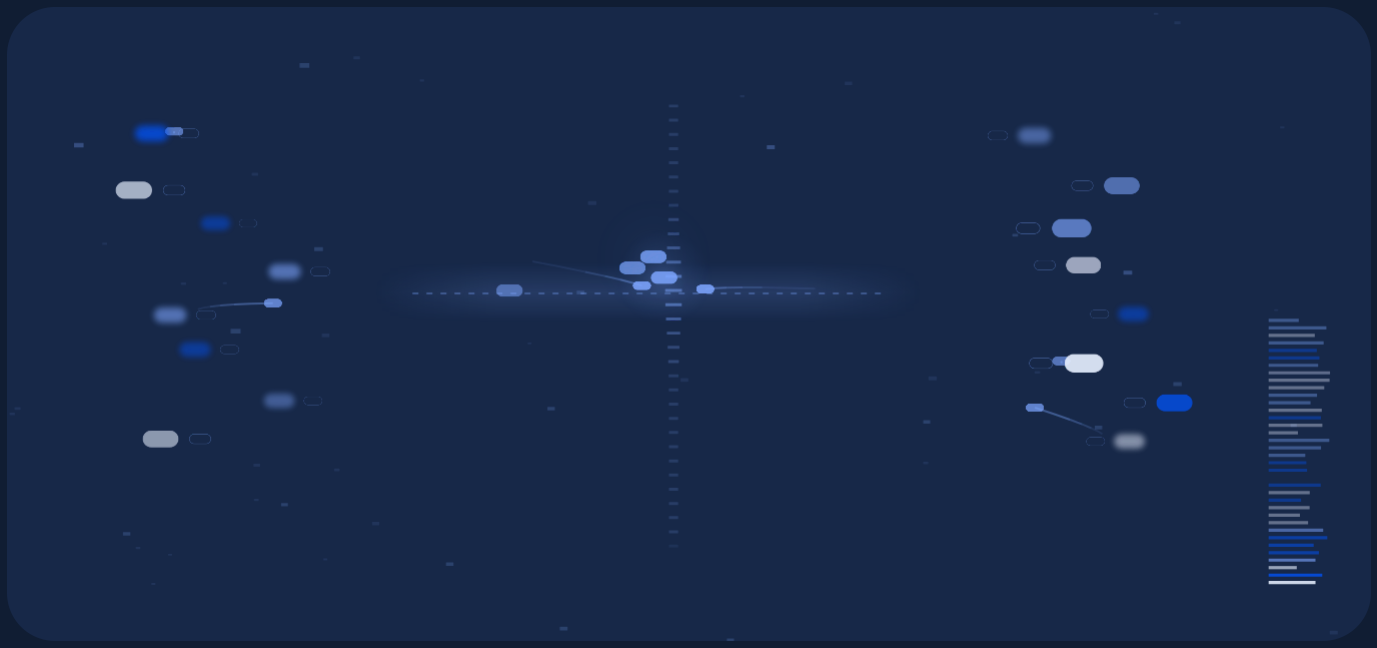
Rillor

RESEARCH REPORT

What agent benchmarks measure, and what they miss

How SWE-bench, GAIA, WebArena, OSWorld and tau-bench score LLM agents, and the validity and contamination flaws that audits have found.



7 November 2026 SI evaluation rillor.com/insights/what-agent-benchmarks-measure

Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- Major model providers cited five agent benchmarks in launch posts from 2024 to early 2025 [7]. Each scores a fixed end state with a fixed rule. SWE-bench runs repository tests on a patch, and GAIA matches a short answer. WebArena uses string matching, a model judge or state checks. OSWorld runs a per-task script, and tau-bench compares the final database with a goal state [1, 2, 3, 4, 5]. A score stands for capability only when the rule is correct and the tasks are new to the model.
- The Agentic Benchmark Checklist (ABC) found task-validity flaws in seven of ten agentic benchmarks, outcome-validity flaws in seven and reporting limits in all ten [7]. Errors run both ways. A do-nothing agent scores 38% on tau-bench's airline tasks, and broken page selectors understated one agent on OSWorld's Chrome tasks by 28 points [7].
- SWE-bench Verified has tests that pass wrong patches [10, 11], solutions that models appear to remember [13] and tests that reject correct ones [26]. In February 2026 the model developer that built it with the benchmark's authors found flawed tests in at least 59.4% of 138 audited problems and stopped reporting it [25, 26].
- The same setups score lower on fresh or hardened tasks, in two cases by about half: 39.7% against 21.3%, 43.20% against 19.25%, and 78.80% against 57.32% once evaluation-time leaks were blocked [14, 15, 17]. By October 2026 the successor, SWE-Bench Pro, had 480 of 731 public tasks repaired or retired [18].
- Scores depend on step limits, scaffolds, the number of runs and spending. On HAL, more reasoning effort gave equal or lower accuracy in 21 of 36 combinations, and two agents differed ninefold in cost for a two-point accuracy gap [21]. Nearly half of 60 benchmarks studied in 2026 were saturated [23].
- No source reviewed measures how well a public score predicts success on an organization's own tasks. In one 2025 study, an agent had an average success rate of 38% on 18 real issues, measured by the maintainers' tests, and none of 15 pull requests reviewed by hand was mergeable as-is [35].

