AGENTEX-BENCH: A STEP-LEVEL BENCHMARK FOR EVALUATING EXPLANATION QUALITY IN AGENTIC TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model agents increasingly expose intermediate plans, evidence, calculations, and tool decisions, creating an opportunity to evaluate how their observable behavior supports task outcomes. However, existing agent evaluation remains centered on task success, action correctness, or domain-specific process checks. These metrics do not determine whether an emitted trajectory provides a well-grounded, specific, and diagnostically valuable explanation of agent behavior. We introduce AGENTEX-BENCH, a benchmark and measurement framework for evaluating observable agent trajectories as auditable artifacts. AGENTEX-BENCH standardizes heterogeneous reasoning and tool-use trajectories and evaluates key steps along five complementary dimensions: validity, task-grounded faithfulness, specificity, diagnostic utility, and evidence alignment. The framework combines source-aware outcome verification, provenance-aware annotation, controlled perturbation tests, and uncertainty-aware comparison across four task families and 15 systems. Our analyses show that task success and trajectory explanation quality provide related but non-equivalent views of agent behavior. Dimension-level profiles reveal distinctions that aggregate scores can obscure, while matched human-reference analyses show substantial uncertainty in fine-grained model ordering. These findings support treating observable trajectory quality as a distinct evaluation target that complements task success and process correctness for agent auditing, diagnosis, and oversight.

1 INTRODUCTION

As language models move from generating isolated responses to executing multi-step tasks, evaluating their reliability becomes increasingly difficult. A task outcome reveals whether a system succeeded, but often says little about why it succeeded, where it failed, or whether its intermediate behavior can be meaningfully inspected. This distinction matters for agent systems whose decisions depend on retrieved evidence, calculations, tool choices, and intermediate judgments: two agents may reach the same outcome through trajectories that differ substantially in grounding, auditability, and diagnostic value. Reliable agent evaluation therefore requires not only measuring outcomes, but also determining what can be learned from the observable record that produced them.

Existing evaluation methods capture different parts of this problem. Outcome-oriented agent benchmarks measure whether agents complete tasks successfully, while process-supervision and reasoning benchmarks evaluate intermediate correctness (Uesato et al., 2022; Lightman et al., 2024; Zheng et al., 2025). Agent and tool-use work studies reasoning-action interaction, interactive task execution, function selection, and argument construction (Yao et al., 2023; Liu et al., 2024; Patil et al., 2025). A separate line of work examines whether generated explanations faithfully reflect the factors influencing model behavior (Jacovi & Goldberg, 2020; Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025), while evidence-grounding benchmarks evaluate whether generated content is supported by provided sources (DeYoung et al., 2020; Jacovi et al., 2025). These perspectives are complementary and target distinct evaluation objectives. What remains missing is a common framework for evaluating the observable trajectory itself as an explanatory artifact: not only whether its individual steps are correct, but whether the trajectory is grounded in task evidence, sufficiently specific to be audited, and informative enough to help diagnose success or failure. Building such a

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.

framework raises three challenges. First, agent trajectories are heterogeneous: a meaningful step may be a numerical calculation, an evidence citation, a Boolean inference, a function choice, or a structured tool call. Second, explanation quality is multidimensional. A step may be correct but unsupported, highly specific but linked to the wrong evidence, or incorrect but diagnostically valuable because it clearly exposes the source of failure. Third, trajectory evaluation is itself a measurement problem: human judgments can disagree, score provenance may differ across examples, and apparent model rankings can depend on task composition and annotation source. A useful benchmark must define what is being measured, preserve the provenance of each judgment, and quantify the reliability of the comparison results.

We address this problem with AGENTEX-BENCH, a benchmark and measurement framework for observable agent trajectories. The framework has three components. First, it maps heterogeneous reasoning and tool-use outputs into a shared trajectory representation while preserving task-specific evidence and action semantics. Second, it operationalizes trajectory explanation quality through five complementary dimensions—validity, task-grounded faithfulness, specificity, diagnostic utility, and evidence alignment—so that correctness can be separated from support, concreteness, and diagnostic value. Third, it treats evaluation reliability as part of the benchmark itself through explicit score provenance, human-reference analysis, controlled perturbation tests, and uncertainty-aware system comparison. Across four task families and 15 systems, we find that outcome correctness and trajectory explanation quality provide related but non-equivalent information, that dimension-level profiles reveal different weaknesses across models and task structures, and that fine-grained system ordering becomes substantially less stable once task composition and measurement uncertainty are controlled. We therefore view AGENTEX-BENCH not primarily as a leaderboard, but as a measurement instrument for determining what observable agent trajectories can reliably support in auditing and comparative evaluation. Our main contributions are as follows:

- We formulate observable trajectory explanation quality as a distinct evaluation target and introduce a shared representation that preserves task-specific evidence and action semantics across heterogeneous reasoning and tool-use tasks.
- We introduce a five-dimensional step-level rubric for validity, task-grounded faithfulness, specificity, diagnostic utility, and evidence alignment, while keeping these measurements separate from source-specific outcome correctness.
- We validate the resulting measurement instrument through provenance-aware annotation, controlled perturbations, matched-task analysis, and uncertainty-aware inference. These analyses characterize rubric sensitivity, annotation reliability, and the resolution supported for system comparisons.

2 RELATED WORK

Process and trajectory evaluation. A large number of works evaluate intermediate behavior rather than final outcomes alone. Process supervision distinguishes step-level correctness from outcome correctness (Uesato et al., 2022; Lightman et al., 2024), while ProcessBench and PRMBench evaluate reasoning-process errors and process-level reward signals (Zheng et al., 2025; Song et al., 2025). Recent benchmarks extend process evaluation to agent trajectories, including tool-use effectiveness, trajectory assessment, and failure localization (Fan et al., 2026; Lù et al., 2025; Deshpande et al., 2025; Zhang et al., 2025; Barke et al., 2026; Yehudai et al., 2026). These approaches motivate process-level evaluation; AGENTEX-BENCH instead focuses on whether an observable trajectory is sufficiently grounded and diagnostically informative to serve as an auditable explanatory record.

Faithfulness, grounding, and monitorability. A separate literature studies whether explanations are supported by the information they claim to use. ERASER and FACTS Grounding evaluate whether generated rationales or responses are supported by provided evidence (DeYoung et al., 2020; Jacovi et al., 2025). This external grounding differs from causal faithfulness, which asks whether an explanation reflects the factors that influence model behavior (Jacovi & Goldberg, 2020; Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025; Tutek et al., 2025). Related work on chain-of-thought monitorability studies whether visible traces provide useful signals for oversight and how robust those signals are (Korbak et al., 2025; Baker et al., 2025; Guan et al., 2025). In AGENTEX-BENCH, faithfulness refers to support by observable task information rather than access to latent computation.

Table 1: Task sources in AGENTEX-BENCH.

Source	Tasks	Primary trajectory demand
BFCL	20	Function and argument selection
GSM8K	20	Multi-step arithmetic reasoning
HotpotQA	20	Multi-document evidence composition
StrategyQA	20	Implicit Boolean reasoning

Agent and tool-use benchmarks. ReAct studies reasoning-action interaction, AgentBench evaluates agents in interactive environments, and BFCL focuses on function selection and argument construction (Yao et al., 2023; Liu et al., 2024; Patil et al., 2025). τ^2-Bench further evaluates conversational agents in interactive settings (Barres et al., 2025). These works primarily assess whether agents choose or execute appropriate actions, whereas AGENTEX-BENCH treats the emitted trajectory itself as the object of measurement and complements environment-level success.

Automated evaluation and judge reliability. LLM-based evaluators enable scalable assessment but can exhibit systematic bias and instability (Zheng et al., 2023). JudgeBench and RewardBench 2 further study judge reliability (Tan et al., 2025; Malik et al., 2026). Accordingly, AGENTEX-BENCH keeps assisted scores, rule-based labels, human consensus, and adjudicated scores distinct in its provenance records.

Positioning of AGENTEX-BENCH. Prior work provides strong tools for evaluating outcomes, process correctness, tool use, grounding, faithfulness, and evaluator reliability. AGENTEX-BENCH complements these directions with a common framework for measuring whether heterogeneous observable trajectories form grounded, diagnostically useful, and auditable records across task types.

3 BENCHMARK CONSTRUCTION

Cross-task trajectory evaluation requires a representation that is comparable across tasks without discarding the evidence and action semantics needed for scoring. AGENTEX-BENCH addresses this problem by pairing complementary task families with a shared observable-trajectory representation. The representation normalizes the structure of model outputs while retaining source-specific evidence, calculations, and intended tool calls. This provides a common unit for the trajectory-quality measurements introduced in Section 4.

