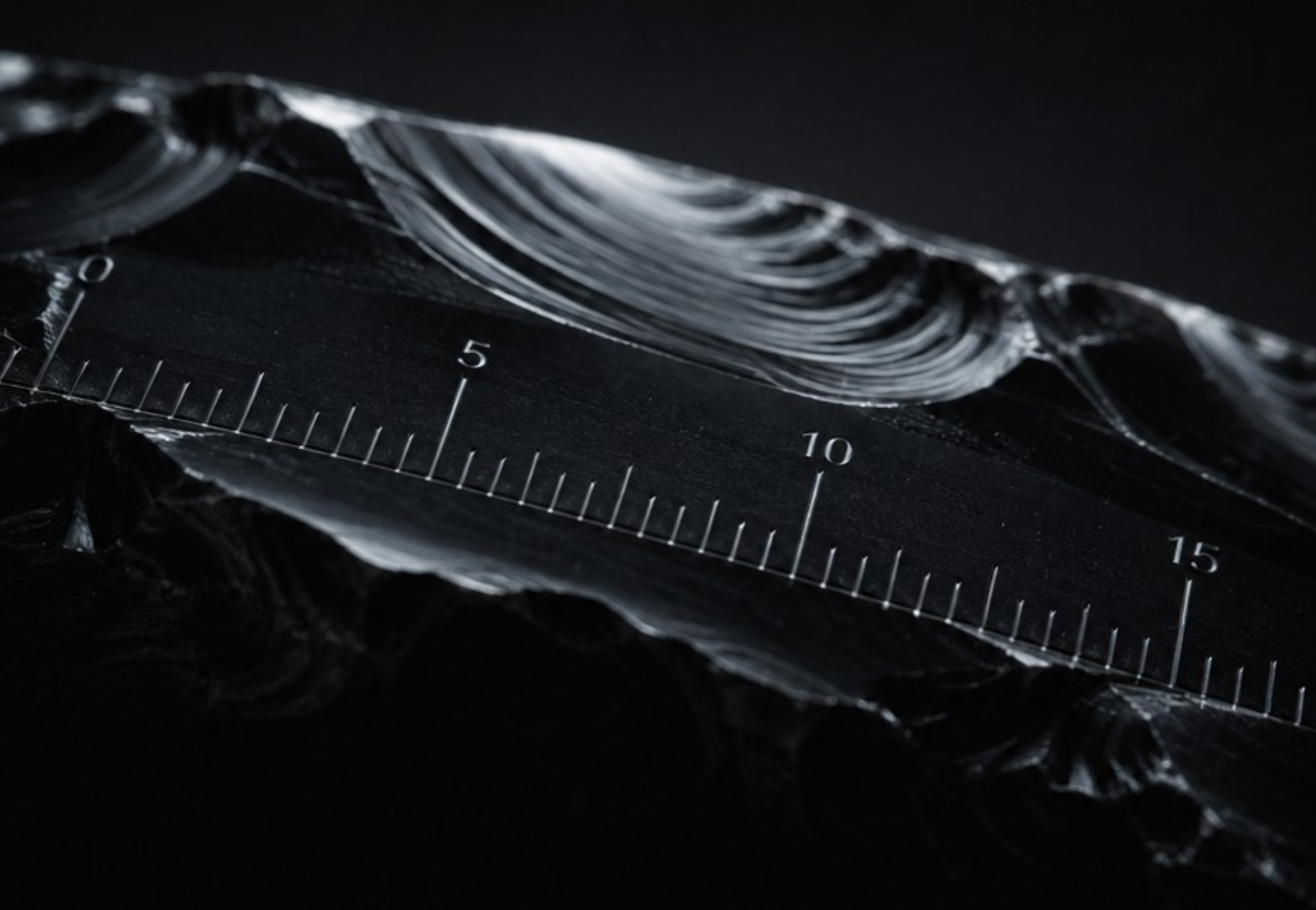

Claude Opus 5.5

The launch benchmark table, read line by line

ANTHROPIC · 22 SEP 2026 · ANTHROPIC.COM/CLAUDE-OPUS-5-5

How to read an AI benchmark table

Every row of the Opus 5.5 launch table decoded, plus a plain-English field guide to the tests every lab reports.

five checks, nine rows, seventeen tests

Luis V

THE SHORT VERSION

One grid, many experiments

A launch benchmark table looks like one experiment. It is a stack of different exams, graded by different referees, under different settings, sharing one grid.

