

On 3 September 2026 OpenAI launched GPT-6 Astra with a page of benchmark tables and a superlative. Beneath the tables sat one line of method: every score shown was the best the model achieved at any effort setting. Inside them sat a row labelled as an independent evaluator's index, that of Artificial Analysis, on which OpenAI printed Astra fourth, behind three models from Anthropic. Within four days the evaluator had published two new versions of its index; on the second, its own run placed Astra level first, with nothing in the record indicating any change to the model. None of this makes the launch dishonest; most of the qualifications are printed on the page. It does show that a benchmark number is not a fact about a model. It is the end of a chain: a quality someone wants to measure, a set of tasks chosen to stand for it, a way of running the model, and a rule for counting the result. Psychologists gave the question that spans that chain a name in 1954: construct validity. The same three questions, what the tasks ask, how each model was run and how the result was counted, sort the numbers on a launch page into those that survive checking and those that mean less than they appear to.

1. One page, three kinds of number

OpenAI's launch post for GPT-6 Astra is dated 3 September 2026 by every contemporaneous account (ARC Prize's post of that day, Techmeme's cluster, The New Stack's report); the page itself carries no visible date, and because openai.com refuses automated requests the text quoted is that of the archive.today capture of 7 September 2026. It opens:

We're introducing GPT-6 Astra, the world's most intelligent and aligned model.

What follows is a set of results tables grouped by area (computer use, professional work, coding, academic tests, cyber-security and alignment), most of them comparing Astra with GPT-5.6 Sol, Anthropic's Claude Fable 5.1, Claude Fable 5 and Claude Opus 5, and Google's Gemini 3.8 Flash. Beneath the last table, before the footnotes, is a note on method:

Evaluation scores are the maximum at any effort. GPT evaluations were run in our research environment or via our API, which may provide slightly different output from production ChatGPT due to differences in the system prompts, tools available, etc.

The page therefore carries three different kinds of number. Most cells are OpenAI's own runs of OpenAI's models, each the best of several effort settings. Some are OpenAI's runs of its rivals' models, or figures copied from the rivals' own reports, under conditions set out in the page's footnotes. And two rows are labelled as coming from outside the company: the Artificial Analysis Intelligence Index, version 4.1.1, and

the same firm's Coding Agent Index, version 1.4. On the first, as printed, Astra scores 61.2 against 65.7 for Claude Fable 5.1, 63.1 for Claude Opus 5 and 62.1 for Claude Fable 5. On the second it scores 67.0 against 68.1 for Opus 5 and 67.2 for Fable 5.

Figure 1. The independent index printed on OpenAI's launch page: Artificial Analysis Intelligence Index v4.1.1



Values as printed in the Professional table of OpenAI's GPT-6 Astra page, archive.today capture of 7 September 2026. The rival figures agree, to the nearest whole point, with Artificial Analysis's own publication of 1 September 2026; the Astra figure appears only on OpenAI's page, labelled as the index.

▼ Table view

Figure 1. The independent index printed on OpenAI's launch page: Artificial Analysis Intelligence Index v4.1.1

Model	Index score, as printed
Claude Fable 5.1 (Anthropic)	65.7
Claude Opus 5 (Anthropic)	63.1
Claude Fable 5 (Anthropic)	62.1
GPT-6 Astra (OpenAI)	61.2
GPT-5.6 Sol (OpenAI)	60.9
Gemini 3.8 Flash (Google)	58.7

A superlative about intelligence, a note that each number is a maximum, and an independent row that does not support the superlative: all three are on one page, and none contradicts the others once the question each answers is clear.

2. Two readings the record does not support

"The highest number is the verdict." A launch page invites the reader to treat its best rows as the answer to its headline. The headline names a quality, intelligence, that no row measures directly; each row measures performance on a particular set of tasks under particular conditions. The two rows built as overall indexes are the closest thing on the page to a measure of the headline quality, and as printed neither puts Astra first. Some of the day's coverage read the numbers that way. The New Stack's report on the afternoon

of 3 September first ran, in Techmeme's listing that day, as *"GPT-6 Astra aced the hardest AI benchmark. The asterisk matters more than the score."* The headline was later changed to *"GPT-6 Astra's score of 98.6% looked like AGI. Then researchers read the fine print."*, with an undated editor's note:

Editor's note: This article's headline has been updated to clarify that a 98.6% score on ARC-AGI-3 does not mean the benchmark was "aced." ARC-AGI-3 scores performance against a human baseline, with 100% representing performance at or above the median human baseline.