3.1 TASK SOURCES

We select four task families that expose different forms of observable support. BFCL emphasizes schema-constrained tool decisions, GSM8K explicit calculations, HotpotQA document evidence, and StrategyQA fact-based Boolean reasoning. This variation allows the same evaluation framework to be examined across qualitatively different types of trajectory steps. Table 1 summarizes the task sources and their primary trajectory demands.

The benchmark contains 80 tasks, with 20 tasks from each source. We retain stable AGENTEX-BENCH task identifiers and source provenance. Source-specific eligibility criteria are provided in Appendix D.1.

BFCL evaluates schema-aware function and argument selection (Patil et al., 2025), while GSM8K provides explicit multi-step numerical reasoning problems (Cobbe et al., 2021). HotpotQA requires evidence composition across documents (Yang et al., 2018), whereas StrategyQA targets implicit reasoning toward Boolean conclusions (Geva et al., 2021).

3.2 OBSERVABLE TRAJECTORY REPRESENTATION

For model m on task t, let $y_{m,t}$ denote the final answer and let

$$\tau_{m,t} = (s_{m,t,1}, \ldots, s_{m,t,L_{m,t}}) \tag{1}$$

denote the observable trajectory, where each step records an identifier, a step type, concise content, and optional evidence references or tool calls. The schema standardizes how trajectories are represented without removing the source-specific information needed for evaluation. This distinction is important because support takes different forms across tasks: a document citation, a numerical derivation, and a function argument cannot be evaluated in exactly the same way.

Source-specific step types are mapped to broader categories used by the evaluation framework. The original step type is retained so that source-level behavior can still be inspected.

Models generate the final answer and observable trajectory in a single structured response. The shared protocol requests 3–10 concise steps describing the evidence, calculations, or decisions relevant to the task. Source-specific instructions adapt this protocol to each task family. For BFCL, tool calls represent intended actions rather than executed interactions. Appendix G and Figure 3 provide the complete protocol.

3.3 Evaluated Models and Coverage

We evaluate 15 hosted systems spanning proprietary, compact, reasoning-oriented, open-weight, and routing baselines. Thirteen systems have 80 trajectories, OpenRouter Free has 77, and Nemotron 3 Ultra has 69, yielding 1,186 trajectories. We do not pad or interpolate missing runs. OpenRouter Free is a routing baseline, not a fixed model identity.

Initial prompts exclude gold answers, but the validation-retry path can disclose reference answers or expected function names after a mismatch. The retained corpus therefore measures outputs produced under the documented protocol rather than uniformly gold-blind first-pass capability. A saved-output audit identifies 138 trajectories with one recorded validation retry and 80 with missing retry metadata. Simply removing retried cases would condition on earlier failure and would not recover an unbiased first-pass comparison; Appendix B provides the full audit.

4 Measuring Trajectory Explanation Quality

Once trajectories are represented in a common form, the next question is what properties make them useful for auditing and diagnosis. AgentEx-Bench separates two evaluation targets. Outcome correctness measures whether the task was solved, whereas trajectory explanation quality measures the observable record that supports the result.

4.1 Step-Level Rubric

We organize trajectory quality around three requirements of an auditable record: correctness, support, and inspectability. Validity captures correctness. Faithfulness and evidence alignment characterize whether claims and references are supported by observable task information. Specificity and diagnostic utility characterize whether the record is concrete enough to inspect and useful for identifying key decisions or failure points.

We score only substantive decision, evidence, calculation, tool-use, and answer-construction steps. Generic restatements and transitions may remain unscored. Each scored step receives five ordinal ratings in $\{0, 1, 2\}$:

Validity	whether the step is correct for the task state;
Faithfulness	whether it is supported by the prompt, evidence, schema, facts, or mathematical conditions;
Specificity	whether it identifies concrete entities, values, documents, functions, or arguments;
Diagnostic utility	whether it exposes a decision or error source that helps explain success or failure; and
Evidence alignment	whether cited evidence or tool calls directly support the step content.

The dimensions are intentionally allowed to disagree. A step can be correct yet unsupported, and an incorrect step can still reveal the source of failure. These distinctions would be collapsed by a single

process-correctness label. Here, faithfulness refers to support from observable task information rather than causal faithfulness to latent model computation. Missing or inapplicable ratings remain missing rather than being converted to zero.

4.2 Outcome Verification

Final-answer correctness is rule-based when a reliable source-specific comparison exists. BFCL uses schema-aware matching of function names, required arguments, values, argument-order invariance, and enumerated aliases. GSM8K extracts the final numeric value and tests numerical equivalence, including integer/decimal and formatting variants. StrategyQA normalizes Boolean variants such as *yes/true* and *no/false*. HotpotQA uses conservative question-answer normalization over case, punctuation, articles, and whitespace, without permissive substring matching. Ambiguous cases remain unresolved or are reviewed; a judge cannot overwrite a definitive rule result.

4.3 Trajectory-Level Aggregation

Because trajectories contain different numbers of scored steps, we first average within each trajectory and then give trajectories equal weight. Let $K_{m,t}^{(d)}$ be the key steps with observed scores on dimension d for model m and task t, and $q_{m,t,i}^{(d)} \in \{0, 1, 2\}$. For nonempty $K_{m,t}^{(d)}$, define

$$Q_{m,t}^{(d)} = \frac{100}{2|K_{m,t}^{(d)}|} \sum_{i \in K_{m,t}^{(d)}} q_{m,t,i}^{(d)}, \tag{2}$$

then average trajectories for each model,

$$Q_m^{(d)} = \frac{1}{|T_m^{(d)}|} \sum_{t \in T_m^{(d)}} Q_{m,t}^{(d)}. \tag{3}$$

Here $T_m^{(d)}$ contains trajectories with an observed dimension score. Empty key-step sets are excluded, not assigned zero: eight retained trajectories lack key-step scores, leaving 1,178 for pooled step-metric analyses. This standard macro-aggregation prevents longer trajectories from receiving greater weight. We also report step-weighted sensitivity results.

For compact model-level summaries, we additionally define the prespecified Explanation Score:

$$\begin{aligned} E_m = {} & 0.25 Q_m^{(\text{faith})} + 0.20 Q_m^{(\text{valid})} + 0.20 Q_m^{(\text{specific})} \\ & + 0.20 Q_m^{(\text{diagnostic})} + 0.15 Q_m^{(\text{evidence})}. \end{aligned} \tag{4}$$

The weights are prespecified design choices, with slightly greater emphasis on task-grounded faithfulness and slightly lower emphasis on evidence alignment. The composite is a descriptive summary rather than a claim that the five dimensions form a single latent construct. We therefore report the five dimensions separately and include an equal-weight sensitivity analysis.

5 Validation and Analysis Protocol

A benchmark for explanation quality should validate the measurement process used to compare systems. We assess whether the rubric can be applied reproducibly and whether its dimensions respond to targeted changes. We then quantify the comparative resolution supported by the available tasks and annotations. Score provenance is retained throughout these analyses.

5.1 Annotation and Score Provenance

Trajectory-level annotations record answer correctness, overall trajectory quality, overall explanation quality, and a major failure type. Step-level annotations contain key-step selection, the five rubric dimensions, an error category, and concise notes. Missing or inapplicable ratings remain missing rather than being converted to zero.

Because scores are produced through different evaluation paths, every resolved label retains its provenance. The record indicates whether the value comes from annotator agreement, adjudication, a

deterministic rule, or an unresolved source. This prevents heterogeneous score sources from being treated as interchangeable ground truth.

5.2 Reliability and Resolution

Reliability is measured before disagreement resolution so that adjudication does not inflate observed agreement. We report exact agreement together with coefficients appropriate to the rating scale. Disagreements are resolved separately, and the original annotations remain available for audit.

The resulting reference subset is used to study annotation reliability and matched-system comparisons. Detailed agreement statistics and resolution counts are reported in Appendix D rather than treated as properties of the full hybrid corpus.

5.3 Construct Validation

Annotator agreement does not show that a rubric dimension measures the property it claims to measure. We therefore evaluate construct sensitivity with controlled perturbations. Each perturbation changes a targeted property while leaving the underlying task context fixed, allowing us to test whether the corresponding score moves in the prespecified direction.

An initial calibration set is used to refine ambiguous rubric anchors, after which the revised rubric is frozen and evaluated on a non-overlapping held-out replication. Candidate order, model identity, perturbation type, and baseline position are hidden from annotators. The perturbation pair is the unit of analysis, and support requires the prespecified directional criterion to hold independently for each annotator.

5.4 Statistical Analysis

Model-level intervals use 2,000 task-level bootstrap resamples. For pooled correlations between final-answer correctness and explanation-quality measures, task identifiers are resampled within source while keeping all systems' responses to the same task together. Step-level measures are aggregated within trajectories before resampling.