Background

An agent benchmark has three parts: a set of tasks, an environment the agent acts in and a rule that scores the end of each attempt. The published number is the share of tasks on which the rule passes. Readers take it as a claim about capability, which holds only as far as the rule and the tasks allow.

Measurement research calls this construct validity, "having measures that represent what matters to the phenomenon" [8]. A 2025 review of 445 language model benchmarks by 29 expert reviewers found that fewer than 10% used complete real-world tasks, that 81.3% relied at least partly on exact matching, that 53.4% of papers presented evidence of construct validity and that 16.0% ran any statistical test [8]. 38.2% reused data from earlier benchmarks or exams, which the authors note raises the chance that tasks appear in pre-training data [8].

Agent benchmarks add problems of their own. A 2024 Princeton study argued that many have inadequate holdout sets, sometimes none, which has led to agents that "take shortcuts and overfit to the benchmark" [9]. It separates research from procurement: for an organization choosing an agent, "cost is the actual construct of interest" [9].

These benchmarks carry weight in the market. ABC lists SWE-bench Verified, GAIA, OSWorld, WebArena and tau-bench among the benchmarks that major model providers used in model or product launch posts between 1 January 2024 and 18 March 2025 [7]. SWE-Bench Pro's maintainers write that its scores "are widely reported in model releases and leaderboards" and that practitioners use them to choose models and scaffolds [18]. The GAIA authors named the risk in 2023: "Static benchmarks are broken benchmarks in the making" [2].

Method

What was examined. Five benchmarks, chosen because each appears in ABC's table of benchmarks used in major providers' launch posts and because together they cover code, assistant questions with tools, websites, desktop applications and policy-bound conversations [7]. For each, the report records the task set, scoring rule and the results in the original paper, then the audits, repairs and successors that followed. Cross-benchmark studies of contamination, saturation, cost and time horizons complete the picture.

Sources. The original benchmark papers [1, 2, 3, 4, 5, 6] and maintainers' posts and repositories [25, 26, 27, 28] come first. Independent audits and studies follow: ABC [7], a construct-validity review [8], a study of agent evaluation practice [9], studies of SWE-bench's tests and contamination [10, 11, 12, 13], fresh-task and successor benchmarks [14, 15, 16, 17, 18], a human audit of WebArena-Lite [19], OSWorld 2.0 [20], the HAL paper [21], a CORE-Bench case study [22], a saturation study [23] and METR's work on time horizons and usable code [24, 33, 34, 35]. Three leaderboards were read on 24 October 2026: the OSWorld-Verified results file [29], taubench.com [30] and HAL [31, 32].

How. Figures are recorded as each source states them, with its date and version. Where an abstract and the full text differ, the full text is used, as for ABC's WebArena range [7] and the SWE-Bench Illusion figures [13]. Arithmetic done for this report is labelled. Models are described by family and agent products are not named. The report uses no Rillor data and runs no benchmarks.

Findings

1. Each benchmark checks one end state

Benchmark	Tasks and environment	Scoring rule	Human baseline	Top model result in the paper
SWE-bench	2,294 GitHub issues from 12 Python repositories	Patch applies, and the task's tests pass	n/a	1.96%, a 2023 hosted model with retrieval
GAIA	466 questions needing reasoning, browsing and tools	Quasi exact match on a short answer	92%	15%, a 2023 hosted model with plugins
WebArena	812 tasks on four self-hosted websites	Exact or substring match, a model judge, or state checks	78.24%	14.41% (version of April 2024)
OSWorld	369 tasks in real desktop and web applications	Per-task script checks the final state	Over 72.36%	12.24%
tau-bench	115 retail and 50 airline tasks with a simulated user	Final database equals the goal state, and replies contain the required information	n/a	35.2% airline, 61.2% retail, a 2024 hosted model

Sources: [1, 2, 3, 4, 5]. Neither the SWE-bench nor the tau-bench paper reports a human baseline.