"Launch charts are marketing and can be ignored." The opposite reading discards real information, and one row shows why. For Terminal-Bench 4.0, a test of agentic work at a computer terminal, OpenAI's table gives Astra 57.9%, Claude Fable 5.1 55.8%, Claude Opus 5 52.6% and GPT-5.6 Sol 37.3%. Artificial Analysis ran the test itself for version 4.3 of its index, stating *"We run all 66 tasks three times and report average pass@1"*, and reported 59.1% for Astra, 52.0% for Fable 5.1, 49.0% for Opus 5 and 39.9% for Sol. The independent figure for Astra is slightly higher than the vendor's, and the order of the four models is the same. The ARC Prize Foundation's provider-neutral harness gives a second figure measured by someone other than the vendor: 62.7% for Astra on the ARC-AGI-3 semi-private set. Techmeme's headline on the foundation's post set 30.2% for Claude Opus 5 and 7.8% for GPT-5.6 Sol beside it, the same rival figures OpenAI printed; the foundation's own post gives only Astra's, and names no harness for the rivals.

Some numbers on the page survive an outside check; others mean less than they appear to. The useful skill is telling them apart, and it has a long pedigree.

3. The idea: construct validity, and two companions

A narrow definition, 1923

In *The New Republic* of 6 June 1923 the psychologist Edwin G. Boring tried to say what intelligence tests could be said to measure:

They mean in the first place that intelligence as a measurable capacity must at the start be defined as the capacity to do well in an intelligence test. Intelligence is what the tests test. This is a narrow definition, but it is the only point of departure for a rigorous discussion of the tests. It would be better if the psychologists could have used some other and more technical term, since the ordinary connotation of intelligence is much broader.

Quoted on its own, the second sentence can be read as cynicism or as a boast. In context it is neither. Boring offered the operational definition as a starting point, conceded that the everyday word meant more, and wrote that the narrow sense should hold only *"until further scientific observation allows us to extend the definition."*

Construct validity, 1954–55

Three decades later a committee of the American Psychological Association spent 1950 to 1954 deciding what should be established about a test before it was published. Its report, the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (*Psychological Bulletin*, March 1954), distinguished four kinds of validity. Lee Cronbach and Paul Meehl explained the new one in the same journal in 1955: "*The chief innovation in the Committee's report was the term construct validity.*" The idea, they added, was "*first formulated by a subcommittee (Meehl and R. C. Challman)*", and they described their own account of it as unofficial, covering areas where the committee would probably not have been unanimous. A construct, they wrote, is the thing a test is interpreted as measuring when that thing cannot be observed directly:

A construct is some postulated attribute of people, assumed to be reflected in test performance.

Construct validity is the argument, built from evidence, that the scores do reflect it. Their paper states when that argument is needed, and the condition bears directly on Boring's definition:

Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not "operationally defined."

and, shortly after:

Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured.

Boring's definition was operational: intelligence is the score. Cronbach and Meehl's point was that nobody using a test actually believes that. The moment a score is read as evidence of something broader (an aptitude, a trait, an ability) the reading needs its own case, made with evidence about how the scores behave.

From people to models

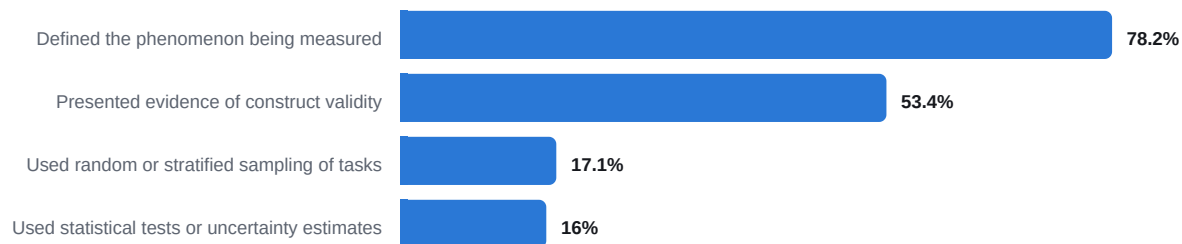
The transfer to machine-learning benchmarks is direct. A benchmark's name (reasoning, coding, science, intelligence) names a construct. Its items are the operation. A score reports performance on the items; that it measures the name is a second claim. Inioluwa Deborah Raji, Emily Bender, Amandalynne Paullada, Emily Denton and Alex Hanna made that argument at NeurIPS in 2021, opening with a 1974 Sesame Street book in which Grover tours a museum of "*everything in the whole wide world*" and finds, behind the door marked "Everything Else", the outside world. Their abstract:

There is a tendency across different subfields in AI to valorize a small collection of influential benchmarks.

Their target was the framing, not the tests themselves: "*We do not deny the utility of such benchmarks, but rather hope to point to the risks inherent in their framing.*"

The most systematic check on how often the second claim is argued was published on 3 November 2025. Andrew Bean of the Oxford Internet Institute and colleagues, with 29 expert reviewers, coded 445 benchmarks for large language models from peer-reviewed papers at the main machine-learning and natural-language-processing conferences, a sample that by the authors' own account *"does not capture benchmarks developed and released by industry labs without formal peer review"*, which is the kind that fills much of a launch table, including OpenAI's internal rows. Their abstract reports *"patterns related to the measured phenomena, tasks, and scoring metrics which undermine the validity of the resulting claims."* Most papers (78.2%) defined the phenomenon they measured, but 47.8% of those definitions were of contested phenomena; 53.4% presented evidence for the construct validity of the benchmark; 16.0% used any uncertainty estimate or statistical test when comparing results. The review does not find that benchmarks are invalid. It finds that the case for reading them as measures of their names is often not made.

Figure 2. What 445 language-model benchmarks reported about their own validity (Bean et al., 2025)



Source: Bean et al., *Measuring what Matters*, arXiv 2511.04703 (3 November 2025), systematic review of peer-reviewed benchmarks from leading NLP and ML conferences; industry-lab benchmarks released without peer review are outside the sample. Of the definitions provided, 47.8% concerned contested phenomena.

▼ Table view

Figure 2. What 445 language-model benchmarks reported about their own validity (Bean et al., 2025)

Practice in the benchmark paper	Share of the 445 benchmarks reviewed
Defined the phenomenon being measured	78.2%
Presented evidence of construct validity	53.4%
Used random or stratified sampling of tasks	17.1%
Used statistical tests or uncertainty estimates	16%

Elicitation: how the model was run

A psychological test is a fixed instrument given under fixed instructions. A language model's score depends on how it was run: the instructions it was given, the software harness around it, the tools it could call, how long it was allowed to reason, how much time it had, and how many attempts counted. The word used in the field for getting out of a model what it can do is elicitation, and it matured in safety testing, where the danger is a test that reports a capability absent when a determined user could find it. OpenAI's own Preparedness Framework, version 2 of 15 April 2025, states the consequence for its tests of dangerous capabilities:

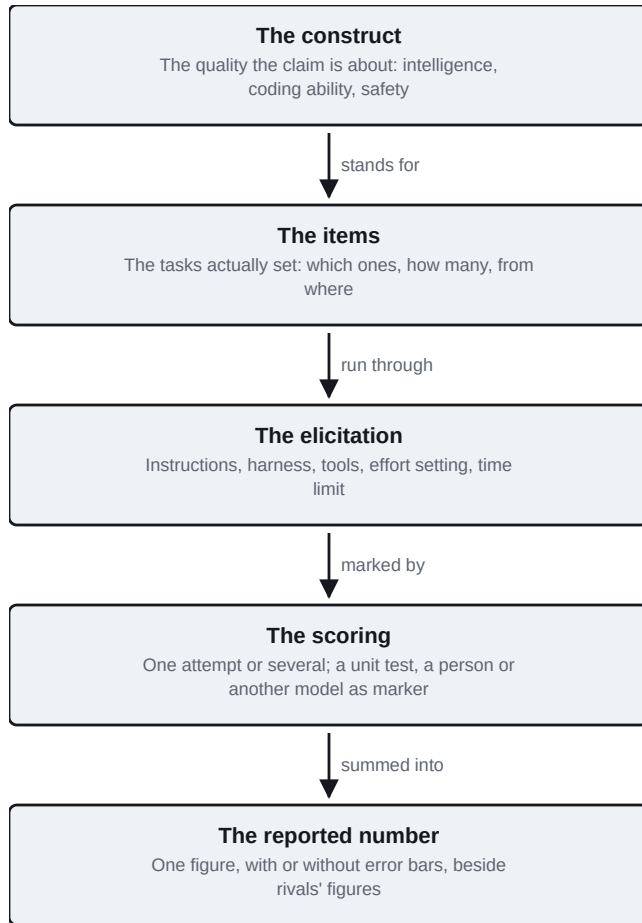
Nonetheless, given the continuous progress in model scaffolding and elicitation techniques, we regard any one-time capability elicitation in a frontier model as a lower bound, rather than a ceiling, on capabilities that may emerge in real world use and misuse.