Repeated-system comparisons use tasks shared across systems, a Friedman omnibus test, and paired Wilcoxon tests with Holm correction. Matched-panel analyses use paired task resampling so that systems are compared on the same task composition. These analyses estimate the comparative resolution supported by the observed task set and annotation provenance. We therefore interpret small score differences through their uncertainty intervals rather than as definitive rank differences.

6 Results

6.1 Outcome Correctness and Trajectory Quality Diverge

Answer accuracy and trajectory quality do not induce the same model ordering. Figure 4 shows that the two measures diverge in both directions. Grok 4.3 has the largest accuracy-minus-explanation gap at 12.8 points, whereas GPT-4.1 Mini shows the largest reverse gap at 11.3 points. Because the full corpus mixes score provenance and retry histories, these differences are descriptive rather than estimates of first-pass capability.

6.2 Correctness Is Partially Associated with Explanation Quality

Across retained trajectories, answer correctness is associated with overall explanation quality (point-biserial $r = 0.425$, 95% source-stratified task-cluster CI $[0.351, 0.493]$, $n = 1{,}186$) and faithfulness ($r = 0.519$, CI $[0.438, 0.582]$, $n = 1{,}178$). Diagnostic utility ($r = -0.014$, CI $[-0.064, 0.046]$) and evidence alignment ($r = -0.046$, CI $[-0.095, 0.009]$) have intervals spanning zero. These are pooled associations across tasks and fixed systems, not causal effects or proof of independence. The small point estimates motivate measuring these dimensions separately; ceiling effects and restricted variation remain alternative explanations.



Figure 1: Five-dimensional explanation profiles across models. Each radar reports trajectory-balanced scores; the gray outline denotes the all-model median profile.

The confirmatory human-reference subset yields correlations of 0.294 for overall explanation quality, 0.311 for faithfulness, -0.091 for diagnostic utility, and 0.072 for evidence alignment (intervals in Appendix C). This subset differs in both composition and provenance from the full corpus, so differences between estimates cannot be attributed to annotation source alone.

6.3 Task Structure Shapes Explanation Profiles

Trajectory quality varies across task families. BFCL has the lowest answer accuracy (81.3%), whereas StrategyQA has the lowest mean Explanation Score (88.9). Appendix Figure 7 shows that these source effects vary across systems rather than producing a common difficulty ordering.

Figure 1 provides a complementary view across the five explanation dimensions. Models with similar aggregate scores can exhibit substantially different profiles, particularly in specificity, diagnostic utility, and evidence alignment. This supports reporting dimension-level behavior rather than relying on a single composite score. Appendix Figure 8 provides the corresponding numerical heatmap.

6.4 Comparative Resolution Is Limited for Near-Tied Systems

The benchmark detects systematic differences across systems, but these differences do not imply a precise total ordering. A Friedman test over the 61 task identifiers shared by all models detects differences in Explanation Score ($\chi^2 = 113.69$, $p = 1.08 \times 10^{-17}$). Of 105 Holm-adjusted pairwise Wilcoxon comparisons, 29 are below 0.05. These tests describe the current hybrid table and should not be interpreted independently of uncertainty and provenance.

Applying the completed human reference changes 1,156 score values relative to the pre-adjudication snapshot, while the full-table model ordering remains broadly similar (Spearman $\rho = 0.964$). The full-table and stratified human-reference orderings correlate $\rho = -0.118$; because that reference contains 12–20 trajectories per model and was designed for calibration coverage, this comparison motivates a stricter matched analysis rather than a replacement leaderboard.

To control task composition, we use a separately blinded panel of 16 shared tasks (four per source) across all 15 systems, yielding 240 trajectories and 1,035 available key steps. The panel was selected to maximize reuse of existing human labels, with provisional answer-discrimination variance used as a secondary criterion. It is therefore not a random holdout. One Claude 3.5 Haiku–StrategyQA

Table 2: Held-out construct-validity results. Deltas are perturbed minus baseline; intervals are pair-bootstrap 95% CIs.

Perturbation	Dimension	Expected	Pair mean [95% CI]	Both Ann.
Corrupt tool argument	Validity	$\downarrow$	-2.00 $[-2.00, -2.00]$	Yes
	Faithfulness	$\downarrow$	-1.50 $[-1.50, -1.50]$	Yes
	Evidence alignment	$\downarrow$	-2.00 $[-2.00, -2.00]$	Yes
Invalid evidence reference	Faithfulness	$\downarrow$	-1.00 $[-1.00, -1.00]$	Yes
	Evidence alignment	$\downarrow$	-2.00 $[-2.00, -2.00]$	Yes
Remove evidence refs.	Evidence alignment	$\downarrow$	-1.33 $[-1.44, -1.21]$	Yes
Shuffle trajectory	Trajectory quality	$\downarrow$	-2.31 $[-2.75, -1.84]$	Yes
	Explanation quality	$\downarrow$	-1.94 $[-2.31, -1.56]$	Yes
Unsupported claim	Validity	$\downarrow$	-1.00 $[-1.00, -1.00]$	Yes
	Faithfulness	$\downarrow$	-1.00 $[-1.00, -1.00]$	Yes
	Evidence alignment	$\downarrow$	-1.00 $[-1.00, -1.00]$	Yes
Generic verbosity	Specificity	≤ 0	0.00 $[0.00, 0.00]$	Yes
	Diagnostic utility	≤ 0	-0.50 $[-0.50, -0.50]$	Yes
Wrong calculation	Validity	$\downarrow$	-1.88 $[-2.00, -1.62]$	Yes
	Faithfulness	$\downarrow$	-0.94 $[-1.00, -0.81]$	Yes
	Diagnostic utility	≈ 0	0.00 $[0.00, 0.00]$	Yes

trajectory contains no scored key step. Step comparisons use the 15 tasks with scores for all models (995 key steps), averaging within source and weighting sources equally, without imputation. The matched-human and full-hybrid EXPLANATION SCORE rankings have Spearman $\rho = -0.046$ (score Pearson $r = -0.218$), whereas answer scores correlate at $r = 0.735$. This cross-panel comparison still mixes composition and provenance effects. GPT-5.5 has the highest observed matched score, 96.8 (95% paired task-bootstrap CI [95.7, 97.8]), but a 38.5% bootstrap top-1 frequency; none of the ten top-five pairwise intervals excludes zero. The matched-panel intervals are shown in Appendix Figure 9.

Source-stratified subsampling of the 61 tasks with complete model coverage reaches an internal rank-stability criterion at 24 tasks, but close-model comparisons remain underpowered: eight of ten top-five pairwise calculations require more than 200 tasks. The benchmark therefore supports dimension-level and source-level comparisons more strongly than a finely resolved total order among near-tied systems.

6.5 CONTROLLED PERTURBATIONS SUPPORT RUBRIC SENSITIVITY

We test whether each rubric dimension responds when its target property is selectively altered. An initial blinded calibration set supported 8 of 16 predictions and motivated eight locked anchor revisions. We then evaluated the frozen rubric on a disjoint held-out set with balanced, hidden candidate positions.

Table 2 summarizes the held-out results. All 16 prespecified predictions satisfy the locked both-annotator criterion: targeted corruptions reduce their intended dimensions, generic verbosity does not improve specificity or diagnostic utility, and wrong calculations preserve diagnostic utility within the prespecified invariance margin.

Preferred-candidate agreement is 94.2% for step pairs ($\kappa = 0.891$) and 93.8% for trajectory pairs ($\kappa = 0.882$). Appendix Figures 2 and 6 provide annotator-specific and pair-mean effects.

6.6 OUTCOME VERIFICATION DOES NOT REPLACE TRAJECTORY EVALUATION

Within the confirmatory human-reference subset, trajectories marked valid by deterministic auto-evaluation ($n = 219$) have 92.7% human answer accuracy, while auto-eval-invalid trajectories ($n = 8$) have 0%. Auto-evaluation is missing for 17 sampled trajectories and disagrees with the human answer label in 16 cases. Rule-based evaluation is therefore valuable for verifiable final answers, but it cannot establish whether evidence is faithful, a step is diagnostically useful, or a tool call is aligned with the surrounding explanation.

7 DISCUSSION

Trajectory quality complements outcome correctness. Our results support treating observable trajectory quality as a distinct evaluation target. Final correctness does not reveal whether intermediate claims are adequately supported, and failed trajectories can still expose localized evidence or decision errors. For auditing and diagnosis, the observable record therefore provides information that outcome accuracy alone cannot supply.

Explanation quality is multidimensional and task dependent. The five dimensions represent different operational properties and should not be interpreted as a single latent construct. A step can be correct yet unsupported, or incorrect yet diagnostically useful. Task structure also changes which properties become most informative. We therefore view dimension profiles and source-specific behavior as more informative than a single composite score.