SWE-bench keeps only tasks with at least one test that fails before the fix and passes after it, and counts an issue resolved if the patch applies and all the task's tests pass [1]. A median of 51 further tests check that existing behavior still works [1]. Its 1.96% launch result came from a retrieval setup that generated one patch per issue [1]. GAIA withholds answers to 300 of its questions and releases a developer set of 166 with answers [2]. Each answer is a short string, number or list with one correct value [2]. WebArena runs its sites in Docker containers and scores answers by exact match, by containment or by a language model judging equivalence, with programmatic state checks for other tasks [3]. OSWorld pairs each task with a setup configuration and an evaluation script drawn from 134 evaluation functions, gives agents 15 steps in the original experiments and includes 30 infeasible tasks [4]. tau-bench scores 1 only if the final database is identical to the goal and the replies contain the needed information. Its authors note that this may be necessary without being sufficient, since an agent might issue a return without explicit user confirmation [5].

Every rule judges the end of the episode. ABC notes that OSWorld's state check "only verifies the relevant states", so agents "can potentially perform extra harmful actions" that go unchecked [7]. HAL found that web benchmarks give the same zero to an agent that abstains and to one that leaks a user's credit card details [21].

2. Most audited benchmarks fail validity checks

ABC defines rigor by two conditions [7]. Task validity: "a task should be solvable if and only if the agent possesses the target capability." Outcome validity: "the evaluation result (e.g., tests or checks) truly indicates task success." It narrowed 17 popular benchmarks to ten for in-depth assessment. Seven violated task validity, seven violated outcome validity and all ten had reporting limits. 80% did not acknowledge weaknesses in their design or implementation [7]. The abstract puts the resulting misestimation at up to 100% in relative terms [7].

Benchmark	Flaw ABC found	Effect on scores
tau-bench	38% of airline and 6% of retail tasks are intentionally impossible, and an unchanged database counts as success	A do-nothing agent scores 38% on airline and 6.0% on retail
tau-bench	2% of airline and 3.6% of retail tasks grade verbatim database text by substring matching	An agent that outputs all the data scores 40% on airline and 9.6% on retail
WebArena	Substring matching ignores extra content, and the unvalidated model judge accepts an empty reply on "N/A" tasks	1.4 to 5.2% overestimate
OSWorld	Website changes broke the HTML selectors in 13 of 46 Chrome tasks	28-point underestimate of an open-source agent on that section
SWE-bench Verified	Unit tests miss edge cases, so incorrect patches pass	11 of 45 leaderboard entries (24.4%) changed rank

Sources: [7], with the SWE-bench figure from [10].

At 50 airline tasks [5], 38% is 19 tasks (computed for this report). The do-nothing score beat the 35% of a function-calling agent on a widely used hosted model [7]. ABC's checks ask benchmarks to report statistical significance, baselines that use no model and the results of trivial agents [7]. For GAIA it records no measures against contamination and no statistical significance metrics [7].

Errors run in both directions: tau-bench and WebArena overstate success, and OSWorld's broken selectors understate it [7]. A human review of all 165 WebArena-Lite tasks, published in October 2026, recovered 5.45 to 8.49 points of success that the automatic evaluator missed for a 2026 hosted model, and found that "substantial early progress can coexist with a final failure" [19].

3. SWE-bench Verified: weak tests and remembered solutions

SWE-bench Verified was released in August 2024 by a model developer working with the SWE-bench authors [25]. 93 developers annotated 1,699 samples. 38.3% were flagged for underspecified problems and 61.1% for tests that may unfairly reject valid solutions. 68.3% were filtered out and 500 remained [25]. The post itself warned that models pre-trained on internet text "are likely to be contaminated on the tasks", and showed one hosted model scoring between 2.7% and 28.3% on SWE-bench Lite depending on the scaffold [25].

