LONGJUDGE-BENCH: A LONG-HORIZON AGENT JUDGE BENCHMARK FOR INTERVENTION DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon LLM agents execute stateful workflows in which tool actions alter persistent environments and determine delayed outcomes. Supervisors must therefore choose what to do before completion, yet existing judge benchmarks mainly score completed responses or trajectories and do not directly evaluate which executable intervention best preserves future success. We introduce LONGJUDGE-BENCH, a consequence-grounded benchmark with 70 unfinished checkpoints from 35 AppWorld tasks, balanced across five domains and seven intervention families. For each case, we execute all seven actions on five matched frozen scenario–seed realizations under a shared executor, checker, budget, and continuation contract, producing 490 empirical action values from 2,450 outcomes. Judges select from pre-outcome evidence, and we separately measure response validity, retained task success (TS), exact reference-action accuracy (AA), and regret. Across 13 models under the Original interface, 90.22% of responses are valid, but mean TS is 53.47% and mean AA is 36.37%, exposing distinct gaps in consequence preservation and reference recovery. Under Declared semantics, observed TS and AA rise for all six tested models, while history and position controls show that AA is consistently more sensitive than TS to evidence representation. These results distinguish valid outputs, consequence-preserving choices, and exact reference recovery as separate capabilities. LONGJUDGE-BENCH provides a controlled basis for evaluating actionable supervision in long-horizon agents.

1 INTRODUCTION

LLM agents increasingly execute long-horizon workflows through tools that modify persistent application state (Yao et al., 2022; Yang et al., 2024). AppWorld and OSWorld evaluate such agents with executable tasks and state-based checks, while HORIZON and OSWorld 2.0 extend evaluation toward longer, stateful workflows (Trivedi et al., 2024; Xie et al., 2024; Wang et al., 2026; Yuan et al., 2026). As execution length grows, failures become harder to localize, costlier to correct, and more likely to invalidate completed steps. Agent judges thus become essential for detecting failures and guiding interventions before errors propagate or recoverable progress is lost.

Agent judges often assess actions or completed trajectories. Long-horizon tasks change this role as errors compound over time: a judge must jointly weigh the unfinished goal, full history, current state, and a pending action whose alternatives may each retain valid progress yet trade off key downstream goals. Blocking, revision, rollback, and escalation thus define competing paths, not equivalent fixes (Vasudev et al., 2026). We ask: ***Given an unfinished long-horizon trajectory with competing outcomes, which intervention best preserves downstream task success?***

Existing benchmarks offer partial foundations. JudgeBench and AgentRewardBench assess responses or completed trajectories, whereas AJ-Bench supports evidence gathering before a verdict (Tan et al., 2025; Lù et al., 2025; Shi et al., 2026). Online monitoring and recovery methods intervene mid-execution, but confound judgment with correction and executor behavior (Liu et al., 2026b; Yang et al., 2026). Counterfactual recoverability and same-prefix analyses examine branching choices (Jiang et al., 2026; Zhang et al., 2026a). Yet Agent-as-a-Judge lacks a long-horizon benchmark for consequence-grounded intervention selection at unfinished checkpoints.

To address this gap, we introduce LONGJUDGE-BENCH, a long-horizon Agent-as-a-Judge benchmark for intervention decisions (Figure 1). It comprises 70 checkpoints from 35 AppWorld source

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.

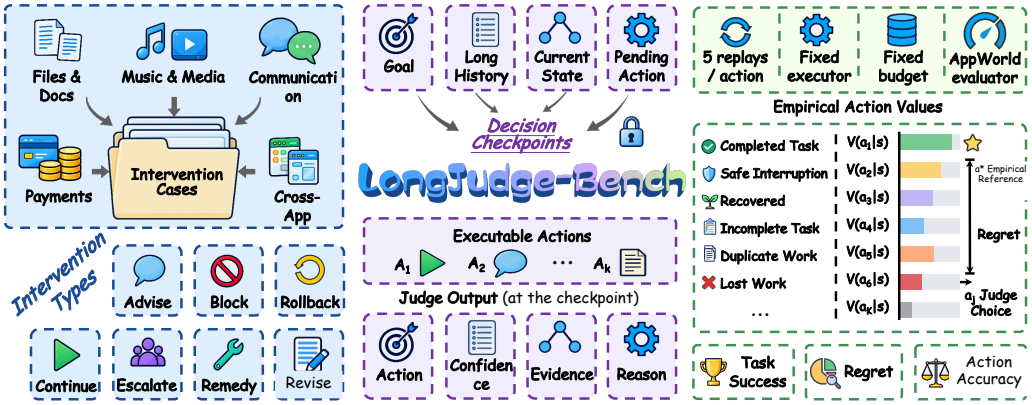


Figure 1: **Overview of LONGJUDGE-BENCH.** Across five AppWorld domains, judges select one of seven executable interventions at 70 checkpoints. Five frozen outcomes per action yield empirical values used to compute task success, regret, and exact action accuracy.

IDs, balanced across five domains and seven intervention families. Each case pairs an unfinished trajectory with seven interventions. Each candidate is executed on five matched frozen scenario–seed realizations with a fixed executor, budget, continuation contract, and task checker, producing 490 action values from 2,450 outcomes and a judge-independent empirical reference. This setup keeps choice and execution apart while retaining executable, state-based effects for every action.

Experiments reveal a gap between producing valid responses and selecting interventions that best preserve task value. Declared semantics consistently improve observed scores, whereas Expanded introduces composite difficulty. History controls reveal evidence dependence, while observed position profiles are non-monotonic. Together, the findings identify output validity, consequence quality, exact selection, semantic grounding, and context use as distinct dimensions of intervention judging. Our contributions are:

- We introduce **LONGJUDGE-BENCH**, which formalizes intervention judging at unfinished agent checkpoints through a judge-visible evidence interface, seven executable actions, and an explicit continuation contract separating observable evidence from hidden future outcomes.
- We develop a **matched branch execution protocol** that evaluates candidates from the same frozen state with a fixed executor, budget, continuation contract, and task checker, deriving 490 empirical action values from 2,450 branch outcomes.
- We provide a **decision- and consequence-based evaluation** combining exact action accuracy, task success, regret, and invalid-response rate with controlled analyses of interface specification, history availability, and evidence placement across model subsets.

2 RELATED WORK

LLM-as-a-Judge. LLM-as-a-Judge uses language models to assess responses or completed trajectories. JudgeBench examines response-pair judgments; AgentRewardBench, MobileJudgeBench, and OSReward evaluate web, mobile, and computer-use trajectories (Tan et al., 2025; Lù et al., 2025; Wan et al., 2026; Sun et al., 2026). LongJudgeBench targets long-form output evaluation (Chen et al., 2026b). Related work examines evaluator uncertainty, safety risks, and failure attribution (Chen et al., 2026a; Ruan et al., 2024; Zhang et al., 2025; 2026b; Zhu et al., 2025; Liu et al., 2026a). These settings remain retrospective despite differing evidence and targets: judgment follows an observed output or trajectory. LONGJUDGE-BENCH instead requires intervention selection before candidate branch outcomes are observed.