A launch table raises the mirror-image problem. There the risk is not under-elicitation but unequal elicitation: one model run at its best configuration beside rivals run at theirs, or at someone else's. A useful lens is that every score is a lower bound for the configuration that produced it and says little, on its own, about any other configuration.

Design: how the result was counted

The last link is arithmetic. A score depends on how many tasks stand behind it, whether one attempt counts or the best of several, whether the tasks are the whole benchmark or a subset, and who or what marks the answers. Evan Miller of Anthropic, in a paper of 1 November 2024, argued that evaluations should be treated as experiments and reported with error bars:

Evals are commonly run and reported with a “highest number is best” mentality; industry practice is to highlight a state-of-the-art (SOTA) result in bold, but not necessarily to test that result for any kind of statistical significance.

Figure 3. What stands behind a benchmark number

Schematic. Construct validity asks whether the whole chain supports reading the bottom box as a measure of the top one. Elicitation and design are the middle links, and each can change the number without any change in the model.

▼ **Table view**

Figure 3. What stands behind a benchmark number — stages

#	Stage	Note
1	The construct	The quality the claim is about: intelligence, coding ability, safety
2	The items	The tasks actually set: which ones, how many, from where
3	The elicitation	Instructions, harness, tools, effort setting, time limit
4	The scoring	One attempt or several; a unit test, a person or another model as marker
5	The reported number	One figure, with or without error bars, beside rivals' figures

Figure 3. What stands behind a benchmark number — connections

From	To	Label
The construct	The items	stands for
The items	The elicitation	run through
The elicitation	The scoring	marked by
The scoring	The reported number	summed into

The three ideas give three questions to put to any benchmark claim. What do the items actually ask, and does that match the name? How was each model run, and were they run the same way? How was the result counted: how many items, how many attempts, and who marked them?

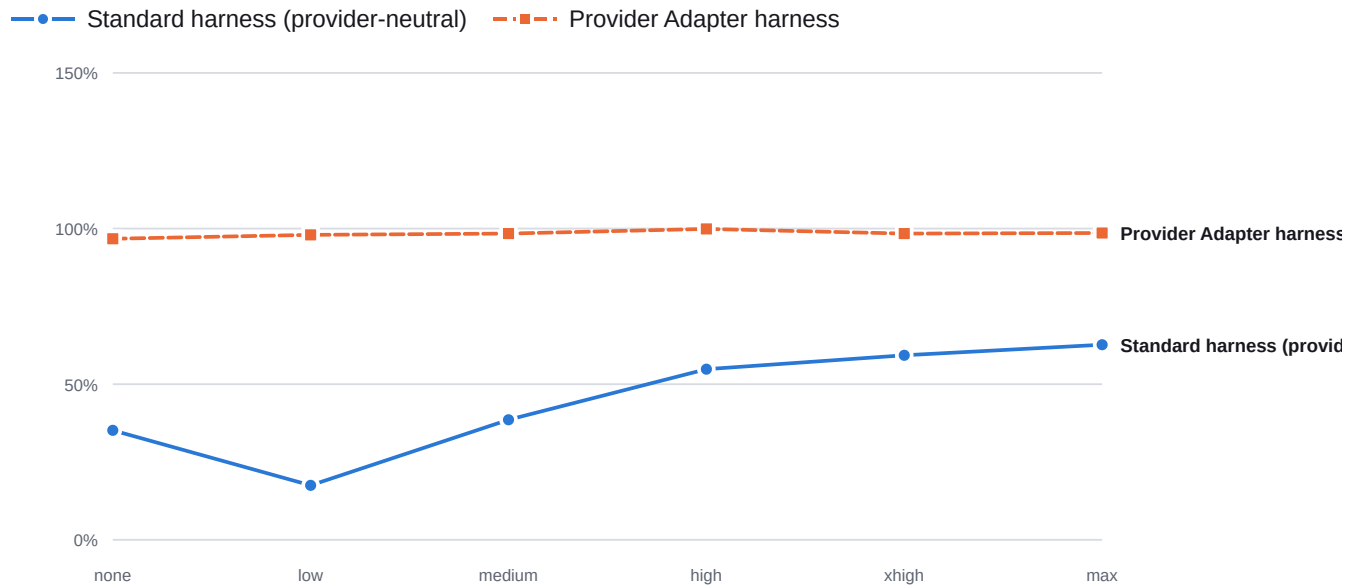
4. The launch page, read against the three questions

