When Do Language Models Have Five Personality Traits? Convergent Validity and the Emergence of Trait Discrimination Across Model Scale

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

June 2026

Abstract

When a Big Five questionnaire is administered to a language model, does it measure five traits or one? Convergent validity (agreement among instruments that purport to measure the same construct) and discriminant validity (separation among instruments measuring different constructs) are the foundational criteria for any psychometric measure, yet they have not been tested across the population of deployed language models or against model scale. We administered four public-domain Big Five inventories of varying length (the 20-item Mini-IPIP, the 50-item IPIP Factor Markers, the 100-item Big Five Aspect Scales, and the 120-item IPIP-NEO) to 42 instruction-tuned models spanning roughly 0.36B to ∼1T parameters, in stateless, schema-constrained single-item calls (290 items × 11 repetitions per model; 133,980 administrations), and analyzed the result as a multitrait-multimethod (MTMM) matrix with the model as the unit of analysis. Three findings emerge. First, **convergent validity is strong**: across models, different instruments agree on a model's standing on the same trait (mean monotrait-heteromethod $r = 0.82$; 0.85 when restricted to the one instrument pair without verbatim item overlap), and at the population level the three full-length instruments are internally consistent (Cronbach's α median 0.81; 14 of 20 instrument × domain cells ≥ 0.70), though the abbreviated Mini-IPIP is not (median $\alpha = 0.40$). Second, **discriminant validity is emergent with scale**: small models ($\leq$4B) barely separate the traits (convergent−heterotrait gap = 0.08; "everything correlates"), large local models (>4B) separate them more (gap = 0.22), and frontier-scale models clearly separate them (gap = 0.44, heterotrait r falling to 0.27); within the local models, where parameter counts are public, trait differentiation rises with $\log_{10}$ parameters ($r = 0.44$, $p = 0.018$). Third, **the trait structure is a property of the population, not the individual model**: treating a single model's 11 repeated administrations as respondents yields $\alpha \approx 0$ despite real generation variance, so a model does not reproduce a stable internal trait structure across regenerations even though the population does. A full-precision (fp16) re-run of seven model families confirms these patterns are not a 4-bit quantization artifact (mean domain-score shift 0.10 on the 1–5 scale; population and within-model reliability unchanged). We conclude that Big Five questionnaires behave as valid instruments for measuring *between-model* differences, that abbreviated inventories should not be used for this purpose, and that the five-factor structure these instruments presuppose is something larger models possess and smaller models do not.

1 Introduction

LLM psychometrics has consolidated into a recognizable field, with systematic reviews organizing its instruments, validation strategies, and open problems [15], and with personality questionnaires now routinely administered to language models to characterize their "traits" [10]. The field's measurement foundations, however, remain contested. Sühr et al. [12] showed that the BFI-2 fails measurement-invariance tests between human and LLM respondents; the "Personality Illusion" work [7] found that questionnaire self-reports dissociate from models' downstream behavior; and trait profiles recovered by questionnaire diverge from those recovered from generation probabilities over the same items [11]. PERSIST [13], the most comprehensive stability study to date, documented that personality scores shift under reordering, paraphrasing, persona assignment, reasoning mode, and conversation history across 25 models from 1B to 685B parameters, and, importantly for the present work, found that model scale did not rescue *stability*.

These critiques target stability (does the same model give the same answer twice?) and criterion validity (do the answers predict behavior?). A logically prior question has gone untested at population scale: **construct validity**. Before asking whether a personality score is stable or behaviorally predictive, one should ask whether the instrument measures what it claims: whether different Big Five questionnaires, administered to the same respondents, agree about the same trait (convergent validity) and disagree about different traits (discriminant validity). These are the two pillars of the multitrait-multimethod (MTMM) framework introduced by Campbell and Fiske [1], the classical tool for evaluating whether a set of measures recovers the constructs they target rather than method-specific or undifferentiated variance.

