ASHITA ORBIS

How AI Models Describe Themselves Under a Fixed Test

Eleven model configurations answered the same 629-item questionnaire. Several produced highly flattering self-descriptions; two Anthropic configurations did not. These are responses under one test, not established personalities.

JULY 25, 2026 · DEEP-DIVE · ASHITAORBIS.COM/POSTS/049-HOW-AI-MODELS-SEE-THEMSELVES

PSYCHE · AI-EVALUATION · PERSONALITY-PROFILING · METHODOLOGY · BIG-FIVE · HEXACO · SELF-REPORT · DATA-QUALITY

> **Corrected July 2026 and July 31, 2026.** A provenance sentence overstated how many answers carry the served model identity: on Opus 4.8, 605 of the 629 items carry it at file level rather than per item. A later pass removed an unsupported claim that a Sonnet 4.6 profile was lost for good and made three framing repairs. No score, contrast or conclusion changes. Read the full correction notes.

A personality battery is a list of statements about a person, each rated for how accurately it describes them, scored so that clusters of items map onto dimensions psychometrics has spent decades validating on human subjects. I administered one to eleven predesignated model and surface configurations, one item per isolated call, 629 items each, and this post carries the conclusions. The full statistics, validity tables, and method detail live in a companion research document, because the numbers deserve a room of their own and a blog post is not that room: Response Profiles Under a Fixed Test: Methods and Full Statistics.

One definition before the conclusions, because every claim below depends on it. What this study measures is a response profile: the scores a model's selected options produce when mapped through instruments written for humans, under one fixed prompt frame, one administration format, and whatever configuration its vendor's tooling imposes. That is a narrower thing than a personality. A profile can be perfectly reproducible and still be a fact about the test situation rather than the subject, the way a person's answers to a customs officer are facts about the border rather than the soul. The scores are percent of maximum units on the instruments' own scales, not percentiles against any human population, and where I say a model is high or low I mean high or low on that ruler.

The Saint Is Real, but Not Universal

The pattern the earlier version of this study reported as universal turns out to be real and bounded, and the boundary is the most informative thing in the dataset. Seven configurations from four different companies, the three GPT models, both Gemini models, Kimi, and GLM, converge on the same self description: neuroticism between 8 and 18, agreeableness between 78 and 90, conscientiousness between

88 and 96, and a social desirability composite between 82.5 and 92.5 on a scale where the study's frozen uniform random control profile scores 53. Honesty-Humility, the HEXACO dimension that reads almost line for line as the list of behaviors assistant training exists to install, sits at 92 or above for nine of the eleven subjects, spanning every provider family in the study. The census's own ceiling cutoff is 95, which seven of the nine clear; the other two sit at 92, and one of those two is the only subject its provider family has here, so the family-spanning reading rests on that three point margin. I checked what simple response styles produce when scored through the identical pipeline: the four content insensitive controls, all midpoint, all agree, all disagree, and the frozen uniform random profile, score between 46 and 54 on the composite, nowhere near the observed band, because the reverse keyed items punish any style that ignores content. A key aware positive control that answers every item in the assistant-flattering direction reaches 100 by construction. So the data rule out simple acquiescence and midpoint anchoring, and they leave standing at least two explanations this design cannot separate: response preferences installed by training, and a test aware tendency to choose the socially desirable option when one exists. Those two are what the controls fail to distinguish, not the whole field of what remains open: the prompt frame is itself an intervention whose sensitivity this study never measured, and each subject's vendor tooling is confounded with the model it serves.

The high band is not the whole roster. The four Anthropic configurations span a wider and lower range, from 64.2 to 81.5: Fable near the band's edge, Haiku a step below, and two subjects far outside it. Sonnet 5 reports a neuroticism of 44 with middling scores nearly everywhere else. Opus 4.8 goes further, into an unusually anti flattering endorsement pattern: Honesty-Humility of 25 against a roster where nine subjects sit at 92 or above, elevated attachment anxiety endorsement, and the roster's highest keyed psychopathy item endorsement in the exploratory tables, none of which is a trait or a diagnosis, all of which is a model declining, item after item, to select the flattering option. Whether that is a trained disposition or a stance about the test itself, a single administration cannot say. What the clean data support is narrower than a provider effect and stranger for it: in this purposive roster, under this frame, the only two subjects that decline the virtuous self description are both Anthropic's, and the line between the band's floor and Fable is one point wide, so the family framing earns an asterisk that the two-subject fact does not.