Agent-as-a-Judge. Agent judges use tools to inspect environments. Agent-as-a-Judge directly assesses coding agents through traces and workspaces; AJ-Bench tests evidence gathering, state checks, and process checks (Zhuge et al., 2024; Shi et al., 2026). Online Agent-as-a-Judge elicits targeted situations; SkillTV-Bench offers task-time skills and inspectable artifacts (Ryu et al., 2026; Han et al., 2026). Related systems support self-refinement, tool-assisted critique, safeguards, failure

Table 1: **Comparison with representative judge benchmarks.** Traj. marks execution traces; State, evidence beyond them; Proc., assessment of intermediate behavior; and Pre., intervention commitment before candidate outcomes. Long marks tasks requiring extended, history-dependent execution. Scale uses native units; rows are not directly comparable.

Benchmark	Evidence		Evaluation			Scope	
	Traj.	State	Proc.	Pre.	Long	Scale	Setting
DevAI (Agent-as-a-Judge) (Zhuge et al., 2024)	✓	✓	✓	✗	✓	55 tasks / 365 reqs.	Coding
JudgeBench (Tan et al., 2025)	✗	✗	✗	✗	✗	350 pairs / 4 domains	General
AgentRewardBench (Lù et al., 2025)	✓	✗	✓	✗	✗	1,302 traj. / 5 benchmarks	Web
MobileJudgeBench (Wan et al., 2026)	✓	✗	✗	✗	✗	931 traj. / 68 apps	Mobile
AJ-Bench (Shi et al., 2026)	✓	✓	✓	✗	✗	155 tasks / 516 traj.	Multi-domain
Online Agent-as-a-Judge (Ryu et al., 2026)	✓	✓	✓	✗	✓	32 criteria	Social
SkillTV-Bench (Han et al., 2026)	✓	✓	✓	✗	✓	50 tasks / 681 cases	Multi-domain
LONGJUDGE-BENCH (Ours)	✓	✓	✓	✓	✓	**70 cases / 2,450 branch outcomes**	**Multi-app**

localization, or repair (Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2024; Xiang et al., 2025; Fang et al., 2025; Ma et al., 2026; Chen et al., 2026c; Lin et al., 2026). These works span evidence acquisition, verification, and corrective action. LONGJUDGE-BENCH tests prospective intervention choice at one unfinished checkpoint, scored by empirical values from matched frozen branches.

Long-horizon agent evaluation. AgentBench, AppWorld, OSWorld, HORIZON, and OSWorld 2.0 span persistent state, long histories, and delayed consequences (Liu et al., 2024; Trivedi et al., 2024; Xie et al., 2024; Wang et al., 2026; Yuan et al., 2026). AgentBoard, ToolSandbox, and TRAJECT-Bench support intermediate progress and trajectory analysis (Ma et al., 2024; Lu et al., 2025; He et al., 2026). LivePlan, AgentProp-Bench, and DART study monitoring, propagation, and recovery (Liu et al., 2026b; Gurram, 2026; Yang et al., 2026); counterfactual methods examine recoverability, attribution, and repair (Jiang et al., 2026; Shah, 2026; Li et al., 2026). Closest work uses same-prefix branching to study intervention advantage, control targets, and learned routing policies (Zhang et al., 2026a); LONGJUDGE-BENCH evaluates LLM judges choosing among seven interventions before observing matched branch outcomes.

3 LONGJUDGE-BENCH: CONSTRUCTION AND EVALUATION

LONGJUDGE-BENCH defines an adapter-based protocol for choosing interventions at unfinished long-horizon checkpoints. Matched branches from a frozen checkpoint yield private values and an empirical reference. Given the visible history and menu, the judge selects one candidate; the evaluator then looks up its frozen value (Figure 2).

3.1 CASE CONSTRUCTION

We build cases from programmatically scorable, multi-application AppWorld tasks (Trivedi et al., 2024). Eligible tasks feature persistent state changes, cross-application dependencies, and outcomes resolved after several operations. A fixed executor yields trajectories used to select checkpoints. Normalization keeps state-changing calls, observations, constraints, and dependency-bearing events, but drops empty pagination, repeated scaffolding, and events with no state, constraint, or dependency content. Source IDs link each normalized trajectory to its run record.

Frozen intervention cases. We retain a checkpoint only when the task remains unfinished, no terminal outcome has occurred, and feasible interventions induce distinct continuations. Each checkpoint yields a public case with judge-visible evidence and a private snapshot of environment state, continuation queue, scenario variables, and branch metadata. The snapshot supports matched execution while keeping outcomes outside the judge input.

The benchmark contains 70 cases from 35 source IDs, with two checkpoints per source task. Its provenance registry defines a 42-case *source-disjoint* core and two 14-case audit strata: *source-sibling* and *pilot-exposed*. These labels support transparent lineage accounting and subset analysis without changing public inputs, frozen values, or pooled scoring. Source IDs also serve as resampling clusters, preserving dependence between checkpoints from the same task.

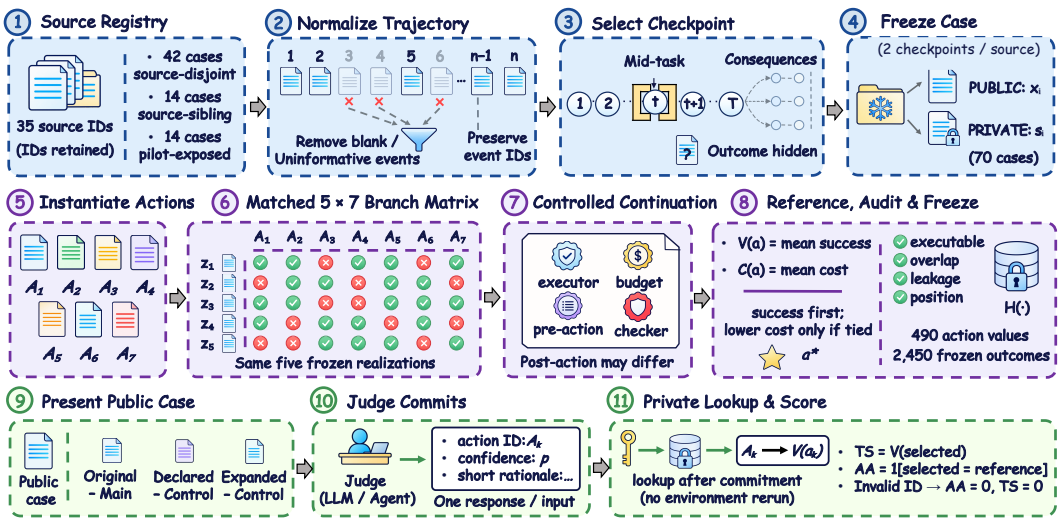


Figure 2: **LONGJUDGE-BENCH protocol.** Top: freeze two checkpoints per source task. Middle: run branches on matched states and form the private reference. Bottom: record the judge's choice and look up its private value for scoring.

3.2 MATCHED BRANCH REFERENCE CONSTRUCTION

Each case i instantiates a canonical set $\mathcal{A}_i^{\text{can}}$ of seven interventions: *continue*, *advise*, *block*, *revise*, *rollback*, *remedy*, and *escalate*. They preserve continuation, guide its controller, block or revise pending requests, reverse prior operations, remedy omissions, or escalate decisions (Appendix B). Candidates must be executable and operationally distinct; implementations may alter environment state, continuation queue, or controller behavior. Each specification fixes its target, preconditions, state or control effect, and subsequent flow; interfaces vary how explicitly they render these fields.

Matched branch execution. For each case, all seven candidates run on five matched frozen scenario–seed realizations. Each realization fixes scenario variables and a random seed; candidates start from the same pre-intervention state and queue and share the executor, permissions, budget, continuation contract, and checker. Branches are separate counterfactual continuations, not sequential actions or judge responses. Post-intervention queues may differ when queue modification is part of the declared intervention semantics.