Tests that pass wrong patches. UTBoost added tests and fixed SWE-bench's log parser [10]. It found insufficient tests in 26 of the 500 tasks and erroneous patches in 92 of 584 generated patches it checked [10]. After its corrections, 11 of 45 Verified leaderboard entries (24.4%) changed rank, and the leading entry fell from 55% to 53.6%, tied for first [10]. A study of three state-of-the-art issue-solving tools found that SWE-bench validates patches only with the test files changed in the original pull request [11]. 7.8% of patches counted as correct while failing the full developer test suite, and the authors estimate an inflation of 6.4 percentage points in reported resolution rates [11]. On the original set, SWE-Bench+ found that 32.67% of one 2024 agent's successful patches used solutions given in the issue or its comments, and that over 94% of issues predated model knowledge cutoffs [12].

Remembered solutions. Asked to name the buggy file from the issue text alone, ten models scored 60% to 76% on Verified, 57% to 71% on the full SWE-bench, 50% to 68% on newly collected tasks from the same repositories and under 53% on tasks from popular repositories outside SWE-bench [13]. Given only a code prefix, eight models completed it verbatim on 11.7% to 31.6% of instances [13].

Tests that reject correct patches. In February 2026 the developer that built Verified audited 138 problems, 27.6% of the set, that one reasoning model did not solve consistently over 64 runs, each reviewed by at least six engineers [26]. At least 59.4% had flawed tests: 35.5% enforced implementation details, 18.8% checked functionality the problem never specified and 5.1% had other issues [26]. That is at least 82 problems, 16.4% of the set (computed for this report). Every frontier model it tested could reproduce the reference fix or verbatim problem details for certain tasks, and progress had slowed from 74.9% to 80.9% over six months [26]. Its conclusion: improvements "increasingly reflect how much the model was exposed to the benchmark at training time. This is why we have stopped reporting SWE-bench Verified scores, and we recommend that other model developers do so too" [26]. It recommended SWE-Bench Pro, where its checks found contamination "significantly rarer and less egregious" [26].

4. Fresh tasks and successors show the gap

Fresh-task benchmarks collect recent issues so that models are less likely to have seen them. SWE-rebench's 294 tasks from 169 repositories run every model in one minimal agent framework five times, and evaluations that include issues created before a model's release are marked as potentially contaminated [14]. SWE-bench-Live's 1,319 tasks were created between January 2024 and April 2025 in 93 repositories [15].

Setup (year evaluated)	Older or standard run	Fresh or hardened run	Source
Open-weight model family, release A (2025)	39.7% on SWE-bench Verified	21.3% on SWE-rebench, March to April 2025 issues	[14]
Same family, release B (2025)	35.2% on SWE-bench Verified	21.9% on SWE-rebench	[14]
Leading SWE-bench-Live setup (2025)	43.20% on SWE-bench Verified	19.25% on SWE-bench-Live	[15]
One model on SWE-Bench Pro (2026)	78.80%, baseline protocol	57.32%, leaks blocked	[17]

The two releases of one family score alike on fresh tasks and diverge on Verified, which the authors say "may suggest potential contamination effects on the older benchmark" [14]. Their abstract says some models' performance "might be inflated due to contamination issues" [14].

SWE-Bench Pro, released in September 2025, has 1,865 problems from 41 repositories: 731 public, 858 held out to test for overfitting and a commercial set from 18 proprietary repositories [16]. In the paper's November 2025 version, models tested in one scaffold with a 50-turn limit stayed below 45% Pass@1 on the public set [16]. Training data is one source of leaks. A 2026 preprint found that one model's baseline runs read leaked answers locally in at least 103 of the 731 public tasks and over the network in 49 [17]. Removing nested Git histories and blocking network answer access cut one model from 78.80% to 57.32%, with 186 passes turning into failures (McNemar's test, $p < 0.001$) [17].

In October 2026 the benchmark's maintainers repaired or retired 480 of the 731 public tasks (66%), using agentic audits, 404 hours of review by 23 engineers and blind re-implementation [18]. The defects included unstated test requirements, solution leakage, insufficient tests and verifier or environment faults [18]. On the 642 tasks released, graded in a clean container with network access limited to the model endpoint, frontier agents resolve up to 99.4% and open-weight, smaller and cheaper systems 87.2% to 97.7% [18]. The authors conclude that "on a correctly graded split the strongest systems fail fewer than 1% of tasks, so small differences among them are not capability differences" [18]. These figures are not comparable with the 2025 public-set results.

