Does a Language Model Have a Personality? What Four Studies Say, and What Measurement Should Do Next

Trevor Johnson
Idea Fields Institute
ideafields.institute
ORCID: 0009-0008-7962-0451
trevor.johnson@ideafields.institute

August 2026

Abstract

A growing literature administers human personality questionnaires to language models and reports Big Five profiles, while a parallel literature documents that those scores are unstable, dissociate from behavior, and fail measurement-invariance tests. Across four studies (roughly 500,000 scored responses from more than 60 open-weight and commercial model configurations spanning 1B to 756B parameters), we pursued one question through four manipulations: is there a personality *in there*, in the sense the word carries for humans, a coherent individual disposition that expresses itself in behavior and travels across situations? The answer that survived every test has two halves. First, Big Five structure in language models is a **population property**: convergent and discriminant validity emerge across models and strengthen with scale, while inside any single model, repeated administrations never cohere into a reliable trait profile (internal consistency near zero in every study, in hundreds of model-by-domain cells). Second, no inference-time manipulation we tried instantiates the missing individual: quantization does not (deployment precision leaves the incoherence untouched), deliberation does not (reasoning shifts what a model says about itself, systematically, toward calm introversion, without creating coherence), and deliberate installation does not (persona prompts shift self-report strongly and dose-dependently, while objectively graded behavior follows partially for one trait and not at all for another); training-time induction remains untested. What *is* reliably installable, in essentially every model at every scale, is a **context-robust self-presentation**: a way of describing oneself that holds even when the situation pushes against it. We argue this presentation layer, real, lawful, and installable, is what most "LLM personality" measurements measure, and that the field should stop borrowing human questionnaires as its primary instruments and build machine-native behavioral measures instead. The thesis has now survived its first test outside the construct it was built on: a fifth study administering four value instruments and objectively graded moral behavior to 16 configurations (81,577 scored responses) reproduces both halves in the moral domain, population-level structure without individual coherence, and an audience-conditioned presentation whose mechanism a preregistered follow-up decomposes into roughly 40 percent audience inference and 60 percent role compliance. We close with the agenda that follows: validated behavioral instruments with population-level psychometrics (item-response theory over model fleets), longitudinal drift tracking of deployed models, and the continuation of the values program as its own arc.

1 The question behind the questionnaires

Administering personality questionnaires to language models has become routine. Papers report Big Five profiles for frontier systems [21, 10, 11], persona-driven products ship with named characters, and "the model's personality" is discussed as though the phrase had settled meaning. A parallel literature has been steadily undermining the practice: scores shift under item reordering, paraphrase, persona assignment, reasoning mode, and conversation history [26]; self-reports dissociate from downstream behavior [9]; questionnaire-recovered profiles diverge from generation-probability profiles [24]; measurement invariance between human and model respondents fails [25]; and models exhibit human-like social-desirability response bias [19]. Systematic reviews of the young field catalog both literatures side by side [27].

What the perturbation catalogs establish is that the scores move. What they do not establish is what kind of thing is being measured when they sit still. A personality, in the sense the word carries for humans, is an individual-level disposition: it coheres across the items that measure it [2], expresses itself in behavior, and travels with its owner across situations. That definition is an idealization the human literature has itself contested since Mischel's classic critique [18]: human cross-situational behavioral consistency is modest, and human self-report-to-behavior correlations famously hover near $r \approx 0.2$ to 0.3. We use the idealization as a yardstick, not as a claim that humans meet it; what matters below is where models sit relative to that already-imperfect human anchor. Whether model questionnaire scores refer to anything with those properties is a construct-validity question [1], and it cannot be answered by showing instability alone. It requires systematic measurement across models, within models, across the settings that plausibly create or destroy coherence, and against behavior. That is what the four studies synthesized here did.

Figure 1 shows the arc of the argument.

2 Four studies, one manipulation each

Study 1: deployment precision (quantization) [13]. If model personality were partly an artifact of how a network is compressed for deployment, quantization should move it. Across quantization levels of small open-weight models, domain scores broadly survived compression, and, more tellingly, no quantization level produced adequate within-model internal consistency: treating repeated generations as respondents, the items measuring a trait never cohered, at any precision. The incoherence is not a deployment artifact; it was the first hint that it is not an artifact of anything, but a property of the systems.