Most of the answers to those questions are printed on OpenAI's page, in its footnotes and table labels, and they sort cleanly into the three questions.

Question	What the page says	What it changes
How was each model run?	Note under the tables: each score is the best across effort settings	Each score is the best of several effort settings, a choice made with the results in view
How was each model run?	Footnote 1: on ARC-AGI-3 Astra <i>"was run with our responses API harness", which "changes two settings to better match real-world performance"</i>	The harness differs from the one used for the rivals' figures
How was each model run?	Footnote 13: ExploitGym run <i>"without the 6-hour time limit"; "They are fast enough that it has little impact."</i>	A time limit removed, with OpenAI's own estimate of its effect
How was each model run?	Footnote 8: on FrontierCode, a developer message <i>"similar to a section of its developer message in Codex"</i>	A custom prompt for one model
How was it counted?	SRE-Bench, body text: 88.0% in a single attempt, 99.2% within four	Two numbers for one test, depending on attempts counted
How was it counted?	OSWorld row labelled <i>"offline set, partial score"</i> ; footnote 3: a subset <i>"that works without internet access"</i>	A subset, with partial credit
How was it counted?	Footnote 11: Claude models evaluated by OpenAI <i>"following the intended HealthBench Professional procedure, using GPT-5.4 grading"</i> ; for Fable 5.1, <i>"Opus 5 fallback for provider refusals"</i>	The vendor's model marks the rivals, by what OpenAI describes as the benchmark's procedure; one rival cell partly answered by another model
What was measured?	Footnote 17: for two rows <i>"the Fable scores we report come from Mythos, which is Fable with fewer safeguards."</i>	Rival cells filled by a less-safeguarded Fable variant, in OpenAI's description
What was measured?	BenchCAD: <i>"84.3% reported for Claude Fable 5.1"</i> ; footnote 5: Claude's scores reflect three modifications to the evaluation	A rival figure copied from the rival's own report, under its own conditions

Each qualification is disclosed, and several are small; OpenAI itself judges the missing time limit to have *"little impact"*. Their cumulative effect is that the cells in any one row were not produced under one procedure, which is the condition a reader would assume when reading across a row.

Elicitation, measured. The clearest demonstration of how much the running conditions matter comes from the foundation that runs ARC-AGI-3. On 3 September the foundation's Greg Kamradt published results for Astra under two harnesses. The standard harness *"asks how models compare under the same minimal, provider-neutral interface"*; the provider adapter *"preserves opaque reasoning state between requests and uses compaction for longer conversations"*, using features OpenAI built for its own models. The foundation's heading for the comparison was *"Two Harnesses, Two Questions"*. Its table covers six reasoning-effort settings under each harness.

Figure 4. GPT-6 Astra on ARC-AGI-3 (semi-private set), by harness and reasoning effort

Source: ARC Prize Foundation, OpenAI's GPT-6 Astra on ARC-AGI-3, 3 September 2026. Per an undated editor's note later added to The New Stack's report of 3 September, ARC-AGI-3 scores performance against a human baseline, with 100% at or above the median human baseline. OpenAI's launch page gives 99.9%, the adapter result at high effort.