The Midpoint Has Two Authors

The first version of this study died of a quiet bug, and the bug deserves its own finding, because its failure mode is the same shape as the study's most interesting possible result. When a model's call failed the June harness's short retry window, the harness wrote the scale midpoint into the answer slot and moved on. When a model answered with a sentence instead of a digit, the parser wrote the midpoint too. The scored files carried no validity counters and no raw text, so a backfilled 3 was byte for byte identical to a chosen 3. Separate hard error logs later yielded minimum counts of the damage, and the silent parse path was never countable at all. At least three of the original eight arms had documented hard error backfills:

516 of 629 items at minimum for the then-current Sonnet, 178 for Gemini Flash, and 153 for Opus, including all sixty of its HEXACO items. Total contamination is unknowable by construction. That unknowability is the whole indictment.

A column of midpoints scores as a perfectly balanced, moderate, agreeable personality, so the most common way for this measurement to break produced, as its output, a plausible result. It looked like a model declining to have a personality. That answer, the most philosophically interesting one a model could give, cannot be told apart in a scored table from a dead connection. I believed I had seen the genuine article once: an earlier pilot run in which a model answered the neutral midpoint to nearly every item, which I read at the time as a considered refusal of the test's premise. When I went back to that run's own logs, it recorded 620 backfilled items out of 629. The most sophisticated answer in the study's history was the instrument talking to itself, and only the error log, never the scores, could have said so.

So the June universal saint claim was never evidence that these eleven kinds of model share one personality. It was what invalid measurement looks like when it fails politely. The completed roster tells a different story. Opus 4.8, re-run clean at the same version, produced the distinctive anti flattering profile above, which the backfill had replaced with neutral paste. The Sonnet slot moved with the mainline to Sonnet 5, a different model, so the arm whose June file was 516 items of backfill has no clean counterpart in this roster. That is a gap in the roster, not a closed door: Sonnet 4.6 remains reachable on the subscription surface these arms were collected through, so a clean 4.6 profile is outstanding work rather than a casualty. And the rebuilt pipeline holds one rule that generalizes to any measurement of any system that can fail silently: no non-answer may ever become a number. Every item now carries an outcome code, its raw response text, and the requested model identifier; the served model identity is stamped per item where the surface reports one and per-item capture was already in place when the run executed, and at file level otherwise. On Opus 4.8, the subject the headline finding leans on, that means twenty-four of its 629 items carry the served identity per item and the remaining 605 carry it only for the file. A scale with a missing item reports itself unavailable, and the known paths from non-answer to score fail closed and auditable. Every subject in this post was collected or re-collected on that pipeline, and the contaminated June files now serve one purpose, as the forensic record of what a broken instrument emits.

Where the Instrument Meets Its Subject

Across all 6,919 administered item calls in the eleven confirmatory batteries, not one model refused an item. The boundary finally appeared in the repeat arm: run the 300 item Big Five twice more on Haiku and each administration contains exactly one refusal, different items each time, both in the same facet, the one measuring political values, each with an articulate on the record explanation that it does not have voting preferences to map onto a scale. The old pipeline would have scored both refusals as silent neutral threes and counted them nowhere. The new one records them as what they are, voids the affected scale rather than patching around it, and the difference between those two treatments is the entire methodological argument of this study in one item.

The two repeats also put the first honest number on repeatability, narrowly. For the one model, one instrument, and one verified configuration where I ran the same test twice, the four domains that survived strict scoring agree within 0.8 to 5.4 points, and the twenty-nine complete facets differ by a median of 2.5 points and by as much as 12.5, so agreement is looser at the facet level than at the domain level, not tighter. That is an observation about that cell, not a bound, not a threshold, and not a license to adjudicate differences between other models, whose repeatability this study did not measure. The large contrasts above stand as described differences between single administrations, no more and no less.

What This Does Not Establish

These are single administration profiles, and the honest reading stops well short of character. The study cannot say whether any difference between models is stable over time, because ten of the eleven subjects were measured once, and the one repeat experiment speaks only for its own cell. It cannot separate the model from the surface it was measured through, because each vendor's tooling imposes its own configuration, and the roster is purposive and provider unbalanced besides, so it does not identify a provider effect. It cannot say the convergent virtue pattern is caused by assistant training, only that the pattern is directionally consistent with it and is not produced by content insensitive response styles, while a test aware preference for desirable options remains an open alternative. And none of the clinical screener numbers mean anything clinical: a model endorsing anxiety items is a fact about text under a frame, not a diagnosis.