Product teams read these tables the way they read a price list: find the biggest number and back it. That works when every cell was produced the same way. In a modern launch table almost none were. On the Claude Opus 5.5 page (22 September 2026), nine rows compare five models, and the footnotes name at least three different referees: Anthropic's own setup, figures "as reported by OpenAI", and runs by Zapier.

BENCHMARK
A fixed set of tasks with an automatic or judged score. An exam for models.

Nothing on that page is hidden. The footnotes are careful and specific. They are also where the real answer lives, and almost nobody reads them. This guide reads them for you, then turns them into a method you can reuse on any launch.

The five checks

1. **Unit.** What kind of number is it? Percent solved, an Elo rating, a similarity score, a pass-every-time rate.
2. **Referee.** Who ran it? The lab, a competitor's own report, a third party, or a test only one company can see.
3. **Settings.** Effort level, tools allowed, number of tries, which harness.
4. **Noise.** Is the gap bigger than the run-to-run wobble?
5. **Fine print.** Dashes, "partial", "with tools", and anything the footnotes say changed the run.

Run them in that order. The first two take seconds and eliminate most bad comparisons before you do any maths.

the first two checks take ten seconds

How to use this guide

Pages 3 to 5 decode the Opus 5.5 table: the full grid, every footnote in plain words, and the noise maths. Pages 6 and 7 are a field guide to the 17 benchmarks that show up most across 2026 launches. Page 8 shows which launches report which tests. Page 9 is the test to run before you switch, and a one-page checklist.

HARNESS
The code wrapped around a model during a test: prompts, tools, retries, time limits. Change it and the score moves.

THE WORKED EXAMPLE

The Opus 5.5 table, all nine rows

Values as published on the launch page. Rings mark the cells the rest of this guide comes back to. Superscripts point to the page's numbered footnotes, decoded overleaf.

BENCHMARK	OPUS 5.5	FABLE 5.1	OPUS 5	GPT-6 ASTRA	GPT-5.6 SOL
Terminal-Bench 4.0 [1] Agentic coding	66.4%	55.8%	52.3%	57.9%	37.3%
FrontierCode v1.1 (Main) Agentic coding	54.4%	50.3%	48.0%	53.3%	47.5%
CursorBench 4.0 Agentic coding	57.8%	51.8%	46.6%	—	41.7%
GDPval-AA v2.1 Knowledge work	1846	1735	1708	1542	1588
AutomationBench [2] Business workflows	40.0%	31.4%	26.9%	41.4%	28.8%
Humanity's Last Exam Multidisciplinary reasoning	67.7%†	65.6%†	63.6%†	57.2%†	—
Terminal-Bench-Science 0.1 [3] Agentic scientific research	58.7%	52.6%	29.0%	64.6%	22.4%
OSWorld 2.0 Computer use	81.8%‡	80.7%‡	74.0%‡	—	—
Chartography Visual chart recognition	89.0%†	88.4%†	83.4%†	—	—

† with tools · ‡ partial · — not reported · Unless noted, Opus 5.5 at adaptive thinking, max effort

1846 sits in a column of percentages

Three things jump out before any maths. First, one row is not in percent at all: GDPval-AA is an Elo rating, so 1846 cannot be compared with 66.4% or read as "out of" anything. Second, the dashes are uneven: OSWorld 2.0 and Chartography only have Claude numbers, so those rows compare Anthropic with itself. Third, the two ringed Terminal-Bench-Science cells are the one place a competitor leads by several points, and the page publishes error bars wide enough to swallow most of that gap (page 5).

None of this makes the table wrong. It makes it a map of where to look, not a verdict.

ADAPTIVE THINKING, EFFORT

The model decides how long to reason before answering; the effort setting caps how much. Higher effort, more tokens, higher bill.

CHECKS 02, 03 AND 05

The footnotes, translated

What the page says, paraphrased closely, then what it means for how you read the row.