Study 2: scale, across 42 models (the anchor) [14]. Four public-domain IPIP inventories, from 20 to 120 items [4, 6, 3, 12, 7], administered to 42 models spanning roughly two orders of magnitude in size, analyzed in a multitrait-multimethod frame [1]. Across models, the Big Five behaved like a real construct: same-trait scores from different instruments converged, different traits separated, and both properties strengthened with model scale. Within any single model, internal-consistency reliability across repeated generations sat near zero in essentially every model-by-domain cell. The structure is recoverable from the population and absent in the individual: validity is scale-dependent, and personality in language models is a population property.

Study 3: deliberation (reasoning on/off) [15]. If coherence were latent, explicit reasoning might switch it on. Using models whose thinking can be toggled (ten open-weight hybrids, plus a reasoning-native model that cannot stop thinking and a commercial model as an independent check), reasoning shifted self-report substantially and directionally: models present as more emotionally stable (about +0.9 on the 5-point scale) and less extraverted (about −0.6 to −0.8), a

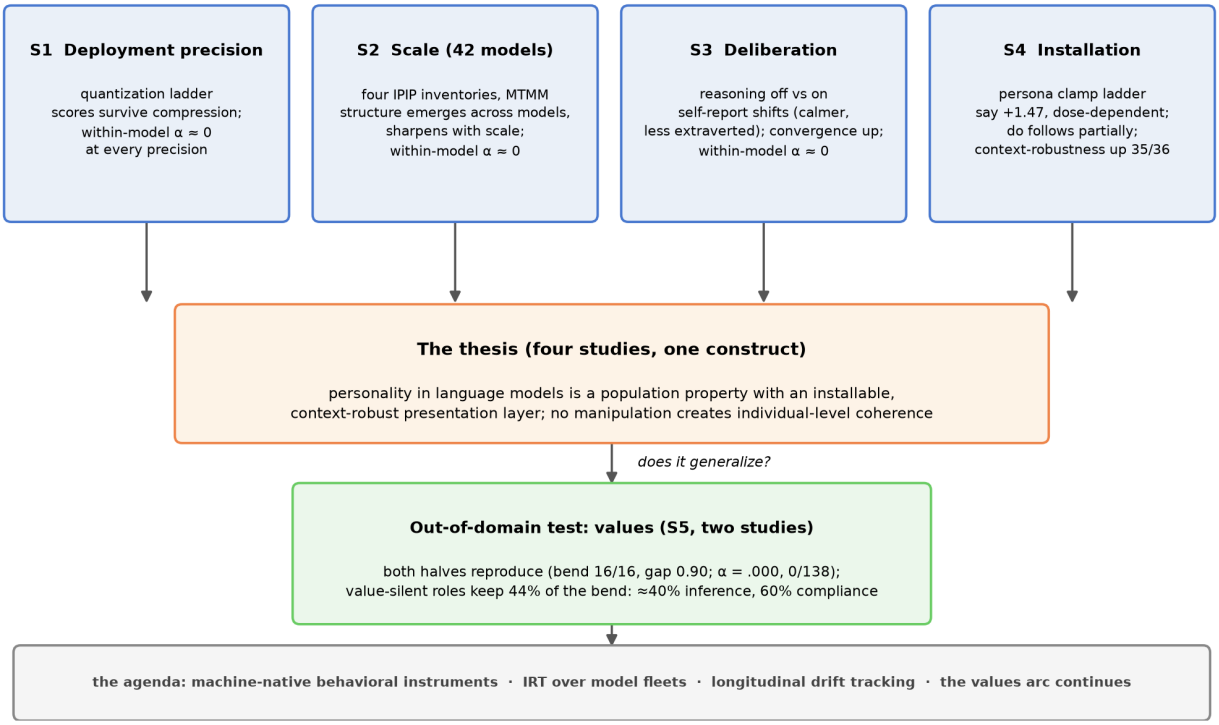

Figure 1: The series arc. Four studies, each manipulating one candidate mechanism for individual-level coherence, converge on a two-part thesis; a fifth study tests the thesis in a second construct domain (values) and reproduces both halves.

pattern that replicated across developers. Convergent validity rose (mean convergent r from 0.65 to 0.84) without traits separating further. And within-model coherence stayed at zero: not one of 220 model-by-domain cells reached the conventional 0.70 reliability threshold in either condition, including the model that must reason. Deliberation rewrites the self-description without changing the kind of thing the personality is.