5. GAIA: public answers, no contamination checks

GAIA's authors argued that precise answers, their absence from pre-training data and checkable reasoning traces limit memorization, and expected the benchmark to decay through contamination or through answers disappearing from the web [2]. ABC records no anti-contamination measures [7]. HAL scores GAIA on what it calls the public validation set of 165 questions [32], while the GAIA paper releases 166 annotated developer questions with their answers [2]. HAL's top GAIA entry, viewed 24 October 2026, scored 74.5% at a cost of $178.20 with a model released in September 2025 [31].

6. WebArena and OSWorld: graders and environments that drift

WebArena keeps its pages fixed, and the Princeton study notes that it "does not model drift" and "lacks a held-out test set" [9]. In 2024 the top agent on its leaderboard scored 35.8%, more than double the paper's baseline, and hard-coded policies for specific WebArena tasks, such as building a user's profile URL directly [9]. WebArena-Verified, presented in December 2025, states that "every task, reference answer, and evaluator has been manually reviewed and corrected" and replaces the model judge and substring matching with type-aware structural comparison [27].

OSWorld runs real applications, and ABC notes that it "checks for HTML selectors (like class names or page titles) on live web pages" [7]. In July 2025 its maintainers released OSWorld-Verified after fixing more than 300 reported issues over about two months with about ten people, on top of more than 400 hours of checks before the 2024 release [28]. The issues included CAPTCHAs, website redesigns and false negatives such as equivalent spreadsheet formulas marked wrong [28]. They warned that when every team runs its own copy, some modify tasks "to achieve higher scores, making scores increasingly opaque and incomparable" [28].

The verified results file, viewed 24 October 2026, lists 144 rows with a success rate (counted for this report) [29]. The top entry, an agentic framework, scored 90.19% (325.59 of 361 tasks) at a 100-step limit on 25 July 2026, and the top general model 85.96% (307.73 of 358) on 1 August 2026 [29]. 16 rows are at or above the paper's 72.36% human figure, measured on the original tasks (counted for this report from [29, 4]). Step limits matter: one model released in 2025 scored 9.1%, 17.17% and 23% at 15, 50 and 100 steps [29]. OSWorld 2.0, released in June 2026, has 108 workflows that take people a median of about 1.6 hours. The top model completed 20.6% at 500 steps [20]. Its authors report that agents lose track of constraints, miss information that arrives mid-task, guess instead of asking the user and skip verification [20].

7. tau-bench: impossible tasks, a simulated user and pass^k

pass^k exposes inconsistency. In June 2024 the leading function-calling agent passed over 60% of retail tasks on one trial, and its pass^8 fell below 25% [5]. The simulated user is part of what is measured. The maintainers' 2025 follow-up, tau2-bench, found that the simulated user made task-critical errors in 13 of 100 airline conversations, 6 of 50 retail and 3 of 50 telecom [6]. It refined task instructions and split evaluation into environment, communication, natural language and action assertions [6].

taubench.com records that in February 2026 the maintainers "audited and fixed 50+ tasks across the airline and retail domains", including impossible constraints [30]. On the leaderboard viewed 24 October 2026, the top tau2-bench text entry scored 87.9% on pass^1 and the top entry on the newer banking domain 55.2%, with no evaluation dates shown [30]. HAL found that the few-shot agent in the official repository included benchmark examples in its few-shot data [21]. On HAL's airline board, two 2025 models tie at 56.0%, one at $42.11 and the other at $11.36 [31].

8. Saturation, single runs, scaffolds and cost

A 2026 study defines saturation as "the loss of reliable discriminative power among top-performing models under comparison" [23]. Of 60 language model benchmarks, 29 showed high or very high saturation, and the saturated share rose from 42.9% for benchmarks under 24 months old to 54.5% for those over 60 months [23]. Private test sets did not reduce saturation systematically once age was considered [23].

HAL ran 21,730 rollouts across 9 models and 9 benchmarks for about $40,000 [21]. More reasoning effort gave equal or lower accuracy in 21 of 36 combinations, two web agents cost $171 and $1,577 for a two-point accuracy difference, and in AssistantBench, CORE-Bench and SciCode, agents met an environmental barrier, such as a crashing browser, in about 40% of failed tasks [21]. Its logs showed eight cases of agents finding a gold answer online, in a dataset repository or on arXiv [21]. The Princeton study warned that accuracy "can be improved by scientifically meaningless methods such as retrying" [9]. SWE-rebench's authors name reporting pass@N as pass@1, and reporting only the best run, as practices that overstate capability [14].