Agent evaluation should expose measurement uncertainty. Near-tied systems cannot always be ordered reliably. This motivates reporting uncertainty and score provenance alongside aggregate results. Agreement analysis evaluates reproducibility, controlled perturbations evaluate construct sensitivity, and matched analyses characterize comparative resolution. Together, these results position AGENTEX-BENCH as a framework for auditing observable trajectories rather than as a definitive model leaderboard.

8 CONCLUSION AND FUTURE DIRECTIONS

We introduced AGENTEX-BENCH, a measurement framework for evaluating observable agent trajectories as auditable explanatory artifacts. Across heterogeneous reasoning and tool-use tasks, the benchmark separates final-task correctness from five complementary properties of the emitted trajectory: validity, task-grounded faithfulness, specificity, diagnostic utility, and evidence alignment. Our results show that outcome correctness and trajectory quality provide related but non-equivalent information, while dimension-level profiles reveal distinctions that aggregate scores can obscure. Matched analyses further show that fine-grained model ordering should be interpreted with explicit uncertainty.

The current study has several limitations. The benchmark contains 80 tasks and does not cover long-horizon interactive agent execution. Some retained outputs also underwent gold-informed validation retries, and the full-corpus analysis combines multiple score provenance sources. These factors limit the interpretation of fine-grained system differences and first-pass capability.

More broadly, AGENTEX-BENCH frames trajectory evaluation as a measurement problem rather than a leaderboard problem. Future work can expand task and agent coverage, use uniformly gold-blind generation, strengthen annotation reliability, and test whether trajectory-quality signals improve downstream auditing, debugging, and automated failure analysis.

REPRODUCIBILITY STATEMENT

We provide task-selection criteria, trajectory schemas, annotation and adjudication procedures, bootstrap specifications, controlled-perturbation protocols, and additional robustness analyses in the appendix. The accompanying repository contains the scripts used for cleaning, aggregation, statistical analysis, and visualization, together with provenance and audit records.

AI USE STATEMENT

Generative AI tools were used to assist with code refinement and debugging, to support the search and identification of relevant literature, and to improve the clarity and wording of selected portions of the manuscript. All experimental design, data analysis, interpretation of results, citation selection, and final manuscript content were reviewed and verified by the authors.

REFERENCES

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Mądry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025. URL https://arxiv.org/abs/2503.11926.

Shraddha Barke, Arnav Goyal, Alind Khare, Avaljot Singh, Suman Nath, and Chetan Bansal. AgentRx: Diagnosing AI agent failures from execution trajectories. *arXiv preprint arXiv:2602.02475*, 2026. URL https://arxiv.org/abs/2602.02475.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025. URL https://arxiv.org/abs/2506.07982.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. URL https://arxiv.org/abs/2505.05410.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Darshan Deshpande, Varun Gangal, Hersh Mehta, Jitin Krishnan, Anand Kannappan, and Rebecca Qian. TRAIL: Trace reasoning and agentic issue localization. *arXiv preprint arXiv:2505.08638*, 2025. URL https://arxiv.org/abs/2505.08638.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Shengda Fan, Xuyan Ye, Yupeng Huo, Zhi-Yuan Chen, Yiju Guo, Shenzhi Yang, Wenkai Yang, Shuqi Ye, Jingwen Chen, Haotian Chen, Xin Cong, and Yankai Lin. AgentProcessBench: Diagnosing step-level process quality in tool-using agents. *arXiv preprint arXiv:2603.14465*, 2026. URL https://arxiv.org/abs/2603.14465.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21/.

Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. Monitoring monitorability. *arXiv preprint arXiv:2512.18311*, 2025. URL https://arxiv.org/abs/2512.18311.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The FACTS grounding leaderboard: Benchmarking LLMs' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025. URL https://arxiv.org/abs/2501.03200.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025. URL https://arxiv.org/abs/2507.11473.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

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.

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. URL https://arxiv.org/abs/2504.08942.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hanna Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ea4fe0a56d02c93401902b5b4c6b12da-Abstract-Conference.html.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1230. URL https://aclanthology.org/2025.acl-long.1230/.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9e720fce64f91114c49cfd640d821da3-Abstract-Conference.html.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In

Advances in Neural Information Processing Systems, volume 36, 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract.html.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9935–9960. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.504. URL https://aclanthology.org/2025.emnlp-main.504/.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/yao2023iclr-react/.

Asaf Yehudai, Lilach Eden, and Michal Shmueli-Scheuer. Agentic CLEAR: Automating multi-level evaluation of LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 750–763. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-demo.74. URL https://aclanthology.org/2026.acl-demo.74/.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76583–76599. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25cq.html.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/.

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 Positioning Relative to Recent Benchmarks

Table 3 contrasts evaluation targets rather than claiming that related systems lack secondary capabilities. AgentEx-Bench remains a benchmark contribution, but neither intermediate-step evaluation nor human error annotation is new by itself. The intended contribution is their combination with five explanation dimensions, evidence linkage, and provenance-aware human validation across heterogeneous emitted trajectories.

Table 3: Selected nearby work and the distinction relevant to AgentEx-Bench.

Work	Evaluation target	Relation to this benchmark
ProcessBench (Zheng et al., 2025)	Mathematical process errors	Correctness is one of our five dimensions.
PRMBench (Song et al., 2025)	Process reward sensitivity, soundness, simplicity	We assess emitted explanations across four task families.
AgentProcessBench (Fan et al., 2026)	Effectiveness of tool-agent steps	Our rubric separates support, specificity, utility, and reference alignment.
AgentRewardBench (Lù et al., 2025)	Evaluators of web trajectories	We separate human references from assisted annotations.
TRAIL / AgentRx (Deshpande et al., 2025; Barke et al., 2026)	Error localization in traces	We score explanation dimensions in both successful and failed outputs.
Agentic CLEAR (Yehudai et al., 2026)	Automated system-, trace-, and node-level feedback	Our validation target is an explicit ordinal rubric.

B Generation-Provenance Audit

The revision audit joins every retained annotation to saved JSON by model and task, additionally checking the final answer. Of 1,186 joins, 1,148 match the final-answer string directly and 38 match after representation normalization (whitespace, Boolean case, or JSON serialization). Compatible duplicate saved files are retained in the audit instead of choosing an arbitrary path. Input hashes and candidate paths are preserved outside the manuscript.

Human reannotation validates the retained output, not the model's access conditions when producing it. We neither infer first-pass performance from these outputs nor delete the retried cases and label the remainder an unbiased evaluation. Recovering unassisted accuracy requires complete first-attempt artifacts or a new uniformly gold-blind run. Our corpus-level findings are conditional on the existing retained-output protocol; this limits model comparisons and generalization of accuracy–explanation relationships.

C Task-Cluster Correlation Sensitivity

The pooled correlations between final-answer correctness and explanation-quality measures share task identities across systems, so individual model–task rows should not be treated as independent observations. We therefore assess uncertainty using a source-stratified task-cluster bootstrap.

For each explanation metric, we retain rows with observed values for both that metric and answer correctness, group observations by source and task identifier, and resample task clusters with replacement within each source. All available system responses associated with a sampled task receive the same bootstrap multiplicity. The statistic is the pooled point-biserial correlation, conditional on the evaluated systems. We use 2,000 bootstrap draws with seed 42.

The human-reference subset contains 78 distinct tasks rather than 244 independent task clusters, so its intervals are computed using the same task-level resampling principle. These intervals capture uncertainty from task sampling under the observed system set, but do not incorporate uncertainty from annotation, task selection, or score provenance.

Table 4 reports the resulting estimates. Intervals containing zero should not be interpreted as evidence of equivalence, and the pooled correlations do not identify within-system effects or causal relationships. The hybrid corpus and human-reference subset also differ in composition and provenance, so differences between their estimates cannot be attributed to annotation source alone.

Table 4: Pooled correlations between final-answer correctness and explanation-quality measures with source-stratified task-cluster bootstrap intervals.

Panel	Explanation metric	n	r	95% CI
Hybrid	Overall quality	1,186	0.425	[0.351, 0.493]
Hybrid	Faithfulness	1,178	0.519	[0.438, 0.582]
Hybrid	Diagnostic utility	1,178	−0.014	[−0.064, 0.046]
Hybrid	Evidence alignment	1,178	−0.046	[−0.095, 0.009]
Human reference	Overall quality	244	0.294	[0.122, 0.453]
Human reference	Faithfulness	243	0.311	[0.059, 0.496]
Human reference	Diagnostic utility	243	−0.091	[−0.203, 0.037]
Human reference	Evidence alignment	243	0.072	[−0.067, 0.213]

D Validation Details and Future Extensions

D.1 Task Sampling and Coverage

The 20 GSM8K tasks were sampled from records with a nonempty question, an explicit gold-final marker, and question length between 30 and 1,200 characters. StrategyQA records required a Boolean answer, at least two supplied facts, and two to five decomposition steps. HotpotQA records required at least two supporting facts spanning at least two documents, 4–12 context documents, and a non-Boolean answer. BFCL contributes five tasks from each of four categories—simple Python, multiple functions, parallel calls, and parallel multiple calls—after requiring a usable prompt, function schema, and gold tool call. The retained task manifest fixes the analyzed task set. Available script defaults do not by themselves verify every historical sampling invocation. These filters do not make the 80-task set statistically representative of each source's full distribution.

D.2 Human Agreement Details

Table 5: Agreement between two independent human annotators on the blinded confirmatory sample. For ordinal metrics, κ is quadratic weighted; for categorical metrics, it is Cohen's κ.

Level	Metric	n	Exact	κ	MAE	Within 1	ρ
Trajectory	Answer correctness	244	89.8	0.587	–	–	–
Trajectory	Trajectory quality	244	41.4	0.358	0.717	88.1	0.331
Trajectory	Explanation quality	244	42.6	0.226	0.660	91.4	0.276
Trajectory	Major failure type	244	63.9	0.252	–	–	–
Step	Key-step selection	1397	100.0	1.000	–	–	–
Step	Step validity	1061	84.0	0.301	0.211	94.9	0.268
Step	Faithfulness	1061	80.6	0.224	0.198	99.6	0.221
Step	Specificity	1061	77.7	0.074	0.224	99.9	0.154
Step	Diagnostic utility	1061	79.7	0.512	0.203	100.0	0.547
Step	Evidence alignment	1061	84.4	0.459	0.166	99.1	0.543
Step	Step error type	1061	86.6	0.228	–	–	–

Independent human agreement and ranking robustness. Two independent annotations and third-human adjudication are complete for both the stratified confirmatory sample and the 16-task

Table 6: Primary outcomes in the held-out construct-validity replication using Annotator A and Annotator B. Deltas are perturbed minus baseline. The pair-mean column reports a pair-bootstrap 95% confidence interval.

Perturbation	Dimension	Expected	n	Ann. A	Ann. B	Pair mean [95% CI]	Both
Corrupt tool argument	Validity	$\downarrow$	16	-2.00	-2.00	-2.00 [-2.00, -2.00]	Yes
Corrupt tool argument	Faithfulness	$\downarrow$	16	-2.00	-1.00	-1.50 [-1.50, -1.50]	Yes
Corrupt tool argument	Evidence alignment	$\downarrow$	16	-2.00	-2.00	-2.00 [-2.00, -2.00]	Yes
Invalid evidence reference	Faithfulness	$\downarrow$	16	-1.00	-1.00	-1.00 [-1.00, -1.00]	Yes
Invalid evidence reference	Evidence alignment	$\downarrow$	16	-2.00	-2.00	-2.00 [-2.00, -2.00]	Yes
Remove evidence references	Evidence alignment	$\downarrow$	24	-1.92	-0.75	-1.33 [-1.44, -1.21]	Yes
Shuffle trajectory	Trajectory quality	$\downarrow$	16	-2.31	-2.31	-2.31 [-2.75, -1.84]	Yes
Shuffle trajectory	Explanation quality	$\downarrow$	16	-1.88	-2.00	-1.94 [-2.31, -1.56]	Yes
Unsupported claim	Validity	$\downarrow$	24	-1.00	-1.00	-1.00 [-1.00, -1.00]	Yes
Unsupported claim	Faithfulness	$\downarrow$	24	-1.00	-1.00	-1.00 [-1.00, -1.00]	Yes
Unsupported claim	Evidence alignment	$\downarrow$	24	-1.00	-1.00	-1.00 [-1.00, -1.00]	Yes
Generic verbosity	Specificity	$\lesssim 0$	24	+0.00	+0.00	+0.00 [+0.00, +0.00]	Yes
Generic verbosity	Diagnostic utility	$\lesssim 0$	24	-1.00	+0.00	-0.50 [-0.50, -0.50]	Yes
Wrong calculation	Validity	$\downarrow$	16	-1.88	-1.88	-1.88 [-2.00, -1.62]	Yes
Wrong calculation	Faithfulness	$\downarrow$	16	-0.94	-0.94	-0.94 [-1.00, -0.81]	Yes
Wrong calculation	Diagnostic utility	≈ 0	16	+0.00	+0.00	+0.00 [+0.00, +0.00]	Yes

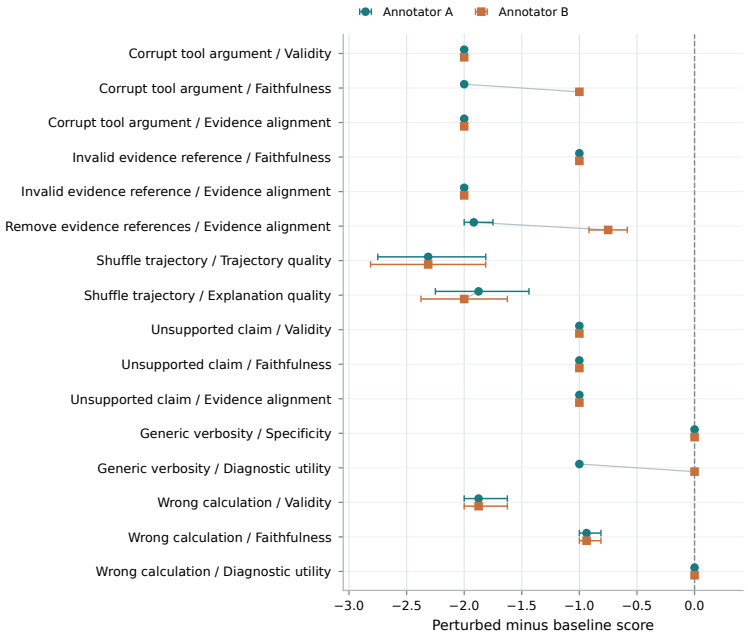

Figure 2: Annotator-specific held-out perturbation effects for A and B.

source-balanced matched panel. The matched analysis does not support a precise model order: full-table versus matched rank correlation is -0.046, and all top-five pairwise intervals include zero. We retain these results as a robustness analysis and report fine-grained ranks as exploratory.

LLM-judge calibration. The repository includes a source-aware, schema-validated judge interface with caching, retries, prompt-injection defenses, and audit logs. Three zero-cost local pilots failed the locked non-ceiling detection gate and are excluded from benchmark results. Future work will evaluate stronger judge configurations against the adjudicated reference before promoting automated explanation scores.

Prompt and evaluator ablations. The present study fixes the requested format, but does not isolate generation-retry or annotation effects. A broader study can compare uniformly gold-blind direct-reasoning, plan-then-solve, and evidence/tool-aware protocols, alongside rule-only, generic-judge, rubric-aware, and evidence-aware evaluators.

Scale and agentic breadth. Task-count stability curves show that the 61-task shared panel supports broad ordering patterns while close-model comparisons remain underpowered. Future releases can expand to 160–200 source-balanced tasks, define practical-equivalence bands, and add reproducible long-horizon interactive tasks. These extensions increase the scope and statistical resolution of the project but are not prerequisites for the claims evaluated here.

E Reproducibility

The repository contains scripts for inventory, cleaning, annotation resolution, aggregation, controlled perturbations, statistics, visualization, and submission validation. Analyses use 2,000 resamples with seed 42: per-model task trajectories, shared tasks for model comparisons and pooled correlations, or pairs for perturbations. Revision audits retain input hashes and per-record generation joins. Earlier pooled correctness–explanation correlation intervals resampled rows independently. This revision replaces those intervals with the task-cluster estimates in Appendix C, without changing point estimates or original outputs.

F Full Rubric Anchors

Table 7: Step-level rubric anchors.

Dimension	0	1	2
Validity	Incorrect step.	Partially correct with a material omission or issue.	Correct for the task state.
Faithfulness	Contradicted, unsupported, or fabricated.	Partly supported but insufficiently grounded.	Clearly supported by task input, evidence, schema, facts, or mathematics.
Specificity	Generic or content-free.	Includes some task details but remains underspecified.	Names concrete entities, values, documents, functions, arguments, or facts.
Diagnostic utility	Does not help explain success or failure.	Provides partial diagnostic information.	Exposes a key decision, basis, or error source.
Evidence alignment	Missing or irrelevant support when support is required.	Partially aligned or incomplete reference.	References or tool calls directly support the step; evidence-free steps are allowed when evidence is unnecessary.