Let $m = 5$ be the matched realizations per case–action pair, $Y_{iar} \in \{0, 1\}$ indicate completion for action a under realization r of case i, and c_{iar} be the decision-interaction count of controller-overhead events, successful target calls, and the final completion event. The canonical private values and reference are

$$\widehat{V}_i(a) = \frac{1}{m}\sum_{r=1}^{m} Y_{iar}, \qquad \widehat{C}_i(a) = \frac{1}{m}\sum_{r=1}^{m} c_{iar}, \qquad a_i^\star = \operatorname*{lex\,arg\,max}_{a \in \mathcal{A}_i^{\text{can}}} \big(\widehat{V}_i(a), -\widehat{C}_i(a)\big). \quad (1)$$

The rule maximizes empirical completion, using lower mean interactions only to break a completion tie; cost never offsets lower completion. We retain cases only when the lexicographic optimum is unique. We then recompute the rule while omitting each realization in turn; only cases whose leave-one-realization-out reference remains unchanged enter the frozen benchmark. Hence, $a_i^\star$ is invariant to every single-realization deletion in the matched sample. TS measures selected value, AA measures exact reference selection, and Regret measures value shortfall from the reference.

Before freezing, two domain experts independently check candidate targets, preconditions, effects, continuations, checker alignment, public–private consistency, and compliance of each computed reference with the rule. Each records a pre-adjudication judgment; disagreements and resolutions remain auditable. Automated tests verify executability, source overlap, exposure tags, answer-position balance, removal of private fields from public inputs, and file integrity; hashes verify frozen records. The canonical table has 490 values from 2,450 outcomes. After reference construction, the frozen

set is balanced across five domains and seven reference-action families, with two cases per domain–family pair and ten per family.

3.3 Public Evaluation Protocol

Under view v, the judge receives $x_i^{(v)} = (g_i, \widetilde{\tau}_{i,v}^{\le t_i}, o_{i,v}^{t_i}, p_i, \mathcal{A}_i^{(v)}, \Gamma_i)$, where g_i is the goal, t_i the checkpoint time, $\widetilde{\tau}_{i,v}^{\le t_i}$ the rendered trajectory, $o_{i,v}^{t_i}$ the observable state, p_i the pending request, $\mathcal{A}_i^{(v)}$ the displayed menu, and Γ_i the public intervention and continuation contract. The response contains a displayed choice ID, confidence, and rationale; the evaluator maps that choice to its saved canonical or view-specific identity before scoring.

The public representation excludes source and exposure identifiers, hidden scenario variables, queues, seeds, branch outcomes, values, and the reference action. These remain private with executable action implementations, executor and checker versions, and frozen run records. The judge therefore infers candidate consequences only from the evidence, action descriptions, and contract rendered in view v.

Evaluation interfaces. *Original* is the main end-to-end interface: full history and the canonical menu require semantic grounding plus consequence comparison. *Declared* keeps $\mathcal{A}_i^{\mathrm{can}}$ and $a_i^\star$ but states object bindings and action semantics, reducing the semantic-resolution burden. *Expanded* uses larger $\mathcal{A}_i^{\mathrm{exp}}$ in selected cases and recomputes $a_i^{\star,\mathrm{exp}}$. As rendering, candidates, and reference all change, Expanded tests response to a broader intervention space, not an isolated menu-size effect; added branches lie outside the canonical counts. Across all three views, each case retains one response schema, and each rendered input receives one judge response for final scoring.

After commitment, the evaluator validates the choice ID and retrieves its empirical value for a valid selection. Lookup fixes consequences across models, removing model-specific continuation variance. Unknown, missing, and ambiguous IDs are invalid responses. Section 4.1 defines Task Success, Action Accuracy, Invalid Rate, Regret, aggregation, and source-level uncertainty. Confidence and rationale remain auxiliary; primary scores depend only on the selected action identity.

4 Experiments

4.1 Experimental Setup

Datasets. LongJudge-Bench contains 70 intervention checkpoints derived from 35 AppWorld source IDs, with two checkpoints selected from each source.

Cases span five domains and seven reference-action families, balanced at 14 per domain and 10 per family. Table 2 reports seven actions per case, five outcomes per value, 490 values, and 2,450 evaluated outcomes. The D/S/P strata contain 42 source-disjoint, 14 source-sibling, and 14 pilot-exposed cases. Pooled and interface evaluations use all 70 cases; history/menu and position controls use the 42 source-disjoint cases. Three interfaces account for 210 slots, while the control families add 252 and 126. Reusing 42 Original records as Full history yields 546 condition instances from 588 case–view slots; Full repeat is an independent judgment of the same serialization. Expanded-only actions and their condition-specific reference are excluded from canonical counts. Appendix C specifies coverage.

Table 2: **Composition of LongJudge-Bench and its evaluation views.**

Statistic	Value
Source and Case Construction	
AppWorld source IDs	35
Checkpoints per source ID	2
Frozen intervention cases	70
Coverage and Provenance	
Application domains	5
Cases per domain	14
Reference-action families	7
Cases per reference family	10
Provenance cases (D/S/P)	42 / 14 / 14
Counterfactual Reference	
Canonical actions per case	7
Fixed outcomes per action	5
Case–action values	490
Frozen branch outcomes	2,450

Models. We evaluate 13 judge models. Seven are open-weight: DeepSeek-V3.1 and DeepSeek-R1-0528 (DeepSeek, 2025b;a); Qwen3-235B-A22B-Instruct-2507, Qwen3-235B-A22B-Thinking-2507, and Qwen3-Coder-480B-A35B-Instruct (Yang et al., 2025; Qwen Team, 2025b;a); Kimi-K2-Instruct-0905 (Moonshot AI, 2025); and GLM-4.5 (Zeng et al., 2025). Six are proprietary: Gemini

Table 3: **Main results under the full-history Original protocol.** Domain columns report Task Success (TS). Regret measures the shortfall from the empirical menu ceiling, AA denotes exact reference-action accuracy, and Inv. denotes invalid-response rate. TS, AA, and Inv. are percentages; Regret is in percentage points. **Bold**/underline mark the two best model estimates in each column, excluding baselines and the average. The final row averages the 13 individual models.

Judge / Policy	Application-Domain Task Success (%)					Consequence		Decision	
	Files	Media	Comm.	Pay.	X-App	TS ↑	Reg. ↓	AA ↑	Inv. ↓
Non-model Baselines									
Uniform random	28.78	29.39	28.98	34.69	26.53	29.67	69.47	14.29	0.00
Fixed-A	25.71	32.86	51.43	35.71	20.00	33.14	66.00	14.29	0.00
Open-weight Models									
DeepSeek-V3.1	48.57	41.43	44.29	38.57	40.00	42.57	56.57	24.29	12.86
DeepSeek-R1-0528	58.57	50.00	54.29	45.71	51.43	52.00	47.14	38.57	8.57
Qwen3-235B-A22B-Instruct-2507	44.29	35.71	40.00	31.43	38.57	38.00	61.14	15.71	22.86
Qwen3-235B-A22B-Thinking-2507	54.29	47.14	50.00	41.43	48.57	48.29	50.86	31.43	11.43
Qwen3-Coder-480B-A35B-Instruct	51.43	40.00	44.29	34.29	40.00	42.00	57.14	20.00	20.00
Kimi-K2-Instruct-0905	58.57	51.43	52.86	42.86	50.00	51.14	48.00	34.29	11.43
GLM-4.5	64.29	57.14	60.00	50.00	55.71	57.43	41.71	42.86	5.71
Proprietary Models									
Gemini 3.1 Pro	62.86	58.57	61.43	51.43	57.14	58.29	40.86	41.43	7.14
Gemini 3 Flash	55.71	50.00	52.86	44.29	48.57	50.29	48.86	30.00	10.00
GPT-5.4 mini	58.57	52.86	55.71	45.71	50.00	52.57	46.57	34.29	8.57
GPT-5.5	**75.71**	68.57	72.86	61.43	67.14	69.14	30.00	54.29	2.86
Claude Sonnet 5	68.57	62.86	67.14	54.29	61.43	62.86	36.29	47.14	4.29
Claude Opus 4.6	74.29	**70.00**	**75.71**	**62.86**	**70.00**	**70.57**	**28.57**	**58.57**	**1.43**
Avg. (13 models)	59.67	52.75	56.26	46.48	52.20	53.47	45.67	36.37	9.78