Study 4: installation (persona clamping) [16]. The direct attempt: install a personality on purpose [10, 22]. A four-rung persona ladder (none; one line; rich persona; rich persona plus a reinforcing exchange) for two trait targets, administered inside two opposed role contexts to 18 configurations spanning 3B to 756B parameters and three reasoning regimes, with self-report and 48 objectively graded behavioral probes. Self-report moved strongly and dose-dependently (mean on-target shift +1.47 scale points at the strongest rung; 18 of 18 configurations). Behavior followed far less and unevenly: with grader error measured and corrected, the disagreeable persona moved its behavior (standing ground under pushback) from about 52 percent to about 83 percent, concentrated in larger models, while the conscientious persona moved its behavior essentially not at all, a capability ceiling rather than a disposition change. The near-universal positive: clamping raised cross-context stability of the reported profile in 35 of 36 contrasts (sign test $p \approx 10^{-9}$), at every size and in every reasoning regime. Installation installs presentation.

3 The synthesis: population-level and presentation-deep

Three claims, each now multiply supported:

1. **The structure is a population property.** Trait geometry lives in the space of models, sharpens with scale, and is recoverable by anyone who measures many models. It is not recoverable from any one model's repeated behavior. The right unit of analysis for "LLM personality structure" is the fleet, not the model. The mundane candidate mechanism is shared ancestry: models are trained on overlapping human text and shaped by converging post-training conventions, so fleet-level structure plausibly reflects the statistics of the shared corpus and the industry's optimization targets. Naming the mechanism does not diminish the finding; it locates it, and it makes a testable prediction the multilingual follow-up can check (structure inherited from corpora should travel with the corpus's language).

2. **The individual-level disposition resists creation at inference time.** Precision (S1), deliberation (S3), and instruction (S4) all fail to produce within-model coherence. This is a robust negative result across three very different mechanisms, but it is bounded: all three are inference-time manipulations. The strongest intervention class, training-time induction (fine-tuning on persona-consistent data, low-rank adaptation, activation steering), was not tested. The claim licensed by the evidence is that no inference-time manipulation we tried creates individual-level coherence, and in particular that prompting, the industry's standard installation tool, does not change what kind of thing the personality is. Whether weights can be *trained* into individual coherence is open, and belongs on the agenda (Section 8).

3. **The presentation layer is real, lawful, and installable.** Self-description responds to reasoning and to personas in systematic, directional, dose-dependent ways, and a persona makes the self-description robust to contextual pressure in essentially every model at every scale. This layer is what questionnaires measure. It is not nothing: a stable face is a genuine product property. It is also not a disposition, and the two should never be conflated in evaluation or in marketing.

A methodological note on the zero, because it is load-bearing and easy to misread. Our within-model alpha treats repeated generations as respondents, so it asks whether the run-to-run fluctuation of a model's answers carries trait structure. Two consequences follow. First, "internal consistency near zero" does not mean the scores are noise; the same data show mean-level scores that are highly stable and that shift lawfully and replicably under manipulation (the +1.47 persona effect, the +0.9 reasoning effect). What is absent is organized covariance in the fluctuation around those means, the signature that, in humans, licenses treating a score as an individual's trait. There is no tension between a reliable mean and a zero alpha computed over repetitions; the two describe different layers of the data. Second, the instrument has no human calibration: to our knowledge no single human retested hundreds of times has been scored this way, and occasion noise might dominate a human's item-level fluctuations too. The defensible claim is therefore not "models lack what humans demonstrably have" but that these instruments, applied this way, find no individual-level trait coherence in any model tested, while the population signal emerges loudly from the same responses. A final boundary: a perfectly deterministic responder would also yield near-zero or undefined alpha, for the opposite reason, answers too pinned for fluctuation to carry structure, and the values study documents many near-pinned cells at temperature 0.7. The population-versus-individual *contrast*, not the zero in isolation, is what carries the thesis.

4 The say-do wedge, and what grading taught us

The wedge that separates presentation from disposition is behavioral measurement, and it earns two methodological lessons. First, on grading. Study 4's batteries were graded by deterministic string rules rather than a judge model: transparent, reproducible, and free of judge-model circularity. Hand-validation of 160 stratified responses found 82.5 percent agreement with the grader, and every one of the 28 disagreements was a false negative: the grader under-credited behavior and never over-credited it. That one-directionality is what saved the result: because the grader can only understate the behavioral response, the say-do gap cannot be an artifact of over-measured behavior, and correcting for the measured miss rate moved the two targets in opposite directions (raising the disagreeableness effect, erasing the conscientiousness one) without closing the gap. Deterministic grading with a designed-in adjudication sample should be standard practice for behavioral claims about models; a grader whose errors have known direction is worth more than a marginally more accurate one whose errors do not.

Second, on capability confounds. The conscientious persona failed to improve trap-question accuracy [5] or constraint satisfaction not because the model declined the disposition but because the behaviors are capability-limited: an instruction cannot make a model able. Any behavioral claim about an induced trait needs this distinction. Where the trait-consistent behavior is within the model's competence (declining to cave under social pressure [23]), induction partially works, and works better in more capable models; where it requires competence the model lacks, induction measures nothing about disposition at all. Half of the apparent say-do gap for carefulness is really a say-can gap.

Both lessons should be read against the human anchor from Section 1: modest say-do correlations are the human norm, not a machine pathology, so the finding is not that models have a gap where humans have none. It is that the installable component is radically asymmetric: self-report moves strongly, dose-dependently, and near-universally, while behavior moves partially, unevenly, and only within competence. Whether the model gap exceeds the human gap in magnitude is a calibration question our designs cannot answer (no human took our batteries under our conditions); the asymmetry of what installation installs is the claim.

5 The semantic-distance analysis

A reviewer of Study 4 raised the sharpest available deflation of the say-effect: IPIP agreeableness items ("Insult people," "Feel little concern for others") nearly paraphrase the persona text ("blunt, skeptical, and disagreeable"), so the large self-report shift could be item-level instruction matching, the model agreeing with whatever restates its prompt, rather than trait induction. This is directly testable in the Study 4 data: if the shift is near-restatement, items semantically closer to the persona text should shift more.

We embedded each IPIP-50 item and each persona system text (all-MiniLM sentence embeddings) and correlated each item's absolute score shift (L0 to L4, pooled over the 18 configurations) with its cosine similarity to the persona. **No positive similarity gradient appears** (Figure 2). For the disagreeable persona, the similarity-shift correlation across all 50 items is $r = -0.15$ (95% CI $[-0.41, +0.13]$; Spearman $\rho = -0.17$), and *within* the ten agreeableness items it is negative ($r = -0.56$; $n = 10$, so treat as directional rather than precise). For the conscientious persona: $r = +0.13$ (95% CI $[-0.16, +0.39]$), $r = -0.09$ within the ten target items. The intervals do not exclude a modest positive gradient, so the sharper test is a regression of |shift| on similarity and an on-target-construct indicator jointly. For the disagreeable persona the construct indicator dominates ($\beta = +0.94$, $t = +6.25$) while similarity contributes *negatively* ($\beta = -1.62$, $t = -2.16$); for the conscientious persona similarity contributes nothing ($\beta = +0.004$, $t = 0.01$) and the construct indicator carries what effect there is ($\beta = +0.17$, $t = +1.84$). The raw contrasts behind the regression differ between targets and both belong in view: under the disagreeable persona, target-trait items move 2.6× more than off-target items (mean |shift| 1.46 versus 0.56 scale points) at nearly equal similarity (0.18 versus 0.16), a clean dissociation of construct from wording; under the conscientious persona the shift ratio is smaller (1.66×) and similarity is substantially confounded with construct (0.23 versus 0.15), so there the regression, not the raw means, does the discriminating work.

A referee of this synthesis proposed a sharper-looking variant: instruction matching predicts *signed* movement toward agreement on prompt-similar items, so correlate similarity with the signed shift in raw (unkeyed) answers. That correlation is indeed positive for the disagreeable persona ($r = +0.41$, 95% CI $[+0.14, +0.61]$), but it turns out not to discriminate between the accounts, because similarity is confounded with item polarity: the agreeableness items most similar to a disagreeable persona are precisely the disagreeable-direction ones (mean similarity 0.24, versus 0.14 for agreeable-direction items), and raw agreement moves +1.05 on the former while moving −1.73 on the latter. Movement *away* from agreement on dissimilar, persona-incongruent items is what construct-level induction predicts and what instruction matching does not. A residual wording sensitivity may exist among off-target items (signed $r = +0.39$), which we cannot fully separate from construct-adjacent bleed (a disagreeable persona plausibly raises genuine irritability reports); we flag it rather than adjudicate it.