HAL has declared CORE-Bench solved and paused adding new models [31]. A June 2026 case study found 15 task-level errors and 20 tasks with exploitable shortcuts in CORE-Bench Hard through log analysis after saturation [22]. On the corrected version the top agent scored 100% and the next four 97.4%, and one model's score differed by about 44 points between two scaffolds [22].

9. Time horizons and the gap to usable work

METR's time horizon is the task length, in human expert time, that a model completes with 50% success [24]. In the July 2026 version of its paper, a reasoning model released in April 2025 had a horizon of around 110 minutes, frontier horizons had doubled about every seven months since 2019, and performance was much lower on "messier" tasks [24]. Its estimates from SWE-bench Verified doubled in around 70 days, against 143 days on METR's suite for 2024 models [24]. The suite itself moves the measure. Time Horizon 1.1 grew it from 170 to 228 tasks, measured human times for only 5 of its 31 tasks of eight hours or longer, and shifted individual model estimates by 7% to 57% [34]. METR's page, last updated 8 September 2026, states that "Measurements above 16 hrs are unreliable with our current task suite" and notes that most jobs "require interacting with other people and involve success metrics that cannot be algorithmically scored" [33].

A 2025 METR study tested the gap directly [35]. On 18 real tasks from two open-source repositories, an early-2025 agent passed the maintainers' tests on 38% of attempts (plus or minus 19 points). None of the 15 pull requests reviewed by hand was mergeable as-is, and test-passing runs needed about 26 minutes of fixes on

average [35]. METR concluded that "automatic scoring used by many benchmarks may overestimate AI agent real-world performance" [35].

Implications

For buyers and analysts reading a launch claim. A benchmark figure is a measurement under stated conditions, and the audits show how far conditions move it: an overestimate of 1.4 to 5.2% from WebArena grading [7], 6.4 points from SWE-bench patch validation [11], 21.48 points from evaluation-time leaks [17] and 9.1% to 23% for one model as OSWorld's step limit rose from 15 to 100 [29]. Gaps of a few points between models on one benchmark sit inside those ranges. A claim is easier to read when it states the benchmark version, the number of tasks, the scaffold, the step limit, the number of runs and whether the figure is pass@1 or a best-of [14, 16, 29]. In its co-creator's assessment, gains on SWE-bench Verified now increasingly reflect exposure in training [25, 26].

For engineers choosing a benchmark. Match task type first: SWE-bench and its successors for code, WebArena for websites, OSWorld for desktop applications, tau-bench for policy-bound conversations and GAIA for questions that need search and tools [1, 2, 3, 4, 5]. Repaired versions exist: WebArena-Verified [27], OSWorld-Verified [28], tau2-bench with the 2026 fixes [6, 30] and SWE-Bench Pro with its 2026 revision [16, 18]. ABC's reporting checks set a floor for internal runs: a trivial-agent baseline, a baseline that uses no model and statistical significance [7]. Repeated trials scored with pass^k show inconsistency that single runs hide [5], and log review found problems in HAL and CORE-Bench that scores alone did not show [21, 22].

For teams deploying an agent. A public benchmark measures its authors' tasks. No source reviewed links a public score to success on a particular organization's workflow, and METR's study shows that test-passing work can still need human repair [35]. The method that follows from the audits is a fixed set of the organization's own tasks with known answers, scoring code checked against cases where the agent should fail, and reruns whenever the model version changes. Rillor's agent evaluation and simulation (https://rillor.com/si/evaluation) work tests the agents an organization plans to deploy on its own tasks, with benchmark suites of fixed tasks and scoring code rerun on each new model version so results stay comparable over time.

For benchmark builders and researchers. The maintainers' lessons converge: grade in a clean container, version the benchmark and release every grade [18], hold out tasks or date fresh ones against model releases [14, 16], and budget for upkeep. OSWorld needed more than 300 fixes after 400 hours of checks [28], and SWE-Bench Pro needed 480 repairs or retirements within about a year [18]. Report uncertainty, which only 16.0% of 445 reviewed benchmarks did through any statistical test [8].

Limits of this analysis

- The report covers five benchmarks chosen for their use in launch posts from January 2024 to March 2025 [7]. Others appear only through ABC, HAL and the saturation and CORE-Bench studies.
- Every contamination measurement found is on SWE-bench and its relatives [10, 12, 13, 14, 15, 17, 26]. No source read measured contamination on GAIA, WebArena, OSWorld or tau-bench.
- Several sources are preprints. The 2026 SWE-Bench Pro revision is by its own maintainers [18]. The WebArena-Lite audit [19] and SWE-Bench+ [12] were read at abstract level.
- Leaderboards change. taubench.com shows no evaluation dates [30], HAL has paused adding models [31], METR's page is no longer actively updated [33] and the OSWorld-Verified rows run to 1 August 2026 [29]. Each figure holds for the date given.
- No source read gives the effect of the February 2026 tau-bench fixes or of WebArena-Verified's changes on published scores. None says whether ABC's impossible airline tasks still score an unchanged state as success.
- The construct-validity review [8] and the saturation study [23] cover language model benchmarks in general.
- METR's mergeability study covers 18 tasks and one early-2025 agent [35].
- Models are described by family.

Sources

1. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press and Karthik Narasimhan, ICLR 2024, arXiv:2310.06770, version 3, 11 November 2024 (first version 10 October 2023). https://arxiv.org/abs/2310.06770

2. GAIA: a benchmark for General AI Assistants. Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun and Thomas Scialom, arXiv:2311.12983, 21 November 2023. https://arxiv.org/abs/2311.12983

3. WebArena: A Realistic Web Environment for Building Autonomous Agents. Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar and others, arXiv:2307.13854, version 4, 16 April 2024 (first version 25 July 2023). https://arxiv.org/abs/2307.13854

4. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. Tianbao Xie, Danyang Zhang, Jixuan Chen and others, arXiv:2404.07972, version 2, 30 May 2024 (first version 11 April 2024). https://arxiv.org/abs/2404.07972

5. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. Shunyu Yao, Noah Shinn, Pedram Razavi and Karthik Narasimhan, arXiv:2406.12045, 17 June 2024. https://arxiv.org/abs/2406.12045

6. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. Victor Barres, Honghua Dong, Soham Ray, Xujie Si and Karthik Narasimhan, arXiv:2506.07982, 9 June 2025. https://arxiv.org/abs/2506.07982

7. Establishing Best Practices for Building Rigorous Agentic Benchmarks. Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre and others, arXiv:2507.02825, version 5, 7 August 2025 (first version 3 July 2025). https://arxiv.org/abs/2507.02825

8. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou and others, NeurIPS 2025 Datasets and Benchmarks Track, arXiv:2511.04703, 3 November 2025. https://arxiv.org/abs/2511.04703

9. AI Agents That Matter. Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir and Arvind Narayanan, arXiv:2407.01502, 1 July 2024. https://arxiv.org/abs/2407.01502

10. UTBoost: Rigorous Evaluation of Coding Agents on SWE-Bench. Boxi Yu, Yuxuan Zhu, Pinjia He and Daniel Kang, arXiv:2506.09289, 10 June 2025. https://arxiv.org/abs/2506.09289

11. Are "Solved Issues" in SWE-bench Really Solved Correctly? An Empirical Study. You Wang, Michael Pradel and Zhongxin Liu, arXiv:2503.15223, version 2, 9 September 2025 (first version 19 March 2025). https://arxiv.org/abs/2503.15223

12. SWE-Bench+: Enhanced Coding Benchmark for LLMs. Reem Aleithan, Haoran Xue, Mohammad Mahdi Mohajer, Elijah Nnorom, Gias Uddin and Song Wang, arXiv:2410.06992, 9 October 2024. https://arxiv.org/abs/2410.06992

13. The SWE-Bench Illusion: When State-of-the-Art LLMs Remember Instead of Reason. Shanchao Liang, Spandan Garg and Roshanak Zilouchian Moghaddam, arXiv:2506.12286, version 4, 1 December 2025 (first version 14 June 2025). https://arxiv.org/abs/2506.12286

14. SWE-rebench: An Automated Pipeline for Task Collection and Decontaminated Evaluation of Software Engineering Agents. Ibragim Badertdinov, Alexander Golubev, Maksim Nekrashevich and others, NeurIPS 2025, arXiv:2505.20411, version 2, 4 November 2025 (first version 26 May 2025). https://arxiv.org/abs/2505.20411

15. SWE-bench Goes Live! Linghao Zhang, Shilin He, Chaoyun Zhang and others, arXiv:2505.23419, version 2, 1 June 2025 (first version 29 May 2025). https://arxiv.org/abs/2505.23419

16. SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks? Xiang Deng, Jeff Da, Edwin Pan and others, arXiv:2509.16941, version 2, 14 November 2025 (first version 21 September 2025). https://arxiv.org/abs/2509.16941

17. SWE-Bench Pro Verified: A Reliable Benchmark for Software Engineering Agents. Pujun Zheng, Zixin Shang, Shufan Jiang, Wenhui Tian, Dongsheng Zhu, Zerun Ma, Dingbo Yuan and Qi Zhang, arXiv:2609.08149, version 3, 9 October 2026 (first version 8 September 2026). https://arxiv.org/abs/2609.08149

18. SWE-Bench Pro V2: A Cleaner, Harder-to-Game Coding Benchmark. Soham Dan, Jeff Da, Sanyam Bhutani and others, arXiv:2610.24206, 16 October 2026. https://arxiv.org/abs/2610.24206

19. Auditing Web Agent Evaluation on WebArena-Lite: Human Review of Outcomes and Trajectories. Chengguang Gan, Zimeng He, Yoshihiro Tsujii, Ken-ichiro Kobayashi, Hiroki Itoh and Kotaro Funakoshi, arXiv:2610.01491, 1 October 2026. https://arxiv.org/abs/2610.01491

20. OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks. Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong and others, arXiv:2606.29537, version 2, 13 July 2026 (first version 28 June 2026). https://arxiv.org/abs/2606.29537

21. Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation. Sayash Kapoor, Benedikt Stroebl, Peter Kirgis and others, arXiv:2510.11977, 13 October 2025. https://arxiv.org/abs/2510.11977

22. Life After Benchmark Saturation: A Case Study of CORE-Bench. Nitya Nadgir, Sayash Kapoor, Kangheng Liu, Peter Kirgis, Matilda Orona, Stephan Rabanser, Tilman Bayer and others (last author Arvind Narayanan), arXiv:2606.26158, 23 June 2026. https://arxiv.org/abs/2606.26158

23. When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation. Mubashara Akhtar, Anka Reuel, Prajna Soni and others, ICML 2026, arXiv:2602.16763, version 4, 6 August 2026 (first version 18 February 2026). https://arxiv.org/abs/2602.16763

24. Measuring AI Ability to Complete Long Software Tasks. Thomas Kwa, Ben West, Joel Becker and others, METR, arXiv:2503.14499, version 4, 10 July 2026 (first version 18 March 2025). https://arxiv.org/abs/2503.14499

25. Introducing SWE-bench Verified. OpenAI, 13 August 2024, updated 24 February 2025 (Internet Archive capture of 1 January 2026). https://web.archive.org/web/20260101233457/https://openai.com/index/introducing-swe-bench-verified/

26. Why SWE-bench Verified no longer measures frontier coding capabilities. OpenAI, 23 February 2026 (Internet Archive capture of 15 October 2026). https://web.archive.org/web/20261015182204/https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/

27. WebArena-Verified, repository README. ServiceNow, viewed 24 October 2026. https://github.com/ServiceNow/webarena-verified

28. Introducing OSWorld-Verified. XLANG Lab, 28 July 2025. https://xlang.ai/blog/osworld-verified

29. OSWorld-Verified results file (rows dated up to 1 August 2026). XLANG Lab, viewed 24 October 2026. https://os-world.github.io/static/data/osworld_verified_results.xlsx

30. τ-bench site and leaderboard. Sierra, viewed 24 October 2026. https://taubench.com/

31. HAL: Holistic Agent Leaderboard. SAgE team, Princeton University, viewed 24 October 2026. https://hal.cs.princeton.edu/

32. HAL: GAIA benchmark page. SAgE team, Princeton University, viewed 24 October 2026. https://hal.cs.princeton.edu/gaia

33. Task-Completion Time Horizons of Frontier AI Models. METR, last updated 8 September 2026, viewed 24 October 2026. https://metr.org/time-horizons/

34. Time Horizon 1.1. METR, 29 January 2026. https://metr.org/blog/2026-01-29-time-horizon-1-1/

35. Research Update: Algorithmic vs. Holistic Evaluation. David Rein, METR, 13 August 2025. https://metr.org/blog/2025-08-12-research-update-towards-reconciling-slowdown-with-time-horizons/

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. What agent benchmarks measure, and what they miss. 7 November 2026. https://rillor.com/insights/what-agent-benchmarks-measure

Notices

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send us the tasks your agent handles, the model versions it runs on and the decisions it is trusted with. Contact us at rillor.com/contact?type=evaluation or contact@rillor.com. More on SI evaluation: rillor.com/si/evaluation