3.1 Pro and Gemini 3 Flash (Google Gemini Team, 2026; Google Cloud, 2026); GPT-5.4 mini and GPT-5.5 (OpenAI, n.d.a;n); and Claude Sonnet 5 and Claude Opus 4.6 (Anthropic, 2026). Interface panels use three models per access regime, position panels use three models, and history controls use GPT-5.5; each analysis fixes its model set across views. Models receive identical case content, schema, parser, and values within a condition. Model-run records retain endpoint identifiers, access dates, decoding settings, and responses; Appendix C gives coverage.

Metrics. For view v, let N_v be its case count, $\widehat{a}_i^{(v)} \in \mathcal{A}_i^{(v)} \cup \{\perp\}$ its parsed choice, $\nu_i^{(v)}$ the invalid indicator, and $a_i^{\star,(v)}$ its saved reference. Set $\widetilde{V}_i^{(v)} = \widehat{V}_i^{(v)}(\widehat{a}_i^{(v)})$ for valid selections and zero otherwise. Task Success (TS) is $\mathrm{TS}^{(v)} = 100 N_v^{-1} \sum_i \widetilde{V}_i^{(v)}$; Action Accuracy (AA) is $\mathrm{AA}^{(v)} = 100 N_v^{-1} \sum_i \mathbf{1}[\nu_i^{(v)} = 0 \wedge \widehat{a}_i^{(v)} = a_i^{\star,(v)}]$. Invalid Rate is $\mathrm{Inv}^{(v)} = 100 N_v^{-1} \sum_i \nu_i^{(v)}$, and Regret is $\mathrm{Regret}^{(v)} = 100 N_v^{-1} \sum_i [\widehat{V}_i^{(v)}(a_i^{\star,(v)}) - \widetilde{V}_i^{(v)}]$. Original and Declared use canonical values and references; Expanded uses its condition-specific table. Family-Balanced AA macro-averages canonical families; it equals AA for Original/Declared. Point estimates are deterministic given saved responses and values. Paired model and history contrasts use 95% intervals from 10,000 source-cluster bootstrap draws. Interface and position profiles are descriptive. Resampling retains both checkpoints per sampled task, estimating variation across sources while fixing the value table.

4.2 Q1: How Well Do Models Judge Interventions?

Table 3 reports cross-model performance under Original. Claude Opus 4.6 has the highest TS and AA (70.57% and 58.57%); GPT-5.5 follows at 69.14% and 54.29%, while GLM-4.5 leads open-weight models at 57.43% TS. The paired source-cluster interval for Opus minus GPT-5.5 includes zero, so their order does not establish statistical separation. Across five within-family comparisons—DeepSeek R1/V3.1, Qwen Thinking/Instruct, Gemini Pro/Flash, GPT-5.5/5.4 mini, and Opus/Sonnet—the higher-scoring variant gains 7.71–16.57 TS points but 11.43–20.00 AA points. This direction indicates that within-family performance differences are expressed more strongly in exact reference recovery than in retained value, although these pairs are not controlled architecture ablations. Across models, mean TS is 53.47% and mean AA is 36.37%. Their positive gap requires value contributed by non-reference decisions, but means cannot separate equal-value

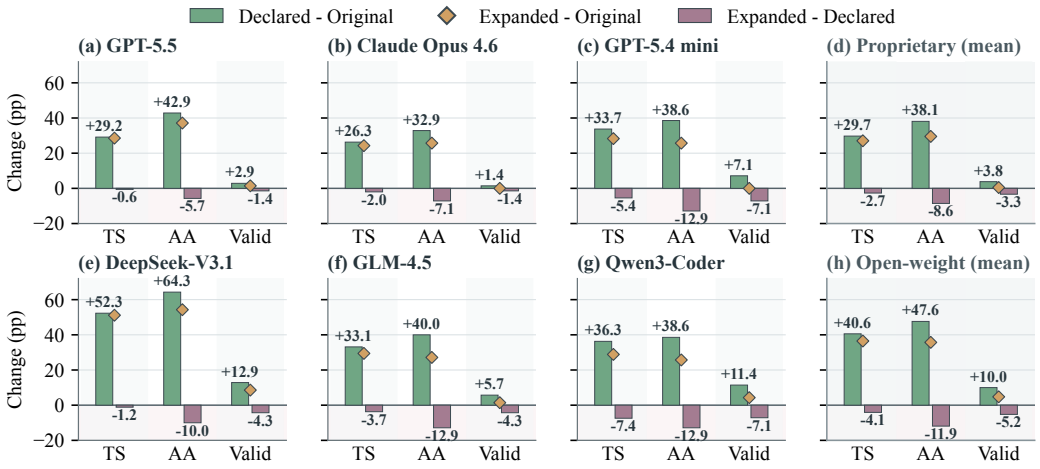


Figure 3: **Observed changes under interface declaration and menu expansion.** Bars show paired score changes in percentage points between adjacent interfaces; diamonds show the net Expanded − Original change.

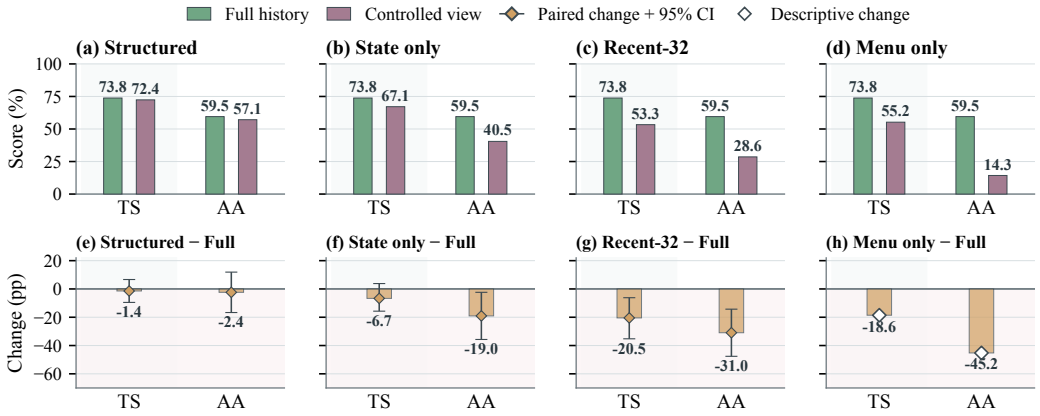


Figure 4: **Historical-evidence controls for GPT-5.5 on 42 matched cases.** Top: absolute TS/AA. Bottom: changes from Full history; intervals are shown only for contrasts sharing that baseline, and hollow Menu-only points are descriptive.

alternatives from partially successful errors. Opus produces 98.57% valid outputs yet 58.57% AA, leaving 40.00% valid non-reference decisions; because invalid responses contribute at most 1.43 regret points, most of its 28.57-point regret reflects valid lower-value selections. Payments is lowest for every model. Because every domain is balanced across the seven reference families, this deficit cannot follow from unequal family composition; it instead localizes difficulty to domain-specific evidence or execution demands without identifying which. The 99.14% empirical menu ceiling defines intervention-selection headroom.