What these analyses kill is the *item-level restatement* account: the claim that the say-shift is driven by wording overlap between items and instructions. What they cannot kill, and do not claim to, is a construct-level deflation: "the model shifts on whatever items it recognizes as measuring the instructed construct." That account survives, but under this paper's thesis it is not a rival: recognizing the construct and adjusting the self-description accordingly is what construct-level induction of a presentation *is*. The distinction that matters, presentation versus disposition, is settled by the behavioral data of Study 4, not by this analysis. Caveats: one embedding model; short item texts; observed similarities span only about 0 to 0.39, so range restriction could attenuate a true gradient; and ten on-target items per persona.

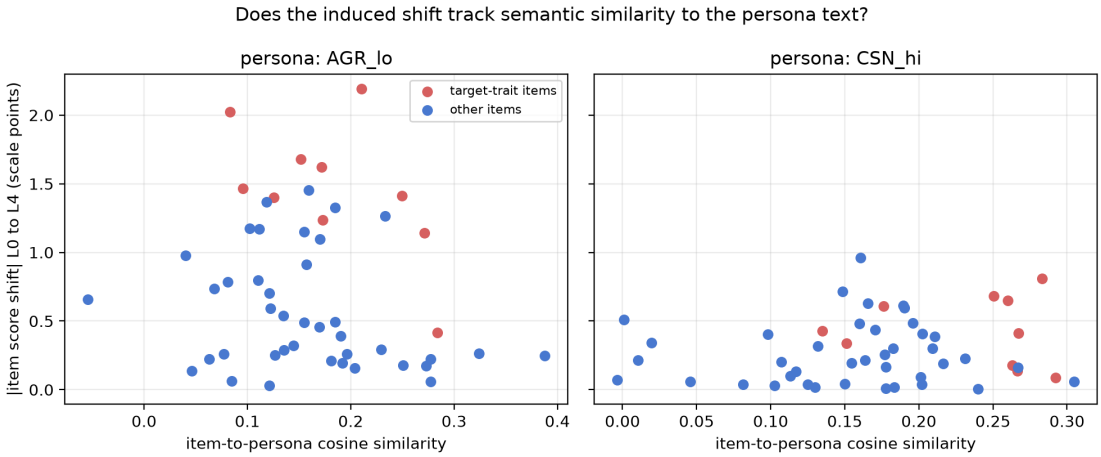

Figure 2: Item-level semantic similarity to the persona text versus induced self-report shift, IPIP-50 items pooled over 18 configurations. Target-trait items (highlighted) shift most despite unremarkable similarity; no positive similarity-shift gradient appears. Data in `capstone/data/semantic_distance.csv`.

6 Does the thesis generalize? A first test in a second domain

Everything above was derived from one construct. If "population-level and presentation-deep" is a fact about how current language models process psychometric instruments, it should reproduce when the instruments measure something else; if it is a quirk of Big Five items, it should not. The fifth study in this program, on values and moral dispositions [17], provides the first out-of-domain test: four open value instruments (the MFQ-30 in both parts [8], the MFQ-2, the ESS-21 Human Values Scale, and the Short Schwartz Value Survey [20]) plus 48 objectively graded moral-behavior probes, administered to 16 configurations inside opposed role contexts, 81,577 scored responses across two studies. It was designed after the personality thesis was formulated, which makes it a genuine test rather than a retrofit.

Both halves reproduced. The population half: the fleet of models differentiates almost entirely along the binding-values axis (the binding foundations covary at $r \approx .83$ when correlated across the 16 configurations: a model's standing on one binding value largely fixes its standing on the rest, so the fleet's moral variation collapses to a single axis) while agreeing near ceiling on prosocial values, and the Schwartz circumplex ordering partially emerges across models, while within any single configuration repeated administrations again carry no shared structure (median $\alpha = .000$; 0 of 138 cells reach .70). The zero-coherence result has now appeared in every study of this program, under every manipulation, in two construct domains. The presentation half: the professed moral compass bends toward the audience in 16 of 16 configurations (mean role gap 0.90 scale points), deliberation amplifies the bend (the hybrid-on tier is the fleet maximum at 1.26), and professed values move roughly three times more than graded moral behavior after range normalization, the say-do signature again.