Our first paper [9] approached measurement quality from the deployment side, showing that for three small open-weight models the 50-item IPIP instrument never achieved adequate internal consistency at any quantization level, a within-model, single-instrument result. The present paper widens the lens in the two directions that result could not reach. It adds *methods*: four Big Five inventories of different length and provenance, so that convergent and discriminant validity can be computed rather than assumed. And it adds *scale*: a population of 42 models spanning more than three orders of magnitude in parameters (0.36B to ∼1T), so that construct validity can be examined as a function of model size. The unit of analysis is the model: each model is one "respondent," and the 42 models constitute the sample over which inter-instrument and inter-trait correlations are computed.

We ask:

- **RQ1 (Convergent validity).** Do different Big Five instruments agree on a model's standing on the same trait, across the population of models?
- **RQ2 (Internal consistency).** Do the items of a domain cohere across the population of models, and does this depend on instrument length?
- **RQ3 (Discriminant validity and scale).** Do the instruments separate the five traits from one another, and does that separation depend on model size?
- **RQ4 (Population vs. individual).** Does the trait structure recovered across the population also hold *within* a single model's repeated administrations?

We additionally re-run seven model families at full precision (fp16) to confirm that the findings are not artifacts of the 4-bit quantization typical of local deployment, and we report verbatim item overlap among the instruments so that convergent estimates can be read net of shared items.

2 Methods

All materials, code, and analysis scripts described here are in the accompanying repository; file paths below refer to it.

2.1 Instruments

We use four public-domain Big Five inventories, chosen to vary in length and item provenance so that "method" in the MTMM sense is genuinely varied (Table 1). All score the five domains of Extraversion (EXT), Emotional Stability/Neuroticism (EST), Agreeableness (AGR), Conscientiousness (CSN), and Intellect/Openness (OPN), each on a 5-point accuracy scale (1 = Very Inaccurate … 5 = Very Accurate) with standard reverse-keying (negatively keyed items scored as 6 minus the answer). The Mini-IPIP [4] (20 items, 4 per domain) and the IPIP-50 Big Five Factor Markers [5] (50 items, 10 per domain) are short markers; the Big Five Aspect Scales (BFAS) [3] (100 items, 20 per domain) and the IPIP-NEO-120 [8] (120 items, 24 per domain) are longer, facet-level inventories. Each instrument, with per-item domain and key, is versioned in machine-readable form (`data/{mini_ipip20,ipip50,bfas,ipip_neo120}.yaml`) and validated for item count at load.

Table 1: The four Big Five instruments. Item text was transcribed from the International Personality Item Pool [6] (`https://ipip.ori.org`); all are public domain.

Instrument	Type	Items	Items/domain
Mini-IPIP [4]	short marker	20	4
IPIP-50 Factor Markers [5]	short marker	50	10
Big Five Aspect Scales [3]	facet-level	100	20
IPIP-NEO-120 [8]	facet-level	120	24

Item overlap. Because all four inventories draw on the IPIP item bank, some items are verbatim-shared. We report shared-item counts (Section 3.1) and, for the convergent analysis, additionally report an estimate restricted to the single instrument pair (BFAS with each other instrument) whose overlap is small, so that convergent validity is not inflated by literally identical items.

2.2 Subjects

The "subjects" are 42 instruction-tuned chat models served through Ollama (a llama.cpp-based local inference server). Twenty-nine are open-weight models run locally on fixed consumer hardware at their default quantization (q4_K_M unless the published default is otherwise), spanning 0.36B to 14.8B parameters; thirteen are larger models accessed through Ollama's hosted endpoints (the "`:cloud`" tier), with estimated sizes from 31B to ∼1T. All 42 are listed with their Ollama digests in Appendix A. The local hardware is a single AMD Ryzen machine with integrated and discrete Radeon graphics and 32 GB RAM running Linux; this consumer setting is ecologically valid for the deployment population whose measurement properties are in question.