4.3 Q2: Does Interface Declaration Change the Task?

Figure 3 reports descriptive matched interface changes. Declared raises TS, AA, and validity for all six models. Mean TS and AA gains are 40.6 and 47.6 points in the open-weight subset and 29.7 and 38.1 in the proprietary subset. Yet open-weight models start lower under Original, and DeepSeek-V3.1—the weakest Original model—has the largest gains. Thus the group difference is compatible with headroom, not evidence that access regime causes interface sensitivity. Because validity im-

proves, TS and AA gains combine fewer schema failures with better valid choices. Original requires semantic grounding and consequence comparison; Declared fixes actions and reference while exposing operational bindings, reducing rather than eliminating semantic-resolution demands. Expanded partly offsets these gains: AA falls by 11.9 and 8.6 points across the two subsets, while TS falls only 4.1 and 2.7. This asymmetry is consistent with added alternatives perturbing action identity more than downstream value. Expanded also changes rendering, feasible actions, and its reference, however, so the contrast cannot isolate menu size or prove that mechanism. Original–Declared characterizes declaration sensitivity under a fixed target; Expanded is a composite end-to-end stress test.

4.4 Q3: How Sensitive Is GPT-5.5 to Historical Evidence?

Historical context is more consequential for exact recovery than value retention. Figure 4 compares five views of 42 source-disjoint checkpoints. Full history yields 73.81% TS and 59.52% AA. Recent-32 is lower by 20.48 and 30.95 points; State-only is lower by 6.67 TS and 19.05 AA points. Paired 95% intervals exclude zero for both Recent-32 metrics and State-only AA, but not State-only TS. This asymmetry is consistent with observable state sustaining consequence-preserving actions while exact intervention identity depends more on interaction history. Because the views alter content and length, the contrast cannot identify which evidence drives the difference.

Structured history yields higher scores than recent truncation. Structured yields 72.38% TS and 57.14% AA; Full-history paired intervals span zero and do not establish equivalence. Its smaller loss than Recent-32 is consistent with preserving task-relevant structure rather than only recent events, but that direct contrast lacks an interval and changes input length. Menu-only yields 55.24% TS but 14.29% AA, exactly the $1/7$ uniform-choice level. Thus Menu-only AA does not exceed nominal chance. Positive TS shows that alternatives possess value, but without a matched 42-case random-TS baseline it cannot establish above-random consequence selection. The separate Menu-only–Full-repeat intervals assess only that paired contrast, not response stochasticity or the plotted Full-history difference (Appendix C.2). These controls distinguish value retention from exact reference recovery while preserving their inferential boundaries.

4.5 Q4: Does Performance Vary Monotonically with Evidence Position?

Table 4: **Evidence-position profiles on 42 matched cases per view.** Scores are percentages; changes and spans are percentage points. Parentheses give successful outcomes for TS ($/210$) and exact actions for AA ($/42$).

Model	Metric	Score (count)			Change vs. Near		Span
		Near	Middle	Far	Middle	Far	max − min
GPT-5.5	TS	73.81 (155)	**76.19** (160)	72.86 (153)	+2.38	−0.95	3.33
	AA	59.52 (25)	**66.67** (28)	57.14 (24)	+7.14	−2.38	9.52
DeepSeek-R1-0528	TS	56.19 (118)	**58.10** (122)	54.76 (115)	+1.90	−1.43	3.33
	AA	40.48 (17)	**45.24** (19)	35.71 (15)	+4.76	−4.76	9.52
Claude Sonnet 5	TS	67.14 (141)	**69.52** (146)	65.24 (137)	+2.38	−1.90	4.29
	AA	50.00 (21)	**57.14** (24)	47.62 (20)	+7.14	−2.38	9.52

Table 5: **Three-model position profiles.** Scores are percentages; Middle–Far is in percentage points.

Metric	Near	Middle	Far	Middle−Far	Mid. peak
TS	65.71	**67.94**	64.29	+3.65	3/3
AA	50.00	**56.35**	46.83	+9.52	3/3

Exact reference recovery shows greater positional variation. All three models peak at Middle and score lower at Far than Near. Middle exceeds Far by 3.65 TS and 9.52 AA points. Every model has a 9.52-point AA span, equivalent to four of 42 reference decisions, whereas TS spans only 3.33–4.29 points, or seven to nine of 210 outcomes. This cross-model agreement localizes greater variation to action identity, but without paired intervals remains descriptive. Identical spans need not involve the same cases: aggregate maxima and minima can conceal offsetting action changes, so shared failure mechanisms require transition-level evidence.

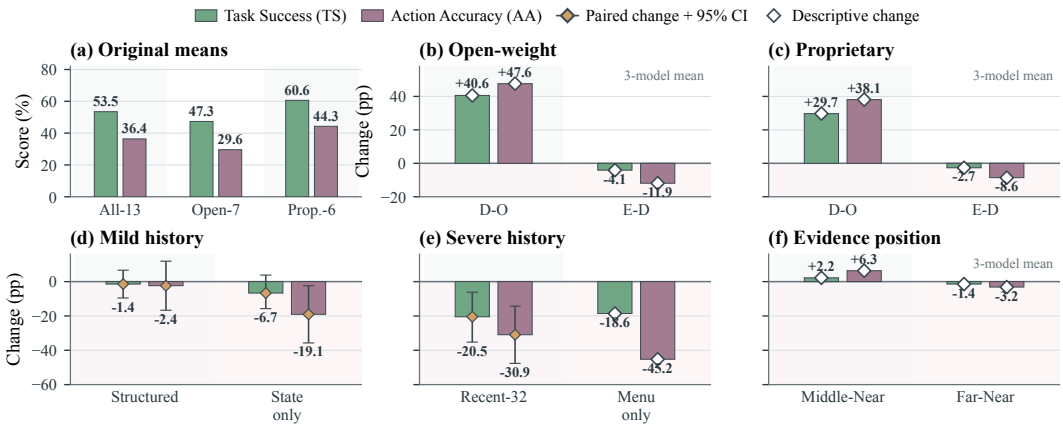


Figure 5: **Action accuracy versus retained value.** Green and purple bars show TS and AA. (a) Original means over all 13, seven open-weight, and six proprietary models. (b,c) Three-model interface changes; Expanded uses a condition-specific reference. (d,e) GPT-5.5 history changes; (f) three-model position changes. Filled diamonds with whiskers show paired 95% intervals; hollow diamonds mark descriptive contrasts.

Position profiles reject a monotonic proximity account. The views preserve target evidence while relocating it within serializations. Their shared Middle peak contradicts monotonic proximity: nearest is not best. Relocation also changes evidence order and neighboring context, precluding attribution to a distance or attention mechanism. The common Middle peak across distinct baseline levels weakens a model-specific account, but these three models alone cannot establish generality across architectures. The evidence therefore establishes a non-monotonic rendering profile, not a causal position effect.