SETTINGS, ALL ROWS

Unless noted, Opus 5.5 ran with adaptive thinking at max effort. On Terminal-Bench 4.0, Opus 5.5 is shown at xhigh effort and GPT-6 Astra at high effort (the OpenAI figure), each described as that model's highest score.

In plain words: Every Opus 5.5 number is a ceiling. The cost of reaching it is not in the table.

SAFEGUARDS, ALL ROWS

Opus 5.5 was tested with its production safeguards on. When they stepped in, cybersecurity tasks were finished by Claude Opus 4.8, and biology and frontier-LLM-development tasks by Claude Opus 5. Anthropic says this likely lowers Opus 5.5's scores.

In plain words: A few tasks in some rows were completed by an older model. Anthropic states the bias runs downward, against itself.

1 · TERMINAL-BENCH 4.0

Standard error ±2.6 points for Opus 5.5 and ±1.6-2 for the other Claude models. The public leaderboard (5 trials per task, Claude Code harness) shows Opus 5 at 51.8%; Anthropic's setup reproduces 52.3%. GPT-6 Astra and GPT-5.6 Sol figures are as reported by OpenAI.

In plain words: Anthropic checked its harness against the public board and landed within half a point. That is the good practice to look for. The OpenAI columns were not rerun.

2 · AUTOMATIONBENCH

Run and reported by Zapier, without fallback models, so any safeguard intervention counted as a failure. Opus 5.5 was scored in Zapier's own early-access evaluation; Opus 5, GPT-5.6 Sol and GPT-6 Astra come from Zapier's public leaderboard.

In plain words: Independent referee, which is good. But two different runs share one row, and Opus 5.5 is penalised for its own safety layer.

3 · TERMINAL-BENCH-SCIENCE 0.1

Standard error ±3.5–5 points per model. The public leaderboard (3 trials per task) shows Opus 5 at 30.0%; Anthropic's setup gives 29.0%. The GPT-6 Astra figure is as reported by OpenAI.

In plain words: The noisiest row in the table, and the one where the headline gap is smallest relative to its error bars.

LABELS · (WITH TOOLS) AND A DASH

Humanity's Last Exam and Chartography scores are marked "with tools". The page does not list which tools.

In plain words: The model could call outside tools while answering, so these are not comparable with no-tools scores elsewhere. And a dash anywhere means no score was reported: not a zero, not a loss.

LABEL · (PARTIAL)

OSWorld 2.0 scores are marked "partial". The benchmark's maintainers publish two scores: binary completion, and a partial score that credits individual requirements met along the way, averaging 27.25 checkpoints per task.

In plain words: The table shows the partial-credit score, which can never be lower than full completion. 81.8% does not mean 81.8% of jobs finished.

CHECKS 04 AND 01

Is the gap bigger than the wobble?

Rerun a model on the same benchmark and the score moves: agents take different paths, tasks time out. The standard error estimates that wobble. Comparing two models needs the wobble of the **gap**, which is larger than either one's own:

$$\text{error of the gap} = \sqrt{(\text{error}_A^2 + \text{error}_B^2)}$$

STANDARD ERROR
How much a score would move if you reran the same test. ±2.6 means reruns usually land within a few points.

Rule of thumb: a gap under about twice that number is not a reliable win. This assumes independent runs.

ROW · COMPARISON	SCORES	GAP (PTS)	ERROR OF GAP	GAP ÷ ERROR	READING
Terminal-Bench 4.0 Opus 5.5 vs Opus 5	66.4 vs 52.3	14.1	3.1 to 3.3	4.3 to 4.6	Clear gain
Terminal-Bench-Science 0.1 Opus 5.5 vs Opus 5	58.7 vs 29.0	29.7	4.9 to 7.1	4.2 to 6.0	Clear gain
Terminal-Bench-Science 0.1 GPT-6 Astra vs Opus 5.5	64.6 vs 58.7	5.9	4.9 to 7.1	0.8 to 1.2	Unresolved
Terminal-Bench 4.0 Opus 5.5 vs GPT-6 Astra	66.4 vs 57.9	8.5	unknown	—	Can't compute
Chartography Opus 5.5 vs Fable 5.1	89.0 vs 88.4	0.6	unknown	—	Treat as tie

The **generation gain** within Anthropic's line clears the noise on both Terminal-Bench rows. The **competitive gap** on Terminal-Bench-Science does not: a 5.9-point lead is about one error of the gap. Where no error is published, the honest reading is "unknown".

no error bar published means no verdict

Reading an Elo gap

GDPval-AA ranks models by blind head-to-head matchups judged by an AI model. Standard Elo maths turns a gap into an expected win rate. Artificial Analysis does not publish this conversion; treat it as a reading aid.

ELO
The chess rating system. The raw number means nothing; only the gap between two ratings does.

ELO GAP	50	100	111	138	304
Expected wins	57%	64%	65%	69%	85%
In this table	·	·	vs Fable 5.1	vs Opus 5	vs GPT-6 Astra

FIELD GUIDE · 1 OF 2

The benchmarks you'll keep seeing

Seventeen tests that recur across 2026 launch pages, in plain words: what a single task looks like, what the score is, who runs it, and the trap to watch for. Grouped by the kind of work they stand in for.

CODING AGENTS

Terminal-Bench

STANFORD & LAUDE INSTITUTE · % OF TASKS SOLVED

An agent is dropped into a computer's command line and must finish a real job there. Automated tests check the result.

Watch for: Version numbers. Labs report 2.0, 2.1 and 4.0; they are different exams.

SWE-bench Verified

SWE-BENCH TEAM · % OF ISSUES RESOLVED

Fix a real GitHub issue in one of 12 Python projects so the hidden tests pass. 500 human-checked tasks.

Watch for: OpenAI stopped reporting it in February 2026, citing flawed tests and training-data leakage.

SWE-Bench Pro

SCALE AI · PASS@1

Long coding tasks that could take a professional hours or days, often touching many files. 1,865 problems across 41 projects.

Watch for: Labs usually report only the public subset.

DeepSWE

DATACURVE AI · % SOLVED

Original long coding tasks written from scratch in 91 open-source projects across 5 languages.

Watch for: Newer and smaller (113 tasks), so expect wider error bars.

CursorBench

CURSOR · CORRECTNESS SCORE

Coding requests drawn from real Cursor sessions, scored for correctness against tokens used.

Watch for: Internal: only Cursor can run it or see the tasks.

AGENTS USING COMPUTERS AND TOOLS

OSWorld

XLANG LAB, HKU · SUCCESS RATE

Operate a real desktop to finish a job, such as editing a spreadsheet. Version 2.0 has 108 long tasks.

Watch for: OSWorld 2.0 has a full-completion score and a partial-credit score. Check which one you are reading.

MCP-Atlas

SCALE AI · PASS RATE

Answer a request by picking and chaining the right tools across 36 real MCP servers, without being told which to use.

Watch for: Half the 1,000 tasks are private, so outside reruns cover only the public half.

Toolathlon

HKUST NLP · SUCCESS RATE

Finish multi-app jobs, such as email plus calendar plus files, using 604 tools across 32 apps. About 20 tool calls per task.

Watch for: Only 108 tasks: a few points is a few tasks.

τ2-bench

SIERRA RESEARCH · PASS^K

Act as a customer-service agent, for example telecom support, while a simulated customer also takes actions.

Watch for: pass^k means succeeding on all k tries. It rewards consistency, so it runs lower than one-try scores.

FIELD GUIDE · 2 OF 2

SEARCH AND LONG CONTEXT

BrowseComp

OPENAI · ACCURACY

Find a hard-to-locate fact by browsing the web. Human trainers solved 29.2% and gave up on the rest within two hours.

Watch for: Short single answers; says little about open-ended research.

MRCR v2

OPENAI · TEXT-SIMILARITY SCORE

A very long chat repeats the same request several times; reproduce one specific instance of it.

Watch for: Labs pick different needle counts and context lengths. Check both before comparing.

REAL WORK

GDPval

OPENAI · WIN RATE VS EXPERTS

Produce a real deliverable (a document, deck or spreadsheet) for one of 44 occupations. Expert graders compare it blind with a professional's version.

Watch for: One-shot: no back-and-forth or clarifying questions.

GDPval-AA

ARTIFICIAL ANALYSIS · ELO RATING

The public GDPval tasks, run with shell and web access; two models' outputs are compared and an AI judge picks the winner.

Watch for: An Elo, not a percentage, and the grader is a model.

REASONING AND KNOWLEDGE EXAMS

Humanity's Last Exam

CAIS & SCALE AI · ACCURACY

2,500 hard expert-written questions across 100+ subjects, some with images.

Watch for: Labs report "no tools" and "with tools" separately. Never mix them.

GPQA Diamond

REIN ET AL. · ACCURACY

198 PhD-level multiple-choice questions in biology, physics and chemistry, built so web search does not help.

Watch for: Epoch AI flagged it as nearing saturation in 2025: top scores bunch up and stop separating models.

ARC-AGI

ARC PRIZE FOUNDATION · ACCURACY, WITH COST

Infer a hidden rule from a few coloured-grid puzzles and apply it (v2), or work out the goal of an unfamiliar game (v3).

Watch for: Versions differ wildly. On v3, frontier AI scored 0.51% at launch in March 2026.

IMAGES

MMMU-Pro

MMMU TEAM · ACCURACY

College-level multiple-choice questions that depend on an image, with 10 options; in one setting the question itself is inside the image.

Watch for: Also reported with and without tools.

Two failure modes apply to every entry. A test **saturates** when the top scores bunch near the ceiling, and it **leaks** when its questions end up in training data. SWE-bench Verified is the cautionary example: OpenAI reported that at least 59.4% of the problems it audited had flawed tests, and that every frontier model it tried could reproduce original fixes. A rising score on a leaked test measures memory, not skill.

SATURATION
When top models all score near the ceiling, so the test can no longer tell them apart.

CONTAMINATION
When test questions leak into training data, so a model can remember answers instead of working them out.

ACROSS LAUNCHES

Who reports what

Seven launch pages from 2026, checked for which benchmarks each one reports. A version label means the page names that version; a tick means the benchmark appears without a version we could pin down.

BENCHMARK	OPUS 5.5	GPT-5.6	GPT-5.5	GEMINI 3.1 PRO	GEMINI 3.8 FLASH	GROK 4.7	DEEPSEEK-V4-PRO	PAGES
Terminal-Bench	4.0	2.1	2.0	2.0	·	4.0	2.0	6
GDPval / GDPval-AA	AA v2.1	AA v2	✓	AA	·	✓	AA	6
Humanity's Last Exam	✓	·	✓	✓	Verified	·	✓	5
SWE-Bench Pro	·	✓	✓	✓	·	·	✓	4
GPQA Diamond	·	✓	✓	✓	·	·	✓	4
BrowseComp	·	✓	✓	✓	·	·	✓	4
MRCR v2	·	✓	✓	✓	·	·	✓	4
OSWorld	2.0	2.0	Verified	·	·	·	·	3
MMMU-Pro	·	✓	✓	✓	·	·	·	3
ARC-AGI	·	v3	v1, v2	v2	·	·	·	3
DeepSWE	·	✓	·	·	✓	✓	·	3
CursorBench	4.0	✓	·	·	·	4.0	·	3
MCP-Atlas	·	·	✓	✓	·	·	✓	3
Toolathlon	·	✓	✓	·	·	·	✓	3
τ2-bench	·	·	Telecom	✓	·	·	·	2
SWE-bench Verified	·	·	·	✓	·	·	✓	2

same name, different exam

There is no shared scoreboard. Only a handful of names appear on three or more of these seven pages, and even the most common one, Terminal-Bench, shows up as three different versions. A 2.0 score and a 4.0 score are not comparable, however similar the name.

What a lab leaves out is also information. When a page drops a widely reported test that its predecessor carried, it is worth asking why, and checking the benchmark's own leaderboard.

Launches: Claude Opus 5.5 (22 Sep 2026) · GPT-5.6 (9 Jul 2026) · GPT-5.5 (23 Apr 2026) · Gemini 3.1 Pro model card (19 Feb 2026) · Gemini 3.8 Flash (2 Sep 2026) · Grok 4.7 (21 Sep 2026) · DeepSeek-V4-Pro (Apr 2026). Meta's latest launch page was not checked.

WHAT TO DO ON MONDAY

The test to run before you switch

A launch table tells you what to test, not what to buy. The decision turns on score against cost per task at the setting you will actually run.