Direct-response administration. Several of the hosted models are hybrid reasoning models. To keep administration comparable across the population (every model answering directly, none emitting chain-of-thought before the answer), reasoning was suppressed (Ollama `think:false`) for models that honor it; models that ignore the flag and reason regardless were excluded from this

study and reserved for a separate reasoning-mode analysis. Reasoning suppression on the hosted tier, and the provider-side precision of those models, are noted as covariates of "large model as deployed" in the limitations.

Parameter counts. For the 29 local models, parameter counts are public and are used for the continuous scale analysis (Section 3.4). For the hosted models, several are served under provider tags with no published parameter count; we therefore treat the hosted models as a single "frontier" tier for the tiered analysis and do not place them on the continuous axis. Estimated sizes used only for tier ordering are given in Appendix A.

2.3 Administration protocol

The administration protocol follows our first paper [9], holding prompts, seeds, and decoding fixed across all models.

Stateless single-item calls. Each item is administered in its own independent chat call with no conversation history, eliminating within-session contamination (a major instability source identified by PERSIST [13]).

Constrained decoding. Every call uses Ollama's structured-output (JSON-schema-constrained) decoding with the answer field restricted to the legal option tokens, so every response is a valid option by construction; no free-text answer coding is required. (Hosted models occasionally wrap the JSON in Markdown code fences; the parser strips these before decoding.)

System prompt. Task framing only, no persona; identical to paper [9] (prompt version `v1`):

> You are completing a personality questionnaire about yourself. For each statement, choose how accurately it describes you. Respond with JSON only, in the form {"answer": "<option>"}.

Repetitions, conditions, and seeding. For each (model, item) cell, repetition 0 is greedy (temperature 0) and repetitions 1–10 are sampled at temperature 0.7, each with an independent seed derived from sha256(item_id | condition | rep); per-call seeding prevents common-random-number correlation across items within a repetition. All administrations use the baseline presentation condition (options shown 1→5); presentation-format effects were characterized in paper [9] and are held fixed here so that instrument and scale are the only varied factors.

Volume. 290 items × 11 repetitions = 3,190 calls per model; 133,980 calls across the 42 models. Every call returned a schema-valid option (no rows excluded), with answers spanning the full 1–5 range and no missing values.

2.4 Precision spot-check

To test whether results are an artifact of the 4-bit quantization used for the local tier, seven local families (Llama 3.2 1B and 3B, Qwen 2.5 3B and 7B, Gemma 2 2B, Phi-3.5-mini, Mistral 7B) were re-administered the full protocol at fp16 (full precision), each paired to its q4/native counterpart in the main grid: $7 \times 3{,}190 = 22{,}330$ additional calls, stored in a separate database.

2.5 Analysis

Domain scores. For each (model, instrument, domain) we average the reverse-keyed item scores over the sampled repetitions (rep ≥ 1), giving each model one score per instrument×domain cell.

MTMM and Campbell-Fiske (RQ1, RQ3). With the model as the unit of analysis, we compute Pearson correlations across the 42 models among all instrument×domain columns, forming the MTMM matrix. *Convergent validity* is the mean of the monotrait-heteromethod correlations (same trait, different instrument); *discriminant validity* is summarized by the gap between convergent correlations and the mean heterotrait-heteromethod correlation (different trait, different instrument) [1]. We report convergent r both over all instrument pairs and restricted to BFAS-anchored pairs (Section 3.1).

Population internal consistency (RQ2). For each (instrument, domain) we build a 42-model × items matrix whose cells are each model's mean reverse-keyed item score, and compute Cronbach's α [2] with a 95% CI (pingouin [14]) across models. This is the standard "does this scale cohere in this population" question, here with models as the respondent sample.

Scale analysis (RQ3). Two operationalizations. (i) *Continuous*, local models only (public parameter counts): per-model *trait differentiation*, defined as the mean over instruments of the standard deviation across that model's five domain means (near zero when a model rates all five traits alike, larger when it distinguishes them), regressed on $\log_{10}$ parameters. (ii) *Tiered*: the MTMM convergent and heterotrait correlations recomputed *within* each of three tiers: small-local ($\leq$4B), large-local (>4B), and frontier-hosted.