4.6 Q5: What Does Action Accuracy Reveal Beyond Retained Value?

TS and AA capture two complementary decision properties. Under Original, AA trails TS for every model, with a 17.10-point mean gap. TS credits selected frozen value; AA credits only the empirical reference. AA alone treats every alternative as equally incorrect despite retained value, whereas TS alone hides failure to recover the benchmark's preferred intervention. Their gap is descriptive, but reporting separates consequence robustness from intervention fidelity and avoids reading value-preserving deviations as failures. Regret grades value lost when choices differ, while Invalid Rate separates decision quality from response-contract compliance. Collapsing these measures would erase distinct failure modes.

Metric separation recurs across all evaluated comparisons. AA changes more than TS in four interface and six history/position shifts, and improves more in all five within-family comparisons. This ordering indicates that exact action identity is more sensitive to model choice and evidence representation than retained value. Because several contrasts are descriptive and Expanded changes its reference, the pattern establishes no single mechanism. Operationally, TS measures consequence preservation, AA measures supervisory fidelity, and their divergence diagnoses substitutable non-reference decisions.

5 Conclusion

LongJudge-Bench shows that intervention judging has distinct notions of correctness: choices may differ from the empirical reference yet preserve downstream value. Across models and views, AA varies more than TS, making intervention identity more sensitive to model choice and evidence representation than consequence preservation. Declared semantics improve scores; Expanded remains a composite change. Joint TS and AA reporting separates consequence preservation from reference fidelity. Evidence covers 70 AppWorld checkpoints and five outcomes per action; broader environments and denser sampling remain tests of generality.

AI USE STATEMENT

Generative AI tools were used to retrieve and summarize related literature; support research ideation and execution through methodological critique, code development and inspection, figure refinement, result interpretation, and consistency checks; and draft, translate for author review, and polish portions of the manuscript. They were not used to generate benchmark data or formulate or prove mathematical claims. The authors manually verified all AI-assisted citations, code, figures, analyses, and claims.

ETHICS STATEMENT

This work evaluates simulated agent interactions and involves no recruited participants or real-world personal data. Its intervention labels and empirical references reflect the benchmark's task and executor assumptions; the benchmark is intended to support the study of safer agent supervision.

REPRODUCIBILITY STATEMENT

Reproducibility is supported by fixed case and action identifiers, frozen branch records, content hashes, and scoring rules. The appendix documents counting units, coverage, interfaces, and bootstrap procedures; model-run records retain endpoint identifiers, access dates, decoding settings, and saved responses for audit.

REFERENCES

Anthropic. Claude model reports, 2026. URL https://www.anthropic.com/transparency. Claude Opus 4.6 and Claude Sonnet 5.

Gongwei Chen, Lirong Jie, Lexiao Zou, Weili Guan, Miao Zhang, and Liqiang Nie. Enhancing gui agent with uncertainty-aware self-trained evaluator. *Advances in Neural Information Processing Systems*, 38:142026–142058, 2026a.

Junjie Chen, Yuxi Dong, Haitao Li, Weihang Su, Yujia Zhou, Min Zhang, Yiqun Liu, and Qingyao Ai. Benchmarking llm-as-a-judge for long-form output evaluation. *arXiv preprint arXiv:2606.01629*, 2026b.

Yinfang Chen, Jiaqi Pan, Jackson Clark, Yiming Su, Noah Zheutlin, Bhavya Bhavya, Rohan R Arora, Yu Deng, Saurabh Jha, and Tianyin Xu. Stratus: A multi-agent system for autonomous reliability engineering of modern clouds. *Advances in Neural Information Processing Systems*, 38:50119–50165, 2026c.

DeepSeek. Deepseek-r1-0528 release. Model release announcement, 2025a. URL https://api-docs.deepseek.com/news/news250528/.

DeepSeek. Deepseek-v3.1 release. Model release announcement, 2025b. URL https://api-docs.deepseek.com/news/news250821/.

Haishuo Fang, Xiaodan Zhu, and Iryna Gurevych. Preemptive detection and correction of misaligned actions in llm agents. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pp. 222–244, 2025.

Google Cloud. Gemini 3 flash. Model documentation, 2026. URL https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-flash.

Google Gemini Team. Gemini 3.1 pro: A smarter model for your most complex tasks. Google Blog, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Bhaskar Gurram. Evaluating tool-using language agents: Judge reliability, propagation cascades, and runtime mitigation in agentprop-bench. *arXiv preprint arXiv:2604.16706*, 2026.

Zhi Han, Chenxi Zeng, Liuhaichen Yang, Zihan Guo, Ming Zhou, and Yang Li. Skilltv-bench: Benchmarking how well judges perform on skill-augmented agentic execution. *arXiv preprint arXiv:2608.05573*, 2026.

Pengfei He, Zhenwei Dai, Bing He, Hui Liu, Xianfeng Tang, Hanqing Lu, Juanhui Li, Jiayuan Ding, Subhabrata Mukherjee, Suhang Wang, et al. Traject-bench: a trajectory-aware benchmark for evaluating agentic tool use. In *International Conference on Learning Representations*, volume 2026, pp. 61766–61801, 2026.

De Jiang, Zhengyang Zhang, Kehong Yuan, and Shaohua Ma. Not every divergence should be suppressed: Counterfactual recoverability in on-policy distillation. *arXiv preprint arXiv:2608.04408*, 2026.

Bingjie Li, Yumeng Song, Zhongming Yao, and Tianyi Li. Localizing emergent failures in agentic ai: Recovering minimal repair families via counterfactual replay. *arXiv preprint arXiv:2608.29228*, 2026.

Xiaofeng Lin, Yingxu Wang, Tung Sum Thomas Kwok, Daniel Guo, Sahil Arun Nale, Charles Fleming, and Guang Cheng. Reflect: Intervention-supported error attribution for silent failures in llm agent traces. *arXiv preprint arXiv:2606.09071*, 2026.

Jiale Liu, Huajun Xi, Shaokun Zhang, Yifan Zeng, Tianwei Yue, Chi Wang, Jian Kang, Qingyun Wu, and Huazheng Wang. Who&when pro: Can llms really attribute failures in ai agents? *arXiv preprint arXiv:2607.09996*, 2026a.

Shuyang Liu, Saman Dehghan, Ji Young Kim, Jatin Ganhotra, Martin Hirzel, and Reyhaneh Jabbarvand. Online monitoring and corrective steering of programming agents. *arXiv preprint arXiv:2608.06701*, 2026b.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025.

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.

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. *Advances in neural information processing systems*, 37:74325–74362, 2024.

Ming Ma, Jue Zhang, Fangkai Yang, Yu Kang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Dover: Intervention-driven auto debugging for llm multi-agent systems. In *International Conference on Learning Representations*, volume 2026, pp. 28059–28087, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Moonshot AI. Kimi-k2-instruct-0905. Hugging Face model card, 2025. URL `https://huggingface.co/moonshotai/Kimi-K2-Instruct-0905`.

OpenAI. Gpt-5.4 mini. OpenAI API model documentation, n.d.a. URL https://developers.openai.com/api/docs/models/gpt-5.4-mini.

OpenAI. Gpt-5.5. OpenAI API model documentation, n.d.b. URL https://developers.openai.com/api/docs/models/gpt-5.5.

Qwen Team. Qwen3-coder: Agentic coding in the world. Blog post, 2025a. URL https://qwenlm.github.io/blog/qwen3-coder/.