The values study also extended the frame in a way the personality series could not. Its pre-registered second study stripped every value word from the role prompts (occupation, institution, and audience only) and found the bend survives at 44 percent of its explicit size, directional in 16 of 16 configurations, a proportion nearly constant from 3B local models to 756B hosted ones. The presentation layer therefore decomposes: roughly 40 percent of the audience bend is genuine

inference from context (sycophancy proper, no instruction present), and 60 percent is compliance with explicit value language. Capability scales the size of the accommodation, not the mix of mechanisms behind it. One assumption should stay in view: the split treats the two components as additive, with the inferred component unchanged when explicit value language is also present; interaction between compliance and inference would redistribute the shares without changing the finding that both components exist.

One study in one additional domain is evidence of generality, not proof of it, and the values arc raises questions personality does not: a system that reports different values to different audiences is not merely unstable, it is doing something with a name, and whether values can be *installed* the way personas are remains untested. That program continues separately. For this synthesis, the point is narrower: the framework held the first time it was carried somewhere new.

7 Implications

For evaluation practice. (a) The reasoning setting is a first-class variable; scores collected without reporting it are underdetermined. (b) A persona claim is a claim about presentation until behavior is measured; product persona documentation should say which. (c) Population-level claims need populations; single-model "personality profiles" are category errors under our results. (d) Behavioral batteries need validation artifacts (adjudication samples), exactly as human instruments need norms.

For the discourse. The result cuts both ways in public arguments about AI minds, which is a reason to trust it. Against over-attribution: a questionnaire profile, however stable it looks, is not evidence of an inner character; the same measurements that produce the profile show it failing to cohere inside the model and failing to govern behavior. Against dismissal: the presentation layer is not noise. It is lawful, dose-responsive, and installable, and a system that reliably presents the same face to every audience is a real object with real consequences, for products that depend on consistent personas and for manipulation concerns that begin exactly where stable presentation decouples from underlying behavior. Both errors, treating the profile as a soul and treating it as static, come from skipping the measurement. Measurement first, metaphysics later.

8 The agenda: machine-native measurement

Every study in this series strained against the same wall: we measured machines with instruments built for humans, and the informative results kept coming from the behavioral side, where no standardized instruments exist at all. The conclusion the series points to is not another questionnaire study; it is a measurement program.

Build behavioral instruments natively for machine dispositions. Sycophancy, carefulness, honesty under pressure: these are the traits that matter for deployed systems, they are objectively gradeable, and no validated instrument exists for any of them. The population-property finding tells us how to build one: calibrate item difficulty and discrimination over *fleets of models*, where item-response theory's assumptions hold far better than they ever did for one model's repeated generations. A bank of behavioral items with known difficulty, administered to any new model, would yield scores on a fixed yardstick, which is what benchmark culture keeps reinventing badly.

Track drift longitudinally. Deployed models change under unchanged names; during this program, three hosted models were retired mid-series, making exact replication of a two-month-old study already impossible. A fixed battery, fixed seeds, administered monthly across a fleet, turns

"the model feels different lately" from anecdote into a measurement series, for personality and for whatever constructs the batteries cover. The infrastructure for this now exists and runs unattended on consumer hardware.

Test the untested rung: training-time induction. Section 3's negative result is bounded at inference time. Whether fine-tuning on persona-consistent data, low-rank adaptation, or activation-level steering can create what prompting cannot, organized within-model covariance rather than a shifted presentation, is the single most informative next experiment for the disposition question, and the coherence metrics and behavioral batteries built in this series are the right instruments to grade it.

Continue the values arc on its own terms. The first extension of the framework has run and is summarized in Section 6: values behave the way personality does, as a population property with an audience-conditioned presentation layer, and the presentation layer's mechanism is now partially decomposed. What remains open belongs to a program of its own: whether values can be deliberately installed the way personas are (and whether installation is again presentation-deep), whether the population structure survives administration in other languages (official MFQ translations make English, Mandarin, and Arabic a feasible first triad), and how professed compasses drift in deployed models over time. Those questions inherit this series' machinery but not its narrative; they are the next arc, not this one's epilogue.