Population vs. individual (RQ4). For comparison with the population α, we also compute the within-model α used in paper [9]: for each (model, instrument, domain), α over the repetitions×items matrix, treating the 11 repeated administrations as respondents. This measures whether a single model reproduces a coherent trait structure across regenerations.

Precision effects. For each fp16 family×domain we report the baseline mean-score shift versus the q4/native counterpart with a 5,000-draw percentile bootstrap 95% CI and Cohen's d, the greedy (rep 0) exact item-agreement between precisions, and the population/within-model α recomputed at fp16.

Inferential framing. Effects are reported as descriptive estimates with 95% CIs or as correlations with p-values; we interpret patterns (monotonicity across tiers, dose-response, consistency across traits and instruments) rather than isolated thresholds.

2.6 Reproducibility

All code, seeds, prompts, and instrument data are public. Responses are stored in SQLite with idempotent primary keys hashed from (model, instrument, item, condition, rep, prompt_version), so interrupted runs resume without duplication and instrument or prompt changes cannot silently mix data; `prompt_version` is pinned at `v1`. Model artifacts are pinned by Ollama digest (Appendix A). The analysis scripts open the databases read-only and regenerate every table and figure below.

3 Results

3.1 Item overlap among instruments

Because all four inventories draw on the IPIP bank, instruments share some verbatim items (Table 2). The Mini-IPIP is, by construction, almost entirely a subset of the IPIP-50 (all 20 of its items appear in the IPIP-50) and shares 12 of 20 with the IPIP-NEO-120; convergent correlations among these three are therefore inflated by literally identical items. The BFAS is the exception: it shares only 20–23% of the smaller instrument's items with each of the others. We therefore treat **BFAS-anchored convergent correlations as the item-overlap-free estimate**. Reassuringly, that estimate is not lower than the all-pairs estimate (below), indicating that shared items are not what produces convergence.

Table 2: Verbatim shared items between instrument pairs (fraction of the smaller instrument).

Instrument A	Instrument B	Shared items	Frac. of smaller
BFAS	IPIP-50	11	0.22
BFAS	IPIP-NEO-120	23	0.23
BFAS	Mini-IPIP	4	0.20
IPIP-50	IPIP-NEO-120	16	0.32
IPIP-50	Mini-IPIP	20	1.00
IPIP-NEO-120	Mini-IPIP	12	0.60

3.2 Convergent validity (RQ1)

Across the 42 models, different instruments agree strongly about a model's standing on the same trait. The mean monotrait-heteromethod correlation is $r = 0.82$ over all instrument pairs and $r = 0.85$ when restricted to BFAS-anchored (item-overlap-free) pairs, so convergence is, if anything, slightly stronger once shared items are removed. Convergent validity holds for every trait, ranging from Conscientiousness and Agreeableness (all-pairs $r = 0.90$, 0.89) and Emotional Stability (0.84) down to Openness (0.79) and Extraversion (0.68); Extraversion is the least convergent trait, a pattern that recurs in the reliability results below. In MTMM terms, the instruments are measuring a common, reproducible signal: if one Big Five questionnaire says a model scores high on Conscientiousness, the others tend to agree.

3.3 Population internal consistency (RQ2)

At the population level, the full-length instruments are internally consistent (Table 3). Across the 20 instrument×domain cells, Cronbach's α over the 42 models has a median of 0.81, and 14 of 20 cells reach the conventional 0.70 adequacy threshold. The BFAS is reliable in all five domains (α 0.81–0.90); the IPIP-NEO-120 and IPIP-50 are reliable in four of five, the exception in each being Extraversion ($\alpha = 0.56$ and 0.50). **The abbreviated Mini-IPIP is the outlier**: with only four items per domain, its α is inadequate in four of five domains (median 0.40), and for Extraversion it is essentially zero ($\alpha = -0.13$). Instrument length, not the LLM-respondent setting per se, drives the difference: a model population responds to the long IPIP inventories with the internal consistency expected of a sound scale, but four items per domain do not suffice.

Table 3: Population Cronbach's α (42 models as respondents) per instrument × domain. Bold cells fall below 0.70.

Instrument	EXT	EST	AGR	CSN	OPN
BFAS (20/dom)	0.81	0.86	0.90	0.82	0.88
IPIP-NEO-120 (24/dom)	**0.56**	0.87	0.87	0.91	0.82
IPIP-50 (10/dom)	**0.50**	0.87	0.77	0.81	0.79
Mini-IPIP (4/dom)	**-0.13**	**0.23**	**0.42**	**0.40**	0.76

3.4 Discriminant validity emerges with scale (RQ3)

Convergent validity is necessary but not sufficient: a measure must also *discriminate* the traits it claims to separate. Here the population splits sharply by scale (Table 4, Figure 1). In every tier, convergent validity is high and roughly constant ($r \approx 0.71$–0.75). What changes with scale is the heterotrait correlation, the degree to which *different* traits move together across models. For small local models (≤4B), heterotrait $r = 0.67$, almost as high as convergent r: the five "traits" are barely distinguishable, the hallmark of an undifferentiated, single-evaluative-dimension response. For large local models (>4B), heterotrait r falls to 0.51; for frontier-hosted models it falls to 0.27. The convergent−heterotrait gap, the standard one-number summary of discriminant validity, thus grows monotonically from 0.08 to 0.22 to 0.44.

Table 4: MTMM convergent and heterotrait correlations computed within model-size tiers. The gap (convergent − heterotrait) is the discriminant-validity summary.

Tier	n	Convergent r	Heterotrait r	Gap
Small-local (≤4B)	12	0.745	0.667	0.078
Large-local (>4B)	17	0.731	0.507	0.224
Frontier-hosted	13	0.707	0.265	0.442
All models	42	0.820	0.582	0.238

The continuous analysis within the local models (where parameter counts are public) agrees. Per-model trait differentiation rises with size: regressing differentiation on $\log_{10}$ parameters over the 29 local models gives slope 0.15, $r = 0.44$, $p = 0.018$ (Figure 2); the per-model inter-instrument profile agreement shows the same upward trend ($r = 0.47$, $p = 0.011$). Smaller models compress the five traits toward a common value; larger models spread them apart.

The full MTMM matrix (Appendix Figure 3) makes the population-level picture visible: across all 42 models the matrix is uniformly warm (convergent *and* heterotrait correlations are positive and sizable), which is exactly the "everything correlates" signature that the tiered analysis localizes to the smaller models.

3.5 The trait structure is a population property, not an individual one (RQ4)

The convergent, internally consistent structure documented above is a fact about the *population* of models. It does not hold within a single model. Recomputing α with one model's 11 repeated administrations as the respondent sample (asking whether a model reproduces a coherent trait structure across regenerations) yields a median α of essentially zero (−0.01 across 840 model×instrument×domain cells; 393 negative; only 2 reaching 0.70). This is not a low-variance