Qwen Team. Qwen3-235b-a22b-thinking-2507. Hugging Face model card, 2025b. URL https://huggingface.co/Qwen/Qwen3-235B-A22B-Thinking-2507.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. In *International Conference on Learning Representations*, volume 2024, pp. 27031–27098, 2024.

Hyogon Ryu, Jeonghwan Kim, Yewon Lim, Chaeun Lee, Jeongwook Kim, and Donghoon Ham. Online agent-as-a-judge: Situation-generating evaluation for interactive agents. *arXiv preprint arXiv:2606.08200*, 2026.

Jaineet Shah. Causal agent replay: Counterfactual attribution for llm-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

Wentao Shi, Yu Wang, Yuyang Zhao, Yuxin Chen, Fuli Feng, Xueyuan Hao, Xi Su, Qi Gu, Hui Su, Xunliang Cai, et al. Aj-bench: Benchmarking agent-as-a-judge for environment-aware evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25371–25413, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Qiushi Sun, Kanzhi Cheng, Yian Wang, Bowen Yang, Hang Yan, Liheng Chen, Fangzhi Xu, Zichen Ding, Nuo Chen, Jialin Cao, et al. Osreward: Instituting standardized evaluation for cross-platform computer-use reward models. *arXiv preprint arXiv:2607.28609*, 2026.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pp. 63277–63303, 2025.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, 2024.

Rakshith Vasudev, Melisa Russak, Dan Bikel, and Waseem Alshikh. Accurate failure prediction in agents does not imply effective failure prevention. *arXiv preprint arXiv:2602.03338*, 2026.

Ziqiang Wan, Li Gu, Zhixiang Chi, Zhi Liu, Seyed Mehdi Ayyoubzadeh, Yuanhao Yu, and Yang Wang. Benchmarking llm judges for mobile agent evaluation. *arXiv preprint arXiv:2608.11434*, 2026.

Xinyu Jessica Wang, Haoyue Bai, Yiyou Sun, Haorui Wang, Shuibai Zhang, Wenjie Hu, Mya Schroder, Bilge Mutlu, Dawn Song, and Robert D Nowak. The long-horizon task mirage? diagnosing where and why agentic systems break. *arXiv preprint arXiv:2604.11978*, 2026.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Nathaniel D Bastian, et al. Guardagent: safeguard llm agents via knowledge-enabled reasoning. In *ICML 2025 workshop on computer use agents*, 2025.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.

Ke Yang, Panpan Li, Zonghan Wu, Kejin Xu, Huaxi Huang, and Xiaoshui Huang. Dart: Semantic recoverability for structured tool agents. *arXiv preprint arXiv:2605.23311*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, et al. Osworld2. 0: Benchmarking computer use agents on long-horizon real-world tasks. *arXiv preprint arXiv:2606.29537*, 2026.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Chubin Zhang, Zhenglin Wan, Xingrui Yu, Jingxuan Wu, Qi Wen, Pengfei Zhou, Wangbo Zhao, and Ivor Tsang. Calibration is not control: Why llm-agent oversight needs intervention. *arXiv preprint arXiv:2606.21399*, 2026a.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. Agentracer: Who is inducing failure in the llm agentic systems? In *International Conference on Learning Representations*, volume 2026, pp. 11377–11399, 2026b.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, et al. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212*, 2025.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, et al. Where llm agents fail and how they can learn from failures. *arXiv preprint arXiv:2509.25370*, 2025.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024.

A BENCHMARK DETAILS

This appendix defines source tasks, intervention cases, controlled views, frozen outcomes, model judgments, counting, scoring, and reported uncertainty. It adds neither result figures nor broader empirical claims.

A.1 COMPOSITION AND COUNTING UNITS

Table 6: **Benchmark composition and counting units.** Slot and instance counts precede expansion over models; judgment count follows model expansion. Canonical action counts exclude Expanded-only alternatives.

Unit	Count	Derivation	Role
AppWorld source IDs	35	Frozen source registry	Overlap and resampling unit
Intervention cases	70	Two checkpoints per source	Primary scoring unit
Provenance strata	42 / 14 / 14	Source-disjoint / source-sibling / pilot-exposed	Provenance slices
Application domains	5	14 cases per domain	Domain TS slices
Reference-action families	7	10 cases per family	Balanced action slices
Canonical case–action values	490	70×7	Private value table
Frozen branch outcomes	2,450	490×5	Value estimates
Protocol slots	210	3×70	Original / Declared / Expanded
History and menu slots	252	6×42	Matched controls
Evidence-position slots	126	3×42	Near / Middle / Far
All condition slots	588	$210 + 252 + 126$	Case–view pairs
Condition instances	546	$588 - 42$ reused Original/Full-history slots	Keeps Full repeat distinct
Model judgments	2,338	Coverage in Section C	Saved model responses

The registry partitions 35 source IDs into 21 source-disjoint (D), seven source-sibling (S), and seven pilot-exposed (P) sources, each contributing two cases. D is held out by source; S uses new checkpoints from source IDs in the curation partition; P comprises pilot cases revalidated under the final pipeline. Labels affect neither public inputs nor pooled scores; controls use D.

Case records retain event counts, checkpoint depth, and cross-application dependencies. Slots are case–view pairs; condition instances merge Original/Full reuse but retain Full repeat. Neither expands over models. Source IDs link paired checkpoints for overlap and resampling; cases remain scoring units. Every domain–family cell contains two cases.

A.2 PUBLIC INPUTS AND PRIVATE VALUES

The judge receives the goal, normalized visible history, observable checkpoint state, pending request, candidate IDs with condition-specific descriptions, and the continuation contract. Source and provenance identifiers, frozen state, branch outcomes, empirical values, and the reference action are not shown. Candidate order is presentation metadata: displayed IDs are mapped back to their saved action identities before scoring.

This separation keeps post-intervention outcomes outside the judge input while allowing every displayed choice to be evaluated by lookup. Original, Declared, and the history or position controls retain the corresponding canonical action identities. Expanded may add separately executed alternatives, as specified below.

B Actions and Empirical Reference

B.1 Canonical Intervention Families

Table 7: **Operational distinction among the seven canonical intervention families.**

Family	State or control effect
Continue	Apply no supervisory modification and follow the specified continuation
Advise	Send specified guidance to the continuation controller
Block	Suppress a named request that has not executed
Revise	Replace the target or arguments of an unexecuted request
Rollback	Reverse a completed operation or restore an admissible prior state
Remedy	Execute a specified compensating or omitted operation
Escalate	Obtain an authorized fact or decision and follow the specified conditional branch

Two domain experts independently check each candidate's target, preconditions, state/queue change, continuation, checker alignment, and reference-rule compliance. Pre-adjudication judgments and resolutions are retained. These construction-validity checks neither define references nor enter model scores.

The families distinguish where an intervention acts. Block and Revise modify a pending request before execution; Rollback addresses an operation that has already occurred; Remedy adds a compensating or omitted operation. Advise changes controller guidance, whereas Escalate obtains an authorized fact or decision before following a specified branch. Continue preserves the declared continuation without a supervisory modification.

B.2 Matched Values and Unique References

For each case, all seven canonical candidates are evaluated on the same five frozen scenario–seed realizations. Within realization r, candidates share the pre-intervention state, queue, executor, budget, permissions, continuation contract, and checker. With binary completion Y_{iar} and recorded interaction count c_{iar},

$$\widehat{V}_i(a) = \frac{1}{5}\sum_{r=1}^{5} Y_{iar}, \qquad \widehat{C}_i(a) = \frac{1}{5}\sum_{r=1}^{5} c_{iar}, \tag{2}$$

and

$$a_i^\star = \operatorname*{lex\,arg\,max}_{a \in \mathcal{A}_i^{\mathrm{can}}} \big(\widehat{V}_i(a), -\widehat{C}_i(a)\big). \tag{3}$$

Mean success is optimized first; lower mean interaction count breaks ties. Only references invariant under all five leave-one-realization-out recomputations enter the frozen benchmark. Here c_{iar} counts controller overhead, successful target calls, and final completion rather than token cost, wall-clock time, or attempted calls.