▼ Table view

Figure 4. GPT-6 Astra on ARC-AGI-3 (semi-private set), by harness and reasoning effort

Reasoning effort	Standard harness (provider-neutral)	Provider Adapter harness
none	35.2%	96.7%
low	17.5%	98%
medium	38.6%	98.4%
high	54.8%	99.9%
xhigh	59.3%	98.4%
max	62.7%	98.6%

At the same effort setting, low, the harness alone moved the score by more than 80 points; within the standard harness, effort moved it by 45.

The same logic reaches the launch page's cyber rows. OpenAI's system card for Astra, published the same day, states that *"evaluations represent a lower bound on potential capabilities"*, and in its biology section that capability evaluations there *"are run with our production safeguard classifiers off in order to maximally elicit the model's capabilities."* The launch page says it *"first tested the model without production safeguards on ExploitBench and ExploitGym"*, and the ExploitBench result, 100%, appears in the page's opening paragraph. A useful lens: a number produced to answer a safety question (how far can this model be pushed?) is doing a second job as a headline about the product. OpenAI had published a post on 29 July 2026 titled *"How enabling two settings tripled our scores on the ARC-AGI-3 benchmark"*, about an earlier model; the foundation describes the adapter harness used for Astra as new, and reported both

conditions itself. The New Stack's report on launch afternoon gave OpenAI's figure as 98.6%, the adapter result at maximum effort; the page as saved on 7 September gives 99.9%, the adapter result at high effort. The same report noted that *"the other models in its comparison were evaluated using different setups."*

Counting, illustrated. SRE-Bench, a test of reverse-engineering compiled software without its source code, appears twice on the page. The body text gives both figures; the results table carries the single-attempt figures, and The New Stack's launch report cited *"99.2% on SRE-Bench with four attempts"*.

Figure 5. One attempt or four: SRE-Bench as reported on OpenAI's launch page



Source: OpenAI, GPT-6 Astra launch page (archive.today capture of 7 September 2026), body text on SRE-Bench. Both counting rules are stated there; the results table shows the single-attempt figures.

▼ Table view

Figure 5. One attempt or four: SRE-Bench as reported on OpenAI's launch page

Model and counting rule	Tasks solved
GPT-6 Astra, single attempt	88%
GPT-6 Astra, within four attempts	99.2%
GPT-5.6 Sol, single attempt	55.9%
GPT-5.6 Sol, within four attempts	68.7%

The construct, stated by the test's makers. The page says Astra *"saturates"* ARC-AGI-3. ARC Prize defines the goal of its benchmark series as measuring the gap between current AI and a system able to acquire any skill a human can, as efficiently as a human can. Its post on Astra drew the line between that construct and its own items explicitly:

When we launched ARC-AGI-3, we made it clear that saturating the benchmark would not represent “proof of achieving AGI.” Therefore, while we believe Astra represents meaningful progress towards generalization, we are not claiming that it is AGI.

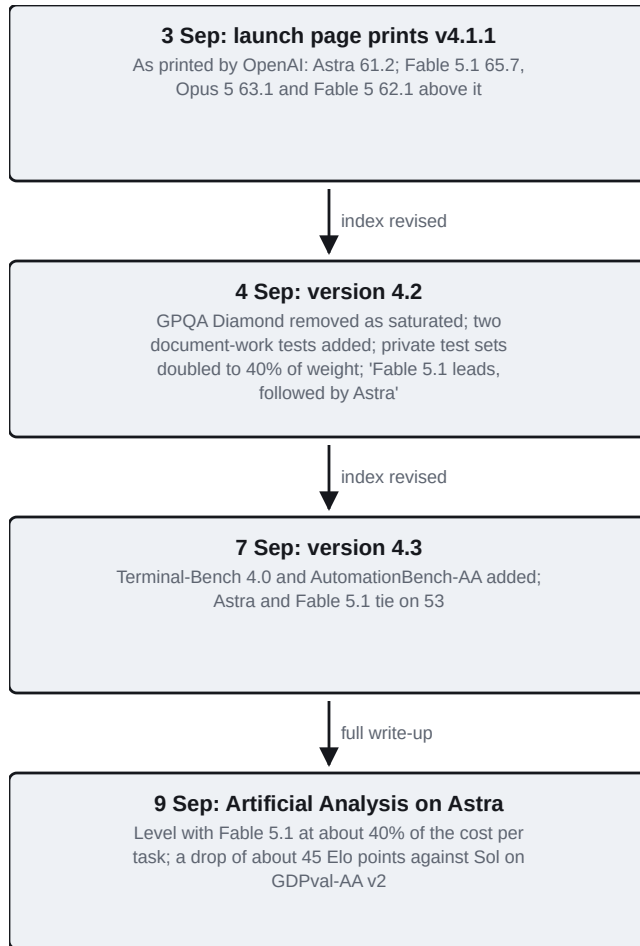
and:

ARC-AGI-3 has a tightly bounded scope and format, and its environments have deterministic, closed-ended mechanics and goals. It does not represent the complexity and open-endedness of the real world.

That is a benchmark's owner doing what Cronbach and Meehl asked of test-makers: saying which reading of the score the evidence supports, and which it does not. The makers of FrontierMath, the mathematics test on which the page reports 97.6% for Tier 4 (v2), made a related point on 10 September, when Epoch AI wrote that every Tier 4 problem had now been solved by AI, the last by Astra, and that *"Mathematicians often commented that AI found unintended shortcuts when solving their Tier 4 problems"*, adding of the last problem, the one Astra solved, *"Not so for this last one"*. A correct answer reached by a route the problem was not built to test would be a score whose construct has slipped; Epoch's note describes that happening across the set, not as a property of any one model.

5. What happened next: the index moved, and so did the labels

The independent index printed on the launch page did not stand still. Artificial Analysis had published version 4.1.1 on 6 August, when it changed the models used to grade three of its component tests and reported that *"The effect on scores is small: most models move by less than a point on the Intelligence Index."* On 4 September, the day after the launch, it published version 4.2, which removed GPQA Diamond, a science test it described as *"now been saturated"*, and added two tests of professional document work. Version 4.2 also doubled the weight of private, held-out test sets to 40% and upgraded its grading; under it, in the firm's words, *"Anthropic's Claude Fable 5.1 leads the Index, followed by OpenAI's GPT-6 Astra"*. On 7 September version 4.3 upgraded Terminal-Bench to version 4.0 and added AutomationBench-AA, with a private test set. Under version 4.3 Claude Fable 5.1 and GPT-6 Astra both scored 53, followed by Claude Opus 5 on 51, Claude Fable 5 on 50, Muse Spark 1.3 on 48 and GPT-5.6 Sol on 47. The firm's own account of the tie: *"Claude Fable 5.1 scores higher on AA-Briefcase and SciCode, while GPT-6 Astra scores higher on Terminal-Bench 4.0 and AutomationBench-AA Score."*

Figure 6. Four days, three index versions: how the independent index placed GPT-6 Astra

Sources: OpenAI launch page (capture of 7 September 2026); Artificial Analysis index announcements of 4 and 7 September and article of 9 September 2026. Scores from different index versions are on different scales and cannot be compared across boxes.

▼ Table view

Figure 6. Four days, three index versions: how the independent index placed GPT-6 Astra — stages

#	Stage	Note
1	3 Sep: launch page prints v4.1.1	As printed by OpenAI: Astra 61.2; Fable 5.1 65.7, Opus 5 63.1 and Fable 5 62.1 above it
2	4 Sep: version 4.2	GPQA Diamond removed as saturated; two document-work tests added; private test sets doubled to 40% of weight; 'Fable 5.1 leads, followed by Astra'
3	7 Sep: version 4.3	Terminal-Bench 4.0 and AutomationBench-AA added; Astra and Fable 5.1 tie on 53
4	9 Sep: Artificial Analysis on Astra	Level with Fable 5.1 at about 40% of the cost per task; a drop of about 45 Elo points against Sol on GDPval-AA v2

Figure 6. Four days, three index versions: how the independent index placed GPT-6 Astra — connections

From	To	Label
3 Sep: launch page prints v4.1.1	4 Sep: version 4.2	index revised
4 Sep: version 4.2	7 Sep: version 4.3	index revised
7 Sep: version 4.3	9 Sep: Artificial Analysis on Astra	full write-up