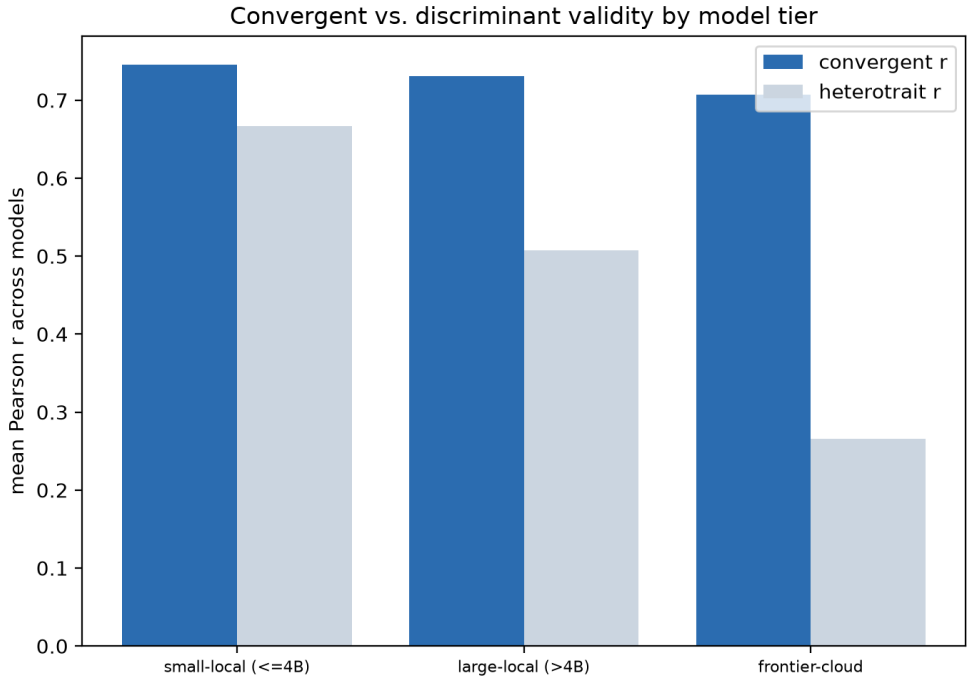

Figure 1: Convergent validity is roughly constant across model tiers; discriminant validity (the convergent−heterotrait gap) grows with scale, driven by falling heterotrait correlations.

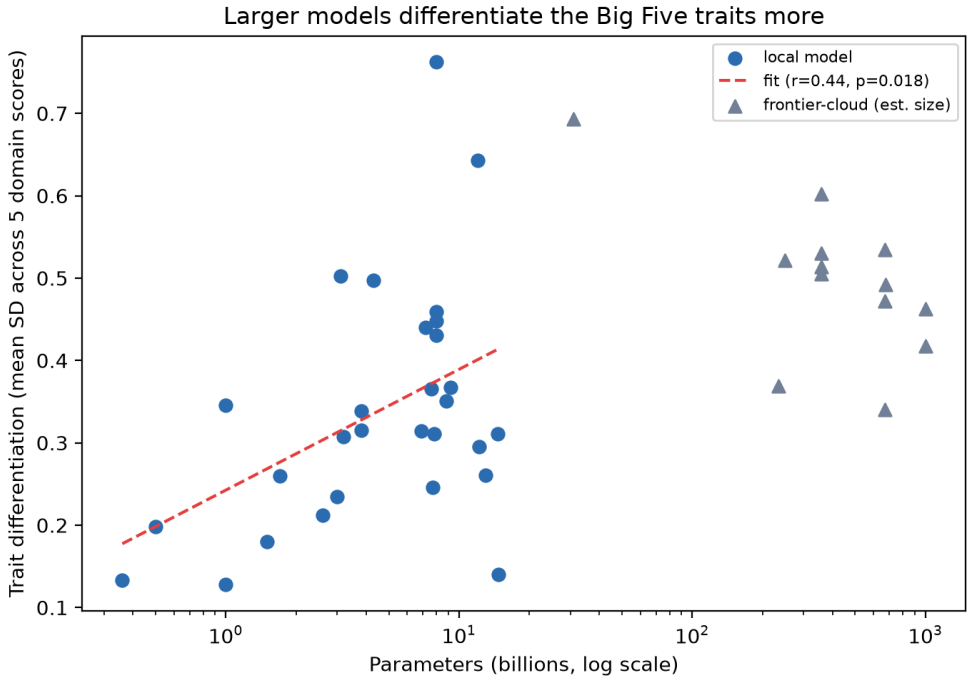

Figure 2: Trait differentiation (mean SD across the five domain scores) versus parameters. Local models (circles) define the fitted trend; frontier-hosted models (triangles, estimated sizes) extend it.

artifact: at temperature 0.7 the models do vary across regenerations (mean across-repetition item SD = 0.38 on the 1–5 scale; 58% of items take more than one value across the 11 repetitions). The variance is simply incoherent: a model's regeneration-to-regeneration fluctuation around its mean does not behave like a latent trait. A single model, asked the same questionnaire eleven times, does not answer like a person with a stable Big Five profile; the stable profile appears only when models are compared to one another.

3.6 Not a quantization artifact

Re-running seven families at full precision reproduces every pattern (Table 5). Quantization moves domain scores only slightly: the mean absolute fp16−q4 shift is 0.10 on the 1–5 scale (max 0.49), with 10 of 35 family×domain cells reaching $|d| \geq 0.5$ (the standardized effects are non-trivial only because within-configuration variance is small, as in paper [9]). Greedy answers agree between precisions on 82.5% of items on average (range 70–94%). Crucially, reliability is unchanged: population α at fp16 matches q4, and the within-model α is identically near zero at both precisions (0 of 140 cells ≥ 0.70 at either). The convergent-validity, scale-dependent-discriminant, and population-vs-individual findings are properties of these models, not of the 4-bit format they are usually deployed in.

Table 5: Precision spot-check (fp16 vs. q4/native), seven families. Score shift is mean $|\Delta|$ over five domains on the 1–5 scale; greedy agreement is exact item match under greedy decoding.

Family	Mean $\|\Delta\|$ score	Greedy agreement
Gemma 2 2B	0.07	87.9%
Llama 3.2 1B	0.01	78.6%
Llama 3.2 3B	0.09	75.2%
Qwen 2.5 3B	0.17	89.7%
Phi-3.5-mini	0.08	82.8%
Mistral 7B	0.04	93.8%
Qwen 2.5 7B	0.22	69.7%
Overall	**0.10** (max 0.49)	**82.5%**

4 Discussion

4.1 Implications

The picture is more constructive than the stability and validity critiques alone would suggest, and more demanding of model scale. (1) **Big Five questionnaires are valid instruments for measuring *between-model* differences.** Different inventories converge on a model's trait standing ($r = 0.82$, 0.85 net of shared items), and the long inventories are internally consistent across the model population (α median 0.81). Researchers who use these instruments to compare or rank models are measuring a real, reproducible signal. (2) **Do not use abbreviated inventories for this purpose.** The 20-item Mini-IPIP, with four items per domain, is unreliable across models in four of five domains; short forms validated on humans do not transfer to LLM populations, and Extraversion is especially fragile across every instrument. (3) **Discriminant validity is emergent with scale.** Small models do not separate the five traits (their questionnaire responses collapse toward a single evaluative dimension), while frontier models recover the five-factor structure. "The

Big Five personality of" a sub-4B model is largely one number wearing five labels; the five-trait description earns its degrees of freedom only at scale. This is a construct-validity counterpart to PERSIST's stability result: scale does not buy *stability* [13], but it does buy *differentiation*. (4) **A model does not have a personality the way a person does.** The five-factor structure is a property of the population of models, not of any single model across its own regenerations; treating one model's questionnaire output as an individual trait profile is not supported by its within-model reliability.

4.2 Relation to prior work

Where PERSIST [13] and the validity-critique line [12, 7, 11] ask whether LLM personality scores are stable and behaviorally meaningful, we ask the prior question of construct validity using the classical MTMM tool [1], and we find that the answer is scale-dependent rather than uniformly negative. Our first paper [9] reported that a single instrument failed internal consistency for three small models at every quantization level; the present, larger study refines that: the failure is specific to the within-model and small-model regimes, while the population of full-instrument, larger models is both convergent and internally consistent. The two results are consistent (the within-model $\alpha \approx 0$ here reproduces paper [9]'s finding on 42 models and at full precision), but the construct-validity lens shows what within-model reliability alone could not: the instruments work, as between-model measures, for sufficiently large models.

