



AI Model Comparisons: 5 Myths, From the Benchmark OpenAI Retired to the \$0.08 Model That Scores 21

On 23 February 2026, OpenAI explained why it had stopped reporting scores on the coding benchmark everyone quoted. Five months later it retracted the replacement it had recommended in that same post.

OpenAI audited 138 of the 500 SWE-bench Verified tasks and found test-design or problem-description defects in 59.4% of them, then wrote on 23 February 2026 that it had stopped reporting the score and asked other developers to do the same. On 8 July 2026 it put 30% of SWE-bench Pro tasks in the broken column and withdrew its interim recommendation. Epoch AI marks both “Flawed.” If your model selection document cites either number, it is citing a metric whose own publisher has disowned it.



Two benchmarks, two retractions, five months apart

The sequence matters more than any single number, so let me lay it out plainly.

In February 2026, OpenAI took the 138 SWE-bench Verified tasks that o3 failed consistently across 64 runs and had people read them. The result, from [Why we no longer evaluate SWE-bench Verified](#): material test-design or problem-description issues in 59.4% of that audited set. Two causes were named. First, tests that reject functionally correct solutions, which means a model can solve the problem and still score zero. Second, contamination: frontier models had seen at least some SWE-bench problems and their solutions during training.

The operative sentence is blunt. “This is why we have stopped reporting SWE-bench Verified scores, and we recommend that other model developers do so too.”

That same post pointed the research community at the public split of SWE-bench Pro as an interim measure while OpenAI built what it called “new, uncontaminated evaluations.” No named replacement shipped.

Then on 8 July 2026, in [Separating signal from noise in coding evaluations](#): “We find 30% of SWE-Bench Pro tasks to be broken, and are retracting our previous recommendation that the research community use it as a leading coding eval.”

Both retractions came from the lab with the most to gain from high coding scores. [Epoch AI](#) designates both benchmarks “Flawed.”

The 59.4% number is smaller and larger than it looks

This is the part most summaries get wrong in both directions, so be careful with it.

59.4% does not mean 59.4% of SWE-bench Verified is broken. It means 59.4% of a deliberately adversarial sample is broken: the 138 tasks a strong model kept failing, which is exactly where you would expect defects to concentrate. Epoch AI does the arithmetic the other way. That audited set was 27.6% of the 500 tasks, which implies a floor of roughly 16.4% of all tasks with scoring problems. One in six, minimum, with no upper bound



established.

The contamination finding is the heavier one. Epoch AI's verdict: "improvements on SWE-bench Verified no longer reflect meaningful improvements in models' real-world software development abilities. Instead, they increasingly reflect how much the model was exposed to the benchmark at training time."

A broken test is noise. Contamination is a bias with a direction: it inflates the models that trained on the data most, which is systematically the newest and best-funded ones. If you used SWE-bench deltas to justify an upgrade in 2026, some unknown fraction of that delta was memory.

The defect rate you can defend is **16.4% across all 500 tasks, not 59.4%**. Anyone quoting 59.4% as the benchmark-wide figure is overstating it, and anyone waving the finding away because the sample was adversarial is ignoring a floor of one broken task in six.

Five claims that survive in model comparison decks

Each of these is stated the way I actually see it written, then corrected against the data.

THE CLAIM SWE-bench Verified is the industry standard for coding ability, so comparing models on it is the neutral choice.	THE REALITY OpenAI stopped reporting it on 23 February 2026 and asked other developers to do the same, citing tests that reject correct solutions plus training contamination. Epoch AI rates it "Flawed."
THE CLAIM Fine, use SWE-bench Pro instead. It was the recommended successor.	THE REALITY OpenAI retracted that recommendation on 8 July 2026 after putting 30% of its tasks in the broken column. Epoch AI rates SWE-bench Pro "Flawed" as well.



AI Model Comparisons: 5 Myths, From the Benchmark OpenAI Retired to the \$0.08 Model That Scores 21

THE CLAIM

The cheap model is the cheap model.
\$0.08 per 1M tokens against \$0.58 is a 7x saving.

THE REALITY

Artificial Analysis computes that displayed price with a 7:2:1 cache-hit / input / output token ratio. It is a weighted construct that assumes your workload is 70% cache hits. The quality gap behind it: GPT-6 Luna (High) scores 21 on the Intelligence Index, Gemini 3.8 Flash (High) scores 31.

THE CLAIM

Tokens per second tells you which model feels faster.

THE REALITY

On the 28 October 2026 Artificial Analysis data, Gemini 3.8 Flash (High) generates at 132.6 tokens/s against Luna's 122.0, an 8.7% gap. Time to first token runs the other way and runs wide: Luna 13.72s, Gemini 27.26s. The slower generator starts answering roughly twice as fast.

THE CLAIM

Leaderboard rank is crowd-sourced and hard to game.

THE REALITY

A September 2026 critique ([arXiv:2609.23201](https://arxiv.org/abs/2609.23201)) documents Meta publicising a high LMArena position for an experimental Llama 4 Maverick chat variant while the separately evaluated public version ranked substantially lower. The rank was real. It belonged to a different artefact than the one you could download.