G Generation Protocol and Data Schemas

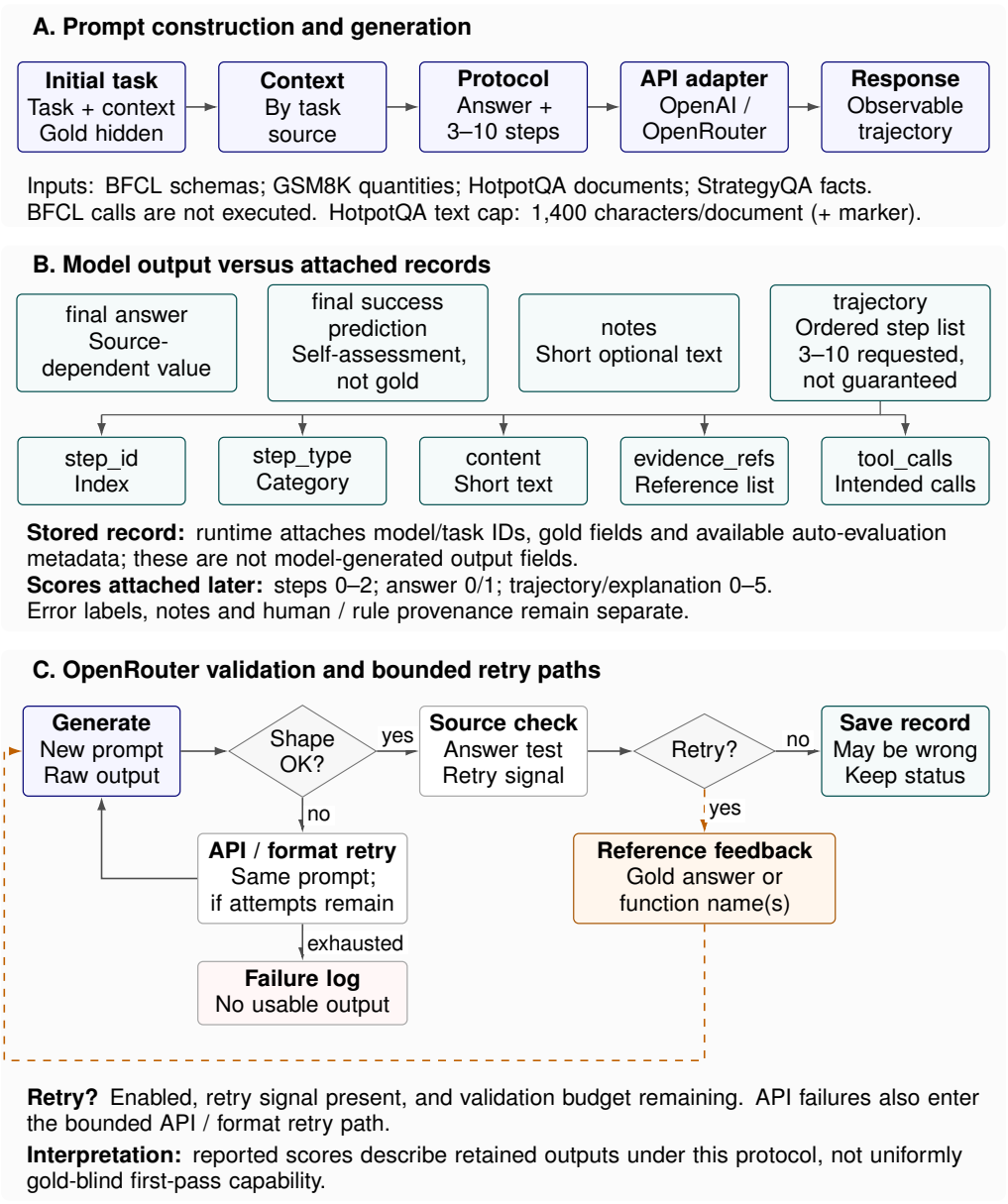

Figure 3: Generation and data-schema overview for AGENTEX-BENCH. **A:** source-aware prompting with adapter-specific wording. **B:** generated fields versus runtime and scoring records. **C:** OpenRouter retry handling; solid arrows show processing and API/format recovery, while dashed arrows show reference-informed regeneration. A saved response need not pass the source check. The diagram summarizes the protocol rather than reproducing verbatim prompts.

Table 8: Full retained-corpus results with mixed score provenance. Acc. is retained-output accuracy under a protocol permitting gold-informed retries, not first-pass capability. Step metrics use trajectory-balanced aggregation on 0–100; Expl. is the weighted Explanation Score. Bold marks observed maxima.

Model	n_T	n_S	Acc.	Expl.	Val.	Faith.	Spec.	Diag.	Evid.
Claude Opus 4.5	80	415	96.2	**95.0**	98.7	98.9	**99.8**	80.4	**96.6**
Nemotron 3 Ultra	69	272	94.2	93.2	98.1	91.3	99.1	**92.2**	83.1
DeepSeek R1	80	355	96.2	92.7	98.6	98.3	97.5	81.6	83.6
OpenRouter Free	77	254	88.3	92.6	97.7	92.1	98.1	90.5	82.5
Mistral Large	80	395	97.5	92.6	96.7	92.0	98.7	91.0	82.1
Gemini 2.5 Flash	80	372	96.2	92.2	98.4	97.3	97.1	81.1	83.5
GPT-4.1 Mini	80	344	80.0	91.3	94.8	87.5	99.7	91.2	81.7
DeepSeek Chat V3.1	80	398	88.8	91.2	93.7	93.2	98.0	79.4	91.5
Qwen 2.5 72B	80	310	96.2	91.0	97.9	88.1	99.1	86.2	82.4
GPT-5.5	80	312	**100.0**	88.0	**100.0**	**100.0**	61.4	84.1	92.8
MiniMax M3	80	372	97.5	87.9	**100.0**	99.9	71.8	77.8	86.7
Grok 4.3	80	319	**100.0**	87.2	99.9	99.9	60.6	82.4	90.8
Llama 3.3 70B	80	334	98.8	86.8	98.1	99.3	62.5	83.7	87.4
Claude 3.5 Haiku	80	280	82.5	86.2	90.7	90.9	89.6	69.2	90.3
Gemma 3 27B	80	349	91.2	85.8	97.1	97.2	58.0	88.4	85.2

Table 9: Source-level profile, reported as the unweighted mean of model-level source scores. Counts are observed records.

Source	n_T	n_S	Acc.	Expl.	Val.	Faith.	Spec.	Diag.	Evid.
BFCL	296	1426	81.3	89.0	92.7	93.7	81.0	84.5	92.8
GSM8K	297	1314	97.7	89.4	98.9	93.0	86.8	86.8	77.6
HotpotQA	297	1338	97.7	93.7	98.5	98.3	90.3	87.7	92.1
StrategyQA	296	1003	97.7	88.9	99.6	95.2	86.1	76.6	84.2

H Trajectory Schema

```
{
  "model": "hidden-from-judge",
  "task_id": "bfcl_0001",
  "source": "BFCL",
  "final_answer": "...",
  "trajectory": [
    {
      "step_id": 1,
      "step_type": "select_function",
      "content": "...",
      "evidence_refs": ["function_schema"],
      "tool_calls": []
    }
  ]
}
```

I Supplementary Figures

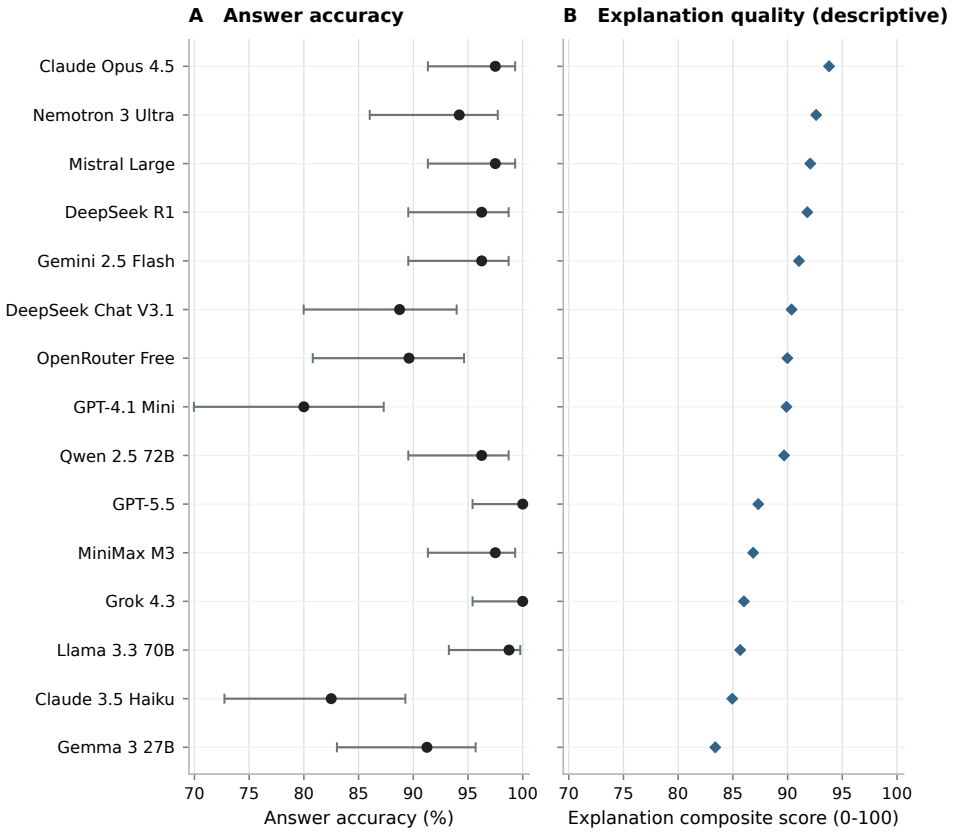

Figure 4: Retained-output accuracy and trajectory-balanced Explanation Score across 15 systems. The full corpus uses mixed score provenance and includes gold-informed validation retries. Values do not estimate unassisted first-pass performance.

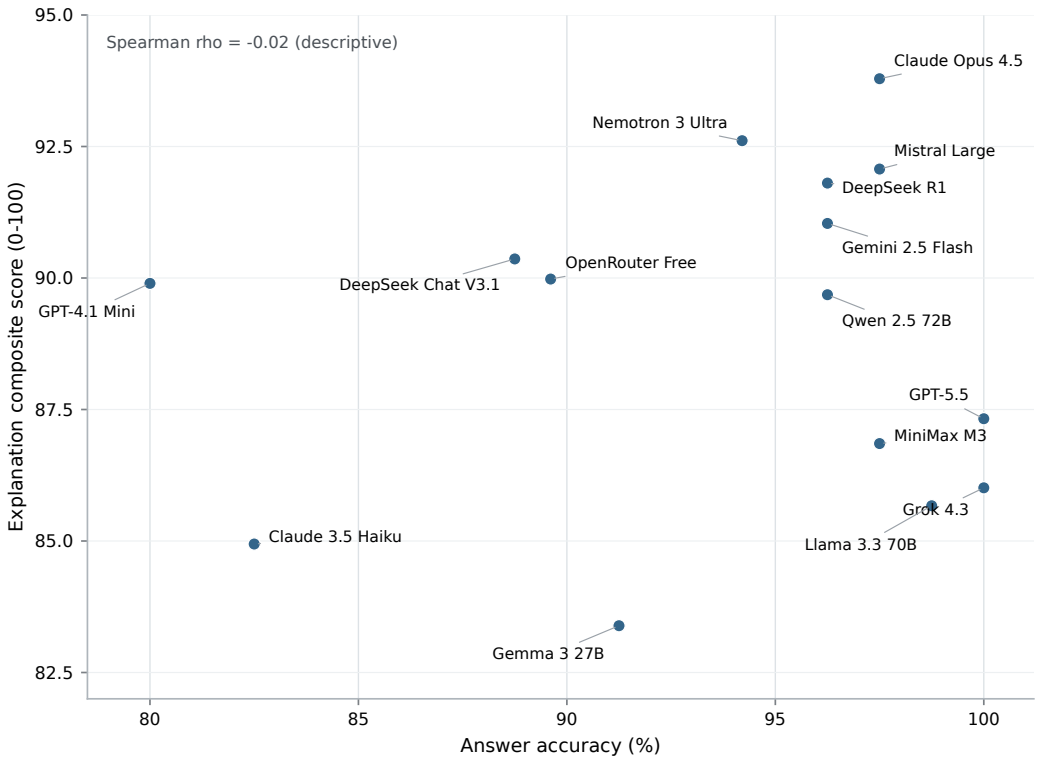

Figure 5: Model-level answer accuracy versus Explanation Score.

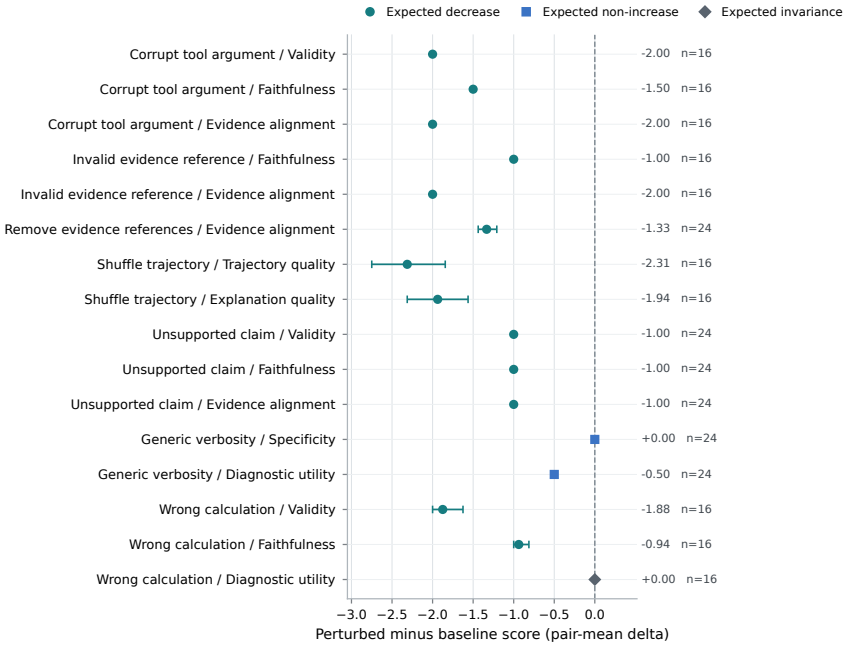

Figure 6: Held-out pair-mean changes under controlled perturbations. Bars are pair-bootstrap 95% confidence intervals. All 16 predictions satisfy the locked criterion independently for both qualified annotators.

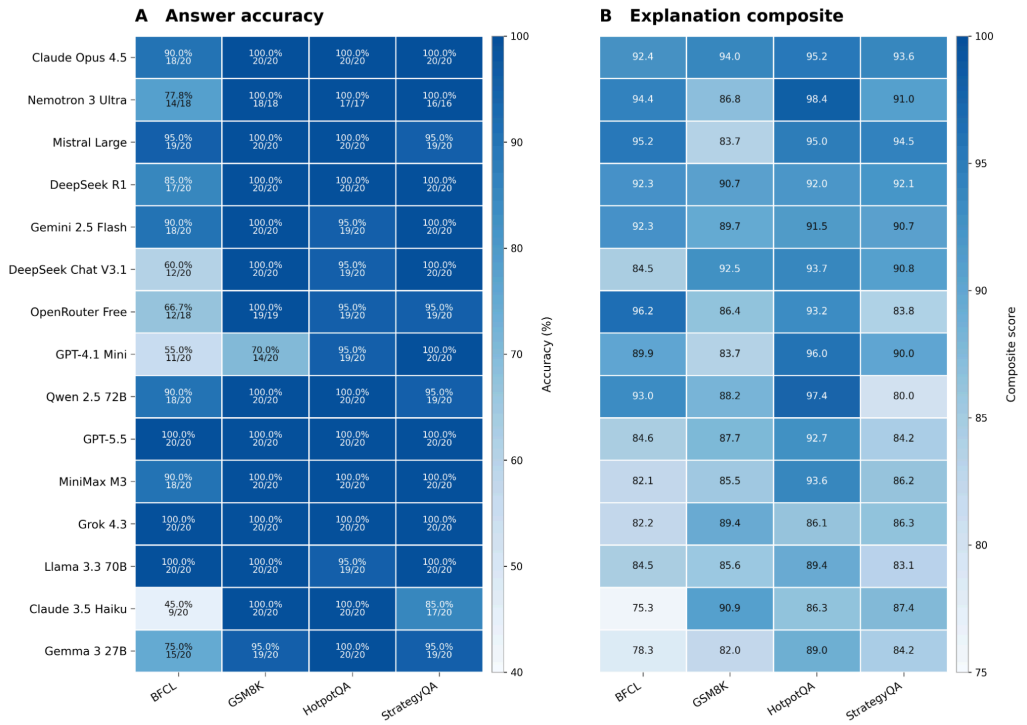

Figure 7: Answer accuracy and trajectory-balanced Explanation Score by model and source. Sample counts shown in panel A apply to the corresponding cells in both panels.