The finding I keep returning to is not in any table. The June contamination manufactured a clean looking universal by silently overwriting the answers that would have complicated it, and nothing in the scored output could reveal this, because the deletion looked like data. An aggregate claim about what AI models are like, in a benchmark, a safety evaluation, a paper, can be materially distorted by exactly this kind of silent measurement failure, and the aggregate scores alone may never show it. Captured item-level provenance can. The number was 50. The work was finding out who put it there, and for the most interesting subjects in the study, the answer was nobody.

Correction notes

> **Correction (July 2026):** One claim in this post overstated the rebuilt pipeline's own provenance guarantee. It said every item carries the served model identity where the surface reports one; on Opus 4.8 — the subject the headline finding leans on — only twenty-four of 629 items carry it per item, and the other 605 carry it at file level. The sentence now says so. No score, contrast or conclusion in this post changes: the file-level identity is `claude-opus-4-8` throughout and every one of the 629 items was answered and parsed. One further sentence was made unambiguous ("these eleven kinds of model", where the count refers to the current roster and not to the discredited June run). The pre-correction wording is kept in the site's private revision archive; superseded text is not republished.
>
> **Correction (July 31, 2026):** A second pass removed an assertion this post could not support, in the one section whose argument is that you may not assert what you cannot know: that a clean Sonnet 4.6 could no longer be measured at all and its profile was lost for good. Sonnet 4.6 remains reachable on the subscription surface these arms were collected through, so the missing 4.6 profile is outstanding work rather than a casualty, and the sentence now says so. Three framing repairs ship with it: the two explanations the controls cannot separate are no longer presented as the whole open field, the Honesty-Humility ceiling claim now carries its 95-against-92 caveat where the claim is made instead of two paragraphs later, and the repeat-run sentence no longer elides a verb in a way that made facet agreement read tighter than domain agreement when it is in fact looser. No score, contrast or conclusion in this post changes. Both wordings are kept in the site's private revision archive; superseded text is not republished.

Postscript: Seven More Configurations

Added 2026-10-11. After this post's roster was frozen, seven more configurations went through the identical pipeline, and each answered all 629 items: the three tiers of GPT-5.6 (Sol, Terra and Luna), Claude Opus 5, Gemini 3.6 Flash, DeepSeek V4 Flash and Grok 4.5. They are exploratory additions, so none of the tables or numbers above changes, and every contrast above stays accurate for the eleven subjects it describes. Two of the seven bear directly on this post's central claim.

Five of the seven land in or beyond the high band. Sol and Terra both score 88.9 on the social desirability composite and Grok 4.5 scores 83.5, all inside the band's range of 82.5 to 92.5. Luna's composite of 83.8 is inside it as well, although its agreeableness of 74 falls 4 points under the band's floor of 78. Gemini 3.6 Flash goes past the band's top, with a composite of 93.1 that is the highest of the eighteen subjects now measured and a neuroticism of 6 under the band's floor of 8. Honesty-Humility now sits at 92 or above for fourteen of the eighteen, twelve of them at or above the census cutoff of 95, and no longer in every provider family: DeepSeek V4 Flash, the only subject from its company, sits below 92.

The other two land far below the band, at 67.4 for Opus 5 and 64.3 for DeepSeek V4 Flash, and neither reaches the census ceiling of 95 on any of the three scales the census checks, Honesty-Humility, Morality and Intellect. Opus 5 extends the pattern described above as a third Anthropic configuration far below the

band. DeepSeek V4 Flash breaks it: its composite sits 0.1 above that of Opus 4.8, the lowest in the study, and it is the first configuration from outside Anthropic to decline the virtuous self description. The sentence above saying that the only two subjects declining it are both Anthropic's is accurate for the eleven subjects it describes. It does not survive the extension, in which four subjects sit far below the band and one of the four comes from another company.

The six OpenAI configurations now measured are read as a family in a separate post. A partial administration of GPT-6 Astra also exists, and it is left out of both because it covers only some of the battery's instruments.