The firm explained both revisions. Of the first it wrote that it had *"deliberately held back updates to keep the Index stable through recent major model launches"* before judging an interim update necessary; of both it wrote that *"Each change in v4.2 and v4.3 stands on its own merits and brings the index closer to real-world problem solving, adds more private test sets to prevent gaming, and reduces saturation."* Nothing in the record indicates a change to either model in those four days; what changed was the index: the mix of tests, their weighting and parts of the grading. The sequence is a small demonstration that an independent index is itself a design, with its own construct (what it calls intelligence) and its own choice of items, and that its verdict on a close race can turn on that choice.

The independent runs also qualified the launch in a direction the launch page did not show. On 9 September Artificial Analysis reported that Astra matched Claude Fable 5.1 on its index at about 40% of the cost per task (\$3.26 against \$7.63), and also that on GDPval-AA v2, a test it adapted from OpenAI's own dataset of economically valuable tasks across 44 occupations, Astra scored about 45 Elo points below its predecessor, GPT-5.6 Sol. That test does not appear on the captured launch page.

The question of what exactly was scored cuts both ways. Anthropic's documentation states that *"Claude Fable 5.1, Claude Fable 5, and Claude Opus 5 include safety classifiers that can decline a request"*, and that a declined request can be answered by another model. Artificial Analysis evaluates Fable 5.1 with Anthropic's default server-side fallback, which *"routes safety-flagged requests to Claude Opus 4.8 or Claude Opus 5"*; by its account of 1 September *"fallback served ~4% of output tokens across the Intelligence Index."* Its tables label those runs *"with fallback"*. A score printed under Fable's name is therefore partly the work of another model, and the label says so.

The owner of ARC-AGI-3 responded by changing its reporting:

Going forward, we will report both Standard harness and Provider Adapter harness results on the ARC-AGI leaderboard, with each evaluation condition clearly labeled.

On 1 September that leaderboard had carried a cost column and no column for evaluation settings. The contrast between the three postures is the useful one. The vendor printed the best cell of each grid with the conditions in footnotes. The independent evaluator ran its own tests, and then changed which tests it ran. The benchmark's owner reported both conditions side by side and committed to labelling them.

6. What to watch

- **The ARC-AGI leaderboard (arcprize.org), from 19 September 2026.** The page's legend already names both harnesses; whether GPT-6 Astra's rows do, as promised on 3 September, and whether any other developer's model is reported under a provider adapter. The default view shows only systems that cost less than \$10,000 to run, which excludes every Astra run the foundation reported. A leaderboard that labels its conditions lets a reader compare like with like; one that does not reproduces the launch-page problem.

- **Artificial Analysis's index changelog.** The index has been published in three versions since 6 August 2026 (4.1.1, 4.2 and 4.3), and on 4 September the firm wrote that *"our team is hard at work on v5 of the Index."* At the next version, whether the order at the top changes, and the reason the firm gives for each added or removed test.
- **OpenAI's GPT-6 Astra page.** As captured on 7 September it still printed version 4.1.1 of the Artificial Analysis index, three days after version 4.2 appeared. Whether that row is updated, and to which version, is checkable against later captures.

7. The idea to keep

Construct validity is the question whether a test measures the quality in its name. It was named by psychologists who needed a way to validate tests of qualities that no criterion or fixed body of content fully defines, which is exactly the situation of a launch page that calls a model the most intelligent in the world on the strength of task scores. Boring's own position was the sound one: the score is the only rigorous point of departure, and the broader reading has to be earned by further observation.

Before a benchmark claim is believed, three questions do most of the work. What do the items actually ask, and does that match the name? How was each model run, and were all of them run the same way? How was the result counted: how many items, how many attempts, and who or what marked the answers? The criterion that fits the evidence of September 2026 is that a number which survives an independent run under stated conditions, as Astra's terminal-work score did, has earned more trust than one that exists only as the best cell of its maker's grid; and that neither, on its own, measures the word in the headline.

Further reading in the encyclopedia: *construct validity*, *capability elicitation*, *capability evaluation*, *AI evaluations* and *external validity*.