1. **Pick 30 tasks** from your own backlog or logs that look like the work you are buying the model for. Write the pass condition before you run anything.
2. **Fix the settings** you will pay for in production: effort level, tools, time limit. Do not test at max effort if you will ship at medium.
3. **Run each task 3 times** per model. The spread between runs is your own standard error.
4. **Record pass rate and cost per task**, not cost per token. More effort means more tokens per task.
5. **Decide by the gap.** If it is smaller than your run-to-run spread, call it a tie and pick on price, latency or vendor fit.

what would change my mind: a gap that survives reruns

ONE-PAGE CHECKLIST · ANY LAUNCH TABLE

- ☐ **Unit.** Percent solved, Elo, similarity score, pass^k? Never compare across units.
- ☐ **Version.** Same benchmark version in every column?
- ☐ **Referee.** Lab-run, “as reported by” a rival, third party, or internal-only?
- ☐ **Settings.** Effort level, tools, number of trials, harness. Same across columns?
- ☐ **Noise.** Error bars published? Is the gap more than about twice the error of the gap?
- ☐ **Fine print.** Dashes, “partial”, “with tools”, safeguard fallbacks, subsets.
- ☐ **Freshness.** Is the test saturated, or has it leaked into training data?
- ☐ **Fit.** Does one task in this benchmark look like one task in your job?

LIMITS AND SOURCES

What this guide leaves out

It reads the table; it does not rerun it. Every number here is the launch page's own. The noise maths uses the page's published errors and assumes independent runs; it is a reading aid, not a significance test.

The field guide compresses. Each benchmark has variants, subsets and settings that change its score. The one-line "watch for" is the trap most likely to mislead a buyer, not the full list.

A clean benchmark can still be the wrong one. No checklist fixes a test that doesn't look like your work. That is what the 30-task test on page 9 is for.

No ranking is implied. Nothing here says which model is better. It says which comparisons in the table hold up and which need your own data.

Sources

1. **Claude Opus 5.5 launch page, Anthropic, 22 Sep 2026.** https://www.anthropic.com/claude-opus-5-5
2. **GDPval-AA v2.1 leaderboard and method, Artificial Analysis.** https://artificialanalysis.ai/evaluations/gdpval-aa
3. **OSWorld 2.0 leaderboard and scoring, Snorkel AI / XLANG Lab.** https://snorkel.ai/leaderboard/os-world-2-0/
4. **Why we no longer evaluate SWE-bench Verified, OpenAI, 23 Feb 2026.** https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/
5. **Terminal-Bench.** https://www.tbench.ai/ · arXiv:2601.11868
6. **SWE-bench.** https://www.swebench.com/
7. **SWE-Bench Pro, Scale AI.** https://scale.com/research/swe_bench_pro · arXiv:2509.16941
8. **DeepSWE.** arXiv:2607.07946
9. **CursorBench, Cursor.** https://cursor.com/blog/cursorbench
10. **MCP-Atlas.** arXiv:2602.00933
11. **Toolathlon.** arXiv:2510.25726
12. **τ2-bench, Sierra Research.** https://github.com/sierra-research/tau2-bench · arXiv:2506.07982
13. **BrowseComp, OpenAI.** https://openai.com/index/browsecomp/
14. **OpenAI MRCR.** https://huggingface.co/datasets/openai/mrcr
15. **GDPval, OpenAI.** https://openai.com/index/gdpval/
16. **Humanity's Last Exam.** https://lastexam.ai/ · arXiv:2501.14249
17. **GPQA.** arXiv:2311.12022 · Epoch AI: https://epochai.substack.com/p/gpqa-diamond-whats-left
18. **ARC-AGI-2 and ARC-AGI-3, ARC Prize.** https://arcprize.org/arc-agi/2 · https://arcprize.org/blog/arc-agi-3-launch
19. **MMMU-Pro.** arXiv:2409.02813
20. **Launch pages compared (page 8).** openai.com/index/gpt-5-6 · openai.com/index/introducing-gpt-5-5 · deepmind.google/models/model-cards/gemini-3-1-pro · blog.google (Gemini 3.8 Flash) · x.ai/news/grok-4-7 · huggingface.co/deepseek-ai/DeepSeek-V4-Pro

Luis V