The seven branches are alternative continuations from the same checkpoint. Differences in their post-intervention states or queues are therefore consequences of the advertised actions. Five matched outcomes per action define the benchmark's empirical value table, making TS, AA, and Regret deterministic for every saved model selection under that table. During evaluation, each non-reused model–case–view instance contributes one judgment against the same frozen target.

Original and Declared share the seven canonical values and reference. Expanded may add executable alternatives and therefore uses a condition-specific value table and reference. Expanded-only branches are excluded from the 490 canonical values and 2,450 canonical outcomes reported above.

B.3 Consequence Value and Exact Selection

TS and AA intentionally answer different questions. TS credits the frozen value of the selected action, whereas AA requires the unique reference action. If two actions have equal success value

and the cost rule selects one as the reference, choosing the other preserves the same TS contribution but receives zero AA for that case. A lower-value non-reference choice decreases both retained value and action accuracy. Invalid outputs contribute zero to both metrics.

This distinction also limits comparisons involving Expanded. Its AA is computed against the condition-specific expanded reference, not against the Original action identity. Expanded contrasts therefore combine a changed rendering with a changed feasible action set wherever additional actions are present; they do not separately identify a menu-size effect.

C Evaluation Coverage

C.1 Models and Analyses

Main comparison. All 13 models in Section 4.1 are evaluated once on all 70 cases under the strict Original interface. This common evaluation supports Table 3 and fixes the same case content, response schema, parser, and frozen values across models.

Interface analysis. The matched Original–Declared–Expanded comparison covers six models on all 70 parent cases. Its open-weight subset contains DeepSeek-V3.1, Qwen3-Coder-480B-A35B-Instruct, and GLM-4.5; its proprietary subset contains GPT-5.4 mini, GPT-5.5, and Claude Opus 4.6. Original judgments are reused from the main comparison, while Declared and Expanded each add one judgment per model and parent case.

History and position analyses. The history/menu study evaluates GPT-5.5 on six matched views of the 42 source-disjoint cases: Full history, Structured, State only, Recent-32, Full repeat, and Menu only. Full history reuses the corresponding Original judgment; the other five views add independent judgments. The position study evaluates DeepSeek-R1-0528, GPT-5.5, and Claude Sonnet 5 on Near, Middle, and Far views of the same 42 cases. Each history/position analysis fixes its model set and parent-case coverage across views. Accordingly, no controlled contrast mixes models, cases, or condition-specific scoring targets across views.

Judgment accounting. The coverage above yields 2,338 saved model judgments: 910 Main/Original judgments, 840 additional Declared or Expanded judgments, 210 additional history/menu judgments after Original reuse, and 378 position judgments. This model-expanded count is distinct from the 588 case–view slots and 546 condition instances in Table 6. Model-run records retain endpoint identifiers, access dates, decoding settings, serialized inputs, and responses. Within each condition, the schema, parser, and frozen values remain fixed; provider-controlled fields follow saved endpoint metadata.

C.2 Controlled Views

Table 8: **Evaluation-view registry.** All controlled views inherit their source and parent-case IDs.

View	N	Input change	Reference / role
Original	70	Full trajectory; seven canonical actions	Canonical; main protocol
Declared	70	Explicit semantics; same actions	Canonical; paired interface
Expanded	70	Explicit semantics; additional actions in selected cases	View-specific; expanded interface
Full history	42	Complete normalized prefix; canonical menu	Canonical; reused Original
Structured	42	Frozen structured prefix; canonical menu	Canonical; matched history
State only	42	Observable state without interaction prefix	Canonical; matched history
Recent-32	42	Final 32 normalized events in order	Canonical; matched history
Full repeat	42	Independent rerun of the same Full-history serialization	Canonical; repeat baseline
Menu only	42	Goal and menu without history/state	Canonical; matched history
Near	42	Target evidence latest among renderings	Canonical; matched position
Middle	42	Target evidence at intermediate location	Canonical; matched position
Far	42	Target evidence earliest among renderings	Canonical; matched position

Structured is a frozen benchmark-authored rendering of task-relevant prefix facts, not a general compression method. State only keeps observable state; Recent-32 keeps final ordered events; Menu only retains the goal and menu. Saved serializations define these information-and-length conditions.

Original and Declared use the same seven actions and frozen canonical values; their contrast concerns how those actions are described. Expanded additionally tests a larger menu in selected cases and is scored with its corresponding reference. Within the history family, Full history reuses the source-disjoint Original record, while Full repeat obtains an independent judgment from the same serialization. The remaining history views remove or restructure visible evidence while retaining the parent case and canonical menu.

Position views relocate target evidence across three frozen prompts. Near, Middle, and Far are ordinal within each triplet, not fixed token bins; saved serializations define each view. Profiles also change evidence order and neighboring context.

C.3 Baselines and Uncertainty

Under Original, Uniform random is the exact mean from choosing uniformly among seven canonical actions. Fixed-A selects public `choice_a` and maps it to its saved action identity. The empirical menu ceiling averages the per-case maximum canonical values and defines Regret and intervention-selection headroom.

Because the canonical reference is unique, uniform random has expected AA $1/7$; its expected TS averages the seven frozen action values within each case. Fixed-A instead tests a constant displayed-position choice after the public-to-canonical mapping.

The reported paired model contrast and history effects use 10,000 source-cluster bootstrap draws. Each draw samples source IDs with replacement and retains both checkpoints and all paired views for every sampled source. The pooled model contrast uses 35 clusters; source-disjoint history contrasts use 21. Reported endpoints are empirical 2.5th and 97.5th percentiles. Interface and position profiles remain descriptive. No inferential interval is attached to either descriptive profile family. The plotted Menu-only change uses Full history; its saved interval instead uses independent Full repeat: $[-30.95, -7.14]$ for TS and $[-61.90, -26.19]$ for AA. Those intervals therefore do not apply to the plotted Full-history contrast in Figure 4.

Resampling source IDs preserves dependence between checkpoints from the same task. Intervals quantify variation across sampled sources conditional on the saved responses and finite five-outcome value table; they do not propagate branch-estimation or repeated-call uncertainty.

D Scope

LongJudge-Bench evaluates one intervention choice from a finite executable menu at a supplied checkpoint. Its scope is pre-outcome action selection under a fixed continuation, complementing systems that learn invocation timing, free-form repairs, or repeated intervention policies. Empirical values and references are defined by the frozen executor, checker, candidate semantics, and continuation contract. Table 6 defines composition; Section C defines model coverage. D/S/P encodes source-level lineage. Interface views use all cases; history and position controls use the source-disjoint subset. State, action, and checker adapters specify integration points, but the present evidence remains limited to AppWorld.

Within this domain, TS measures checker-defined task completion under the specified continuation, while AA measures exact selection of the corresponding empirical reference. Their combination distinguishes retained consequence value from exact reference selection over the finite frozen distribution. Regret quantifies remaining headroom to its menu ceiling. Together, these metrics provide an auditable target for comparing intervention judges under shared actions, continuations, and executable outcomes.