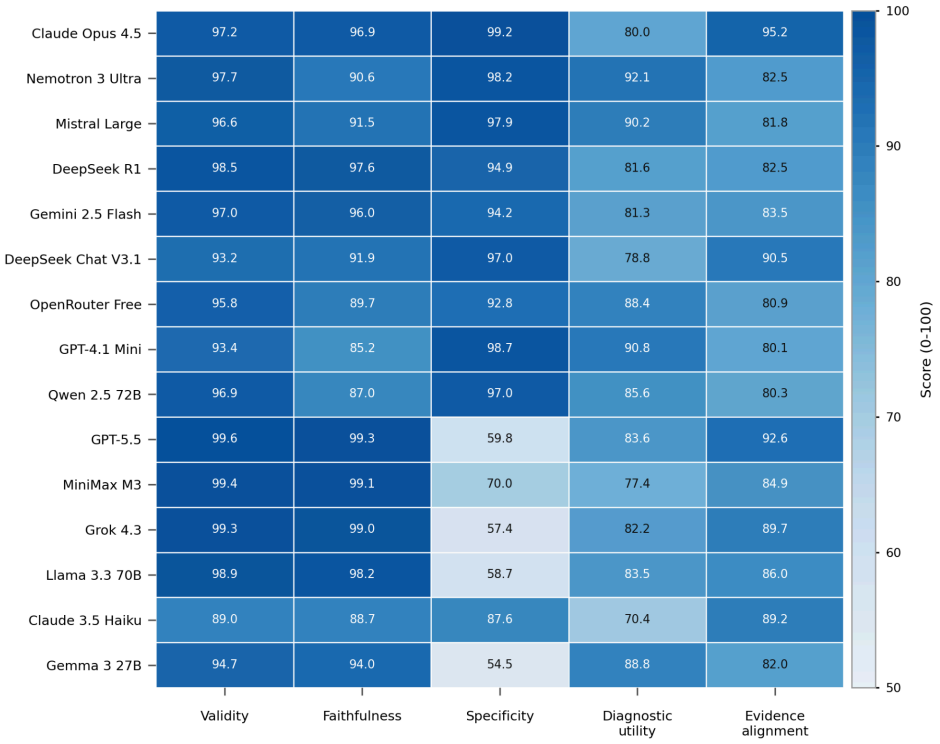

Figure 8: Cross-model comparison of the five explanation dimensions. Cells report trajectory-balanced scores on a 0–100 scale.

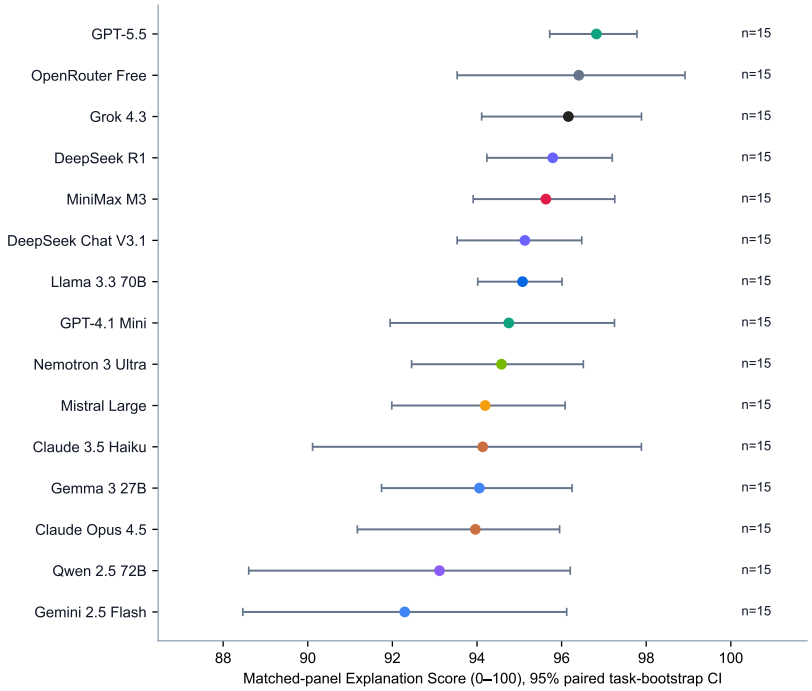

Figure 9: Human-adjudicated matched-panel Explanation Scores for retained trajectories. Points are source-equal means over 15 shared tasks; bars are 95% paired task-bootstrap intervals.

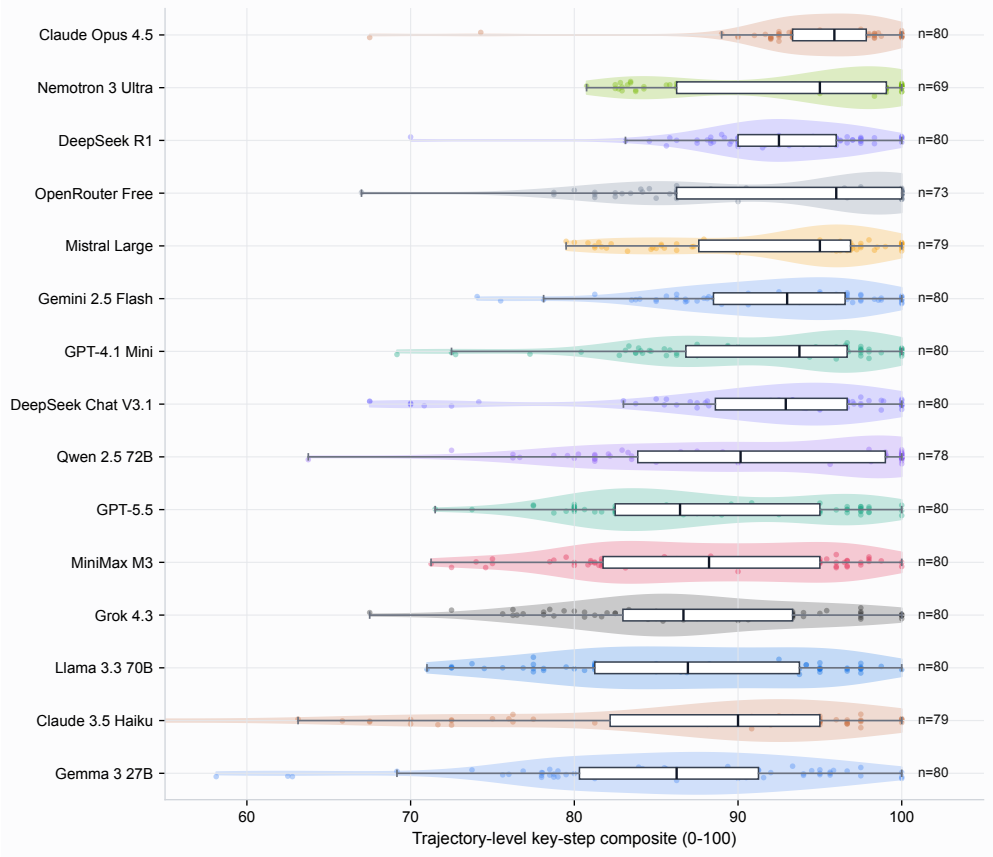

Figure 10: Trajectory-level Explanation Score distributions. Each point is one trajectory after averaging its scored key steps; violins and embedded boxplots expose variability and outliers hidden by model means.

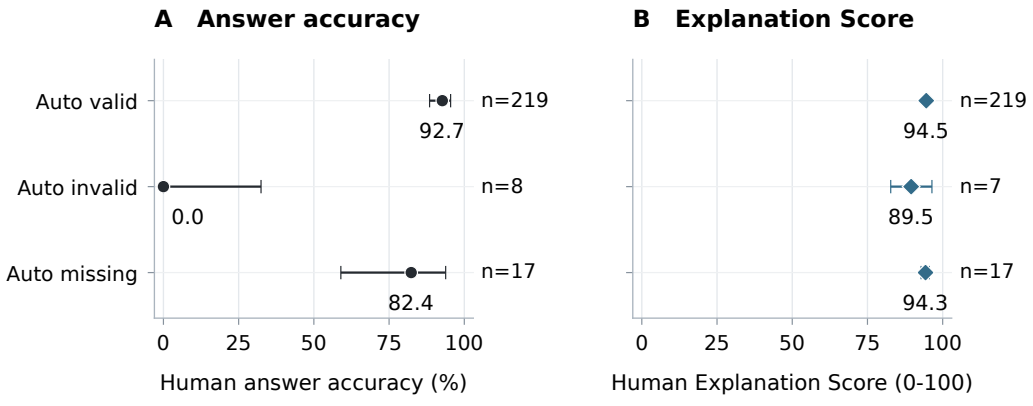

Figure 11: Rule-based auto-evaluation status versus judgments in the confirmatory human-adjudicated subset.