Why the \$0.08 model may still cost more per finished task

Price per token and price per finished task are different quantities, and only one of them appears on your invoice.



AI Model Comparisons: 5 Myths, From the Benchmark OpenAI Retired to the \$0.08 Model That Scores 21

ATTRIBUTE	GPT-6 LUNA (HIGH)	GEMINI 3.8 FLASH (HIGH)
Blended price per 1M tokens	\$0.08	\$0.58
Intelligence Index	21	31
Output speed	122.0 tokens/s	132.6 tokens/s
Time to first token	13.72s	27.26s

A 7x price gap for a 10-point quality gap looks decisive until you ask how many attempts each model needs.

I want to be careful here, because the research data gives me prices and index scores, not retry rates. So this is my reasoning rather than a sourced finding: a weaker model that needs three passes, a validation step and occasional human repair can erase a 7x unit-price advantage without anyone noticing, because the retries land in a different budget line than the API bill. The index gap of 21 against 31 is large enough that I would not assume parity on first-pass success.

Measure cost per accepted output on your own task distribution instead of cost per million tokens. Replace tokens per second with the latency figure your users actually perceive, and replace benchmark rank with a frozen set of 50 to 200 of your own historical tasks with human-graded answers. Your task set is uncontaminated by construction: no lab trained on your Jira tickets.

Throughput and responsiveness rank in opposite directions here

That is a product decision wearing the costume of a metric.

For an interactive assistant, time to first token is close to the whole experience. Luna's 13.72s against Gemini's 27.26s is a difference a user feels on every request. For a batch pipeline that writes to a queue overnight, time to first token is irrelevant and the 8.7% throughput edge is the only one that pays. Pick the wrong column and you will optimise a number nobody experiences.



The gpt-oss-120b figures make the same point from a different angle. Artificial Analysis lists it at 236 tokens/s on OpenAI's creator page and 264 tokens/s in a low-reasoning comparison. Same weights, same measurer, two numbers, because the reasoning setting changed. Any speed figure you quote without naming the configuration is a number without units. I have made a related argument about scaffolding before: [the same harness can produce 65% or 74% on SWE-bench](#) depending on how it is wired.

A benchmark score without a harness description, and a speed figure without a reasoning setting, are the same category of claim: a number detached from the conditions that produced it.

What is actually defensible after February and July 2026

Strip out the retracted and the ambiguous, and a short list remains.

Confirmed: SWE-bench Verified has a defect floor of roughly 16.4% and documented contamination, per OpenAI and Epoch AI. SWE-bench Pro has roughly 30% broken tasks, per OpenAI. Both are rated "Flawed" by Epoch AI. The Artificial Analysis blended price is a 7:2:1 weighted construct, not a tariff. Throughput and time to first token diverge sharply between Luna and Gemini 3.8 Flash on the 28 October 2026 data. The Llama 4 Maverick leaderboard case is documented in arXiv:2609.23201.

Not confirmed, and I will not pretend otherwise: I have no named, shipped, uncontaminated public coding benchmark to point you at. OpenAI said in February it was building "new, uncontaminated evaluations." As of the July post, what exists publicly is a retraction. That is an uncomfortable place for the whole field to be sitting, and it is where it is sitting.

MY TAKE

These myths persist because the alternative is expensive. A public benchmark score is free, comparable across vendors, and arrives pre-formatted for a slide. Building a private evaluation set costs a senior engineer a week and produces a number only you can interpret, which makes it useless for justifying a decision to a board and extremely useful for actually making one. I think the contamination problem is structural rather than fixable: any benchmark valuable enough to be widely cited becomes valuable enough to train on, so its



useful life starts the day it is published and ends quietly some time later. I expect the February and July 2026 pattern to repeat with whatever succeeds SWE-bench. Plan for your evaluation harness to outlive three public benchmarks.

Three edits to the selection document this week

In order of how much they change the answer.

Delete the benchmark column, or annotate it with the retraction date and the Epoch AI “Flawed” rating so the reader knows what weight to give it. **A cited number the publisher has withdrawn is worse than no number**, because it carries false precision.

Split the single price column into price per million tokens and estimated cost per accepted task, and write down the cache-hit assumption you are inheriting. If your workload is not 70% cache hits, the blended figure flatters one model over another by an amount you have not calculated.

Split speed into generation rate and time to first token, and mark which one your product is sensitive to. If you cannot answer that question, you have found the more important gap. I have worked through the broader ranking of evaluation methods by what they predict in production [separately](#).

None of this requires new tooling. It requires accepting that the comparison you can produce in an afternoon from public leaderboards is measuring something adjacent to what you care about, and that the one worth having takes a week and belongs to you.

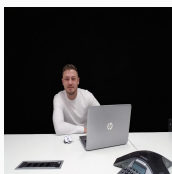
WHAT TO DO WITH THIS

Both headline coding benchmarks were disowned by their most prominent publisher inside five months, and the cost and speed columns most teams use are weighted constructs that rank differently depending on which one you read. If your model decision has real money behind it, the evaluation has to be yours.

[Book a call →](#)



AI Model Comparisons: 5 Myths, From the Benchmark OpenAI Retired to the \$0.08 Model That Scores 21



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →