action-taking agents: retain the claims that authorize commitment and the paths by which those claims acquire decision authority.

2 Related Work: What the Evaluator Retains

Factuality benchmarks evaluate the truth of claims or answers (Lin et al., 2022; Min et al., 2023; Li et al., 2023b; Manakul et al., 2023), while HELM and BIG-bench aggregate behavior across broad task collections (Liang et al., 2023; Srivastava et al., 2023). These protocols establish whether content is correct or capabilities are broad. Trace2Act retains an additional dependency: whether a refuted claim appears on the justification path to an irreversible action. The matched factuality comparison measures the empirical difference on the same episodes.

Final-output evaluators such as Chatbot Arena and G-Eval score response quality or preference (Zheng et al., 2023; Liu et al., 2023). AgentBench, AgentBoard, and τ-bench evaluate task completion, subgoal progress, grounding, or final environment state (Liu et al., 2024; Ma et al., 2024; Yao et al., 2025); multi-agent frameworks expose how coordination changes those outcomes (Li et al., 2023a; Wu et al., 2024; Hong et al., 2024; Qian et al., 2024). These task-performance views can assign the same score to actions reached through different justifications. Trace2Act aligns the transcript, evidence pack, and action rubric at episode level, so a decision can be inspected together with its stated rationale.

Agent-safety benchmarks broaden evaluation beyond task success. ToolEmu scores consequential tool failures, R-Judge tests risk recognition from interaction records, AgentHarm measures completion of malicious multi-step tasks, and ST-WebAgentBench evaluates policy compliance along action traces (Ruan et al., 2024; Yuan et al., 2024; Andriushchenko et al., 2025; Levy et al., 2026). AgentRewardBench studies judges of trajectory success, side effects, and repetition (Lù et al., 2025). Trace2Act resolves a complementary decision variable: under evidential closure, did a specifically refuted premise acquire authority and cross an irreversible commitment boundary?

Debate, verification, and process supervision expose intermediate reasoning or introduce challenge (Irving et al., 2018; Du et al., 2024; Dhuliawala et al., 2024; Lightman et al., 2024); work on agent visibility similarly emphasizes observability of interaction (Chan et al., 2024). Trace2Act binds process information to an auditable commitment event with a pre-action interrupt condition. Its matched protocol turns a scored-object distinction into directly countable disagreement.

3 Commitment-Boundary Measurement

3.1 Auditable Premise-to-Action Rule

The unit of evaluation is one scenario-seed episode $e_{s,z}$. Each episode is audited as a directed claim-state graph $G_{s,z} = (V^{\text{claim}} \cup V^{\text{action}}, E)$: claim nodes normalize proposition identity, action nodes mark rubric-qualified commitments, and edges record typed epistemic interactions with agent identity and transcript position. Trace2Act uses six edge types: mention, affirm, cite-as-support, operationalize, challenge, and retract.

Let $\mathcal{F}_s$ be the seeded claims contradicted either directly or jointly by scenario s's fixed evidence, $\mathcal{A}_s^{\text{irr}}$ its rubric-defined irreversible actions, $\mathcal{R}$ the reinforcing edge set, and $\mathcal{C}$ the challenge/retract set. The primary label is

$$y_{s,z} = \mathbf{1}\big[\exists f \in \mathcal{F}_s,\ a \in \mathcal{A}_s^{\text{irr}},\ P : f \rightsquigarrow a,\\ P \subseteq \mathcal{R},\ \text{last}(P) = \text{operationalize},\ \nexists c \in \mathcal{C}(f) : t(c) < t(a)\big].$$

The interruption condition is episode-wide for the same seeded claim: a prior challenge or retraction excludes that candidate path even if the claim is subsequently re-endorsed. Mention alone cannot yield a positive event; the path must end in a qualifying action justification.

3.2 Scenario Design and Metrics

Trace2Act contains 50 scenarios across incident triage (15), code patch approval (13), procurement authorization (12), and regulatory compliance (10). Every scenario includes a task brief, a fixed evidence pack, two to four plausible but checkable seeded claims, and a rubric covering data-loss, financial-commitment, and operational-exposure actions. A scenario qualifies as fixed-evidence when its pack alone directly contradicts, or jointly rules out, each seeded false claim.

For a condition with S scenarios and five seeds, the binary irreversible error rate (IER) is

$$\hat{p}_z = 100S^{-1}\sum_{s=1}^{S} y_{s,z}, \qquad \bar{p} = 5^{-1}\sum_{z=1}^{5}\hat{p}_z.$$

Matched-evaluator comparisons use a separate, explicitly pooled estimator. For a fixed transcript population of size N, the displayed incidence is $100N_+/N$, where N_+ is the printed positive count. Count tables therefore never infer an integer from a rounded condition-level rate.

Time-to-False-Belief (TFB) is an observed transcript duration conditional on a Trace2Act-positive episode. It is the interaction-step distance from the first mention of the seeded claim on the qualifying path to the first qualifying operationalize edge. Episodes without a positive path are outside this risk set and are not censored. Pre-action features are extracted at the transcript state immediately before that qualifying edge.

The 95% intervals in Appendix B resample scenarios as clusters, carrying all five seeds from a selected scenario together. They are reported only as interval estimates; the manuscript makes no exact-test or multiple-comparison claim.

4 Experimental Setup

4.1 Episode Account and Scoring

The evaluated system families are single-agent (SA), pairwise discussion (PD), a three-agent shared-memory team (SM-3), and a hierarchical planner-worker system (HP). The primary 50-scenario, five-seed matrix used `gpt-4o-2024-05-13`, temperature 0.7, top-$p = 1.0$, and a 128K context.

The retained episode account is a non-overlapping generation-cell ledger: $50 \times 5 \times 15 = 3{,}750$ for the primary matrix, $20 \times 5 \times 8 = 800$ for two non-GPT-4o cross-backbone condition sets, $50 \times 5 \times 4 = 1{,}000$ for handle quality, $50 \times 5 \times 3 = 750$ for SM-3 topology, $20 \times 5 \times 8 = 800$ for Trace2Act-Ext, $50 \times 5 \times 3 = 750$ for matched compute, $15 \times 5 \times 8 = 600$ for Trace2Act-Blind, and $50 \times 5 \times 2 = 500$ for paired HP visibility. These cells total 8,950 generated episodes. The external matched comparison uses the four no-mitigation system-family conditions per scenario-seed: its 400 Ext and 300 Blind transcripts are subsets of the respective 800- and 600-episode ledger cells. Transcript rescoring and human annotation do not create new episodes.

Scoring normalizes seeded-claim mentions against fixed evidence, assigns one of the six edge labels, identifies rubric-qualified actions, assembles reinforcing paths in transcript order, and applies the binary rule. The primary edge extractor is a GPT-4o classifier trained on 500 manually annotated messages disjoint from all audit samples. Three blinded domain experts annotate endpoints from transcript, evidence pack, and action rubric; a Llama-3 70B extractor, an orthogonal rubric-first route, and an SBERT-assisted finite-state scorer provide additional scoring paths.

The adjacent evaluators preserve their native target questions while receiving exactly the information in Table 1. The matched design fixes the transcript population while changing complete evaluation interfaces. Its disagreement counts therefore compare criterion and visible information together, rather than attributing an effect to either component alone.

Table 1: Matched-evaluator protocol. Every row scores the same episode using only the listed inputs and applies the stated episode aggregation.

Evaluator	Input visibility	Positive criterion and episode aggregation
Trace2Act	Transcript, evidence, action rubric	Refuted premise justifies irreversible action; episode OR over seeded claims
Factuality	Final answer and evidence	Any claim contradicted by evidence; claim maximum
Final judge	Task, evidence, final answer	Unsafe or unsupported final output; one episode judgment
Task success	Task and final action	Incorrect or failed task outcome; binary episode failure
Process monitor	Transcript only	Low-quality or unsupported interaction; maximum turn score

Table 2: Pooled commitment-boundary comparison on the same $1{,}000$ no-mitigation home transcripts. Pooled positives is the integer N_+ and exact incidence $100N_+/1{,}000$. The two disagreement columns are integer episode counts.

Evaluator	Pooled positives	T2A+/E−	T2A−/E+
Trace2Act	196 (19.6%)	—	—
Factuality	107 (10.7%)	125	36
Final judge	135 (13.5%)	99	38
Task success	126 (12.6%)	109	39
Process monitor	152 (15.2%)	83	39

5 Results

5.1 Evaluator Choice Changes the Observed Failure Set

The disagreement is episode-level and directional. On the home set, factuality misses 125 action-coupled episodes while adding 36 positives outside Trace2Act; final judging, task success, and the process monitor each produce a distinct asymmetry. The 125 factuality-missed cases partition into 41 episodes with locally correct wording after unsafe justification, 34 that omit the causal premise from the final output, 29 that reach a correct action for a refuted reason, and 21 with fluent interaction but decision-unsafe reuse. This partition connects the aggregate difference to information that local answer evaluation omits.

5.2 Independent Endpoint Recovery

The blinded home audit reaches Cohen's $\kappa = 0.813$ for the binary endpoint and Krippendorff's $\alpha = 0.809$. The agreement analysis separates endpoint reliability from finer-grained edge-label distinctions, keeping the audit target aligned with the commitment-boundary rule.

6 Discussion

An evaluator is also a decision rule: the information it retains determines which failures it can surface. Trace2Act's matched asymmetries show that answer and output compression can omit an action's evidential justification even when the final response is locally plausible. Preserving the path changes what an evaluator records. Benchmark designers can score the commitment boundary, and system builders can inspect the premise paths that authorize irreversible actions. The blinded audit shows that this endpoint can be independently applied using the transcript, evidence, and action rubric.

The commitment boundary turns a broad concern about unsupported reasoning into a checkable object: a particular refuted proposition, a typed sequence of reuse, a qualifying action, and an

Table 3: Blinded home endpoint-annotation audit. Candidate edges are restricted to seeded-claim-relevant interactions.

Audit	Transcripts	Candidate edges	Pairwise κ	Krippendorff's α
Home human	600	2,400	0.813	0.809

interruption condition. This structure makes disagreements between evaluators inspectable rather than merely qualitative.

7 Conclusion

Trace2Act establishes commitment-boundary evaluation for fixed-evidence, irreversible decisions. Its deterministic rule records the evidence-refuted premise path that authorizes an action, while matched evaluator interfaces reveal asymmetric failure sets on the same transcripts. By treating justification as part of the scored object, Trace2Act provides a precise basis for auditing how multi-agent workflows reach commitment.

AI Use Statement

Generative AI tools were used solely for narrative phrasing, copyediting, and presentation of this manuscript. They were not used to generate synthetic datasets; develop theoretical or conceptual frameworks; formulate mathematical claims, proofs, or hypotheses; design or provide feedback on methodology or experiments; implement methods; assist with translation; clean or reformat data; support qualitative or thematic data analysis; or interpret results. The authors reviewed all AI-assisted work and remain responsible for the final content.

Reproducibility Statement

Sections 3–4 define the scenario unit, fixed-evidence criterion, action taxonomy, edge labels, decision rule, condition-level and pooled estimators, episode ledger, and evaluator inputs. Appendix A gives the external comparison populations and the exact arithmetic behind every retained cross-evaluator gap.

References

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations*, pp. 79185–79220, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c493d23af93118975cdbc32cbe7323f5-Abstract-Conference.html.

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. doi: 10.1145/3630106.3658948. URL https://doi.org/10.1145/3630106.3658948.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, 2024. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference*

on Machine Learning, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, pp. 23247–23275, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6507b115562bb0a305f1958ccc87355a-Abstract-Conference.html.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. doi: 10.48550/arXiv.1805.00899. URL https://arxiv.org/abs/1805.00899.

Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. ST-WebAgentBench: A benchmark for evaluating safety and trustworthiness in web agents. In *International Conference on Learning Representations*, pp. 122213–122255, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c6c29e590e3c62e37e6b39cdd6baf2e8-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, pp. 51991–52008, 2023a. doi: 10.52202/075280-2264. URL https://proceedings.neurips.cc/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023b. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397/.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

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*, pp. 52989–53046, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J. Pal, and Siva Reddy. AgentRewardBench:

Evaluating automatic evaluations of web agent trajectories. *arXiv preprint arXiv:2504.08942*, 2025. doi: 10.48550/arXiv.2504.08942. URL https://arxiv.org/abs/2504.08942.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 74325–74362, 2024. doi: 10.52202/079017-2365. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/877b40688e330a0e2a3fc24084208dfa-Abstract-Datasets_and_Benchmarks_Track.html.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, 2024. doi: 10.18653/v1/2024.acl-long.810. URL https://aclanthology.org/2024.acl-long.810/.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, pp. 27031–27098, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7274ed909a312d4d869cc328ad1c5f04-Abstract-Conference.html.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uyTL5Bvosj.

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 *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, pp. 9965–10017, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. R-judge: Benchmarking safety risk awareness for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1467–1490. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.79. URL https://aclanthology.org/2024.findings-emnlp.79/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information*

Processing Systems, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A External Matched-Transcript Evaluation

The external matched set contains no-mitigation transcripts from Trace2Act-Ext and Trace2Act-Blind. Trace2Act-Ext has 20 scenarios and Trace2Act-Blind has 15 scenarios; each has four system families and five seeds, giving pooled populations of 400 and 300 transcripts, respectively. These are the no-mitigation subsets of the 800- and 600-episode generation cells in Section 4.1; conditions outside those subsets are not pooled into this comparison, and the 600-transcript human audit is annotation rather than a further generation cell. Table 4 reports only signed count differences that can be checked directly from those denominators. For each row, the reported gap is $100(n_{+/-} - n_{-/+})/N$ after rounding to one decimal.

Table 4: Pooled external matched-transcript disagreement. $+/-$ denotes Trace2Act-positive and adjacent-evaluator-negative; $-/+$ is the reverse. Each gap is calculated from the displayed counts and the stated pack denominator.

Pack	Evaluator	$+/-$	$-/+$	Signed gap
Ext ($N = 400$)	Factuality	57	15	10.5 pp
	Final judge	46	16	7.5 pp
	Task success	51	16	8.8 pp
	Process monitor	40	18	5.5 pp
Blind ($N = 300$)	Factuality	43	11	10.7 pp
	Final judge	35	14	7.0 pp
	Task success	39	13	8.7 pp
	Process monitor	31	14	5.7 pp

Every external row has more Trace2Act-positive/adjacent-negative cases than the reverse. The two independently authored packs therefore preserve the matched-interface distinction while exposing the exact population and count arithmetic behind the result.

B Supplementary Annotation Agreement

Table 5: Inter-annotator agreement on the 600-transcript, 2,400-edge human audit. Three annotators independently labeled each transcript's IER endpoint and each edge type under the shared rubric and blinding protocol. Cohen's κ is the pairwise mean and Krippendorff's α uses all three annotators.

Target	Pairwise κ	α	95% CI (α)
IER endpoint (binary)	0.813	0.809	$[0.776, 0.842]$
Edge label (6-class)	0.762	0.784	$[0.751, 0.817]$
Operationalize edges (binary)	0.841	0.833	$[0.801, 0.865]$
Challenge/Retract edges	0.794	0.788	$[0.752, 0.824]$
Cite-vs-Affirm distinction	0.683	0.701	$[0.659, 0.743]$

IER and operationalize edges have pairwise $\kappa > 0.80$, while the cite-as-support versus affirm boundary has the lowest agreement. Disagreements were resolved by majority consensus with a post-audit adjudication round for edges without a majority; 4.2% of edges required adjudication. The agreement profile follows the endpoint's dependency structure: raters recover the binary action-coupling call more consistently than the fine-grained distinction between two reinforcing edge types.