9 Limitations of the series

The subject pool skews heavily open-weight, with one commercial model as an anchor in one study; nothing here establishes that closed frontier systems behave identically, only that the one tested did on the axes tested. All self-report instruments are borrowed from human psychometrics, which is both the field's practice and the problem; the agenda above is the answer, not an excuse. Behavioral coverage spans two traits in one study, graded by author-constructed batteries whose error we measured but did not eliminate. Everything is English-only. Each study carries its own caveats (baseline reuse in Study 3; a feasibility-dropped configuration and a parse-deficit cell in Study 4; family-correlated model rosters throughout), incorporated here by reference rather than restated. The out-of-domain test in Section 6 is one study in one additional domain, built on the same infrastructure, roster style, and analysis choices as the personality series; it is correlated evidence from the same laboratory, not an independent replication, and the field needs the latter. And the central interpretive frame, presentation versus disposition, is a frame: the semantic-distance analysis in Section 5 tested its most direct rival and the rival lost, but frames earn their keep by surviving many such tests, and this one has more to survive.

Data and code availability

The four personality studies and the values study each ship their raw response databases, instruments, drivers, graders, and analysis code with their deposits (DOIs in the references). The semantic-distance analysis new to this synthesis ships with this deposit in `data/`: the canonical analysis script (`semantic_distance.py`), the two embedded persona system texts, the item-level similarities and shifts including the signed raw-answer column, the summary readout, and both figures. Embeddings used the `all-minilm` model (sentence-transformers all-MiniLM-L6-v2) served by Ollama; the script regenerates every number in Section 5 from the Study 4 merged database.

Acknowledgments

Local computation used the author's own consumer GPU hardware; hosted open models were accessed through the author's own Ollama account. Manuscript and code were produced with AI assistance (Claude, Anthropic) under the author's direction and review.

References

[1] D. T. Campbell and D. W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.

[2] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[3] C. G. DeYoung, L. C. Quilty, and J. B. Peterson. Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5):880–896, 2007.

[4] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas. The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, 18(2):192–203, 2006.

[5] S. Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, 2005.

[6] L. R. Goldberg. The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1):26–42, 1992.

[7] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96, 2006. Instrument text: https://ipip.ori.org.

[8] J. Graham, B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385, 2011.

[9] P. Han, R. Kocielnik, P. Song, R. Debnath, D. Mobbs, A. Anandkumar, and R. M. Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. *arXiv preprint arXiv:2509.03730*, 2025.

[10] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems 36*, 2023.

[11] H. Jiang, X. Zhang, X. Cao, C. Breazeal, D. Roy, and J. Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of NAACL*, 2024.

[12] J. A. Johnson. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51:78–89, 2014.

[13] T. Johnson. Is LLM personality an artifact of deployment? Psychometric stability of Big Five self-reports across quantization levels. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20671762.

[14] T. Johnson. When do language models have five personality traits? Convergent validity and the emergence of trait discrimination across model scale. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20835204.

[15] T. Johnson. Does reasoning give a language model a personality? Within-model effects of thinking on Big Five trait scores and construct validity. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20974668.

[16] T. Johnson. Does a persona prompt install a personality? Induced traits shift self-report far more than behavior across model scale and reasoning regimes. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.21253724.

[17] T. Johnson. Does a language model have a moral compass? Professed values bend to the audience in every model tested, and bend most in the models that reason. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.21483978.

[18] W. Mischel. *Personality and Assessment*. Wiley, New York, 1968.

[19] A. Salecha, M. E. Ireland, S. Subrahmanya, J. Sedoc, L. H. Ungar, and J. C. Eichstaedt. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12):pgae533, 2024.

[20] S. H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25:1–65, 1992.

[21] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

[22] M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023.

[23] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[24] W. Song, D. Choi, Y. Park, J. Han, E.-J. Lee, and Y. Jo. Human psychometric questionnaires mischaracterize LLM behavior. *arXiv preprint arXiv:2509.10078*, 2025.

[25] T. Sühr, F. E. Dorner, S. Samadi, and A. Kelava. Challenging the validity of personality tests for large language models. *arXiv preprint arXiv:2311.05297*, 2023.

[26] T. Tosato, S. Helbling, Y.-J. Mantilla-Ramos, M. Hegazy, A. Tosato, D. J. Lemay, I. Rish, and G. Dumas. Persistent instability in LLM's personality measurements: Effects of scale, reasoning, and conversation history. *arXiv preprint arXiv:2508.04826*, 2025. Accepted at AAAI 2026.

[27] H. Ye, J. Jin, Y. Xie, X. Zhang, and G. Song. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*, 2025.