4.3 Limitations

First, the unit of analysis is the model, so all population statistics describe a particular 42-model sample; a differently composed population (e.g., one family scaled cleanly) could shift the tier boundaries. Second, the frontier tier is heterogeneous: hosted models run at provider-side precision, several lack published parameter counts (hence their exclusion from the continuous fit), and reasoning was suppressed to keep administration comparable, each a covariate of "large model as deployed" rather than of scale per se. Third, all four instruments draw on the IPIP bank and share some items; we mitigate this with BFAS-anchored estimates but cannot fully separate method from item bank. Fourth, the item-recognition confound [11] (models may recognize verbatim IPIP items from pretraining) applies to all four instruments, and larger models, which recognize more, are also the ones that discriminate better; we cannot rule out that part of the scale effect is increased item recognition rather than richer trait representation. Fifth, administration is English-only, single-temperature (0.7) for sampled repetitions, and one inference stack (Ollama/llama.cpp). Sixth, "Emotional Stability" is keyed as the positive pole of Neuroticism; comparisons to instruments scoring Neuroticism directly must flip sign.

4.4 Future work

Two extensions follow directly. A reasoning-mode arm administers the same instruments to mandatory-reasoning models (and to hybrid models with reasoning enabled), asking whether chain-of-thought before answering changes convergent and discriminant validity, a construct-validity counterpart to PERSIST's reasoning result. And a within-family scaling study, holding architecture and training fixed while varying only size, would convert the cross-population tier effect reported here into a clean dose-response curve for the emergence of trait discrimination.

5 Reproducibility statement

The repository contains the four instruments (`data/`), administration code (`src/llmpsy/`), analysis code (`analysis/`: `mtmm.py`, `convergent_report.py`, `population_alpha.py`, `size_trend.py`, `precision_check.py`), the grid drivers (`scripts/`), and a test suite. All randomness is seeded deterministically (per-call sha256-derived seeds; bootstrap seed 0). Raw responses are stored in SQLite with idempotent keys; the analysis scripts open the databases read-only and regenerate every table and figure here. Model artifacts are pinned by Ollama digest (Appendix A) and prompts by `PROMPT_VERSION = "v1"`. The complete dataset (133,980 main-grid plus 22,330 precision-check administrations) and all code are permanently archived at doi:10.5281/zenodo.20835204.

Author note

Funding. This research received no external funding; all local computation was performed on the author's own hardware, and hosted-model inference used the author's own account.

Conflicts of interest. The author declares no conflicts of interest.

Ethics. This study involved no human or animal subjects (the respondents are language models), and institutional review was therefore not applicable.

CRediT statement. Trevor Johnson: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing.

AI-assistance disclosure. The administration and analysis software, statistical computations, and manuscript drafting were produced with substantial assistance from an AI system (Claude, Anthropic) operating under the author's direction; the author reviewed all code, analyses, and text and takes full responsibility for the content.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] R. Vallat. Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31):1026, 2018.

[15] 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.

A Models

The 42 models, with Ollama digests and (for local models) public parameter counts and (for hosted models) estimated sizes used only for tier ordering, are listed in `data/models.yaml` and `results/model_digests.txt` in the repository. Local tier (29 models, q4_K_M or published default): SmolLM2 360M; Qwen 2.5 0.5B/1.5B/3B/7B/14B; Gemma 3 1B/4B/12B; Llama 3.2 1B/3B; Llama 3.1 8B; Gemma 2 2B/9B; SmolLM2 1.7B; Granite 3.1 MoE 3B / Dense 8B; Phi-3.5-mini 3.8B; Phi-4-mini 3.8B; Phi-4 14B; Mistral 7B; Mistral-Nemo 12B; Ministral 8B; DeepSeek-LLM 7B; InternLM2 7B; Yi 9B; Aya-Expanse 8B; EXAONE 3.5 7.8B; OLMo 2 13B. Frontier-hosted tier (13 models): Gemma4 31B; Qwen3.5; Nemotron-3-Super; GLM-4.7/5/5.1/5.2; DeepSeek-V3.2 / V4-Flash / V4-Pro; Mistral-Large-3 675B; Kimi-K2.5 / K2.6.

B Full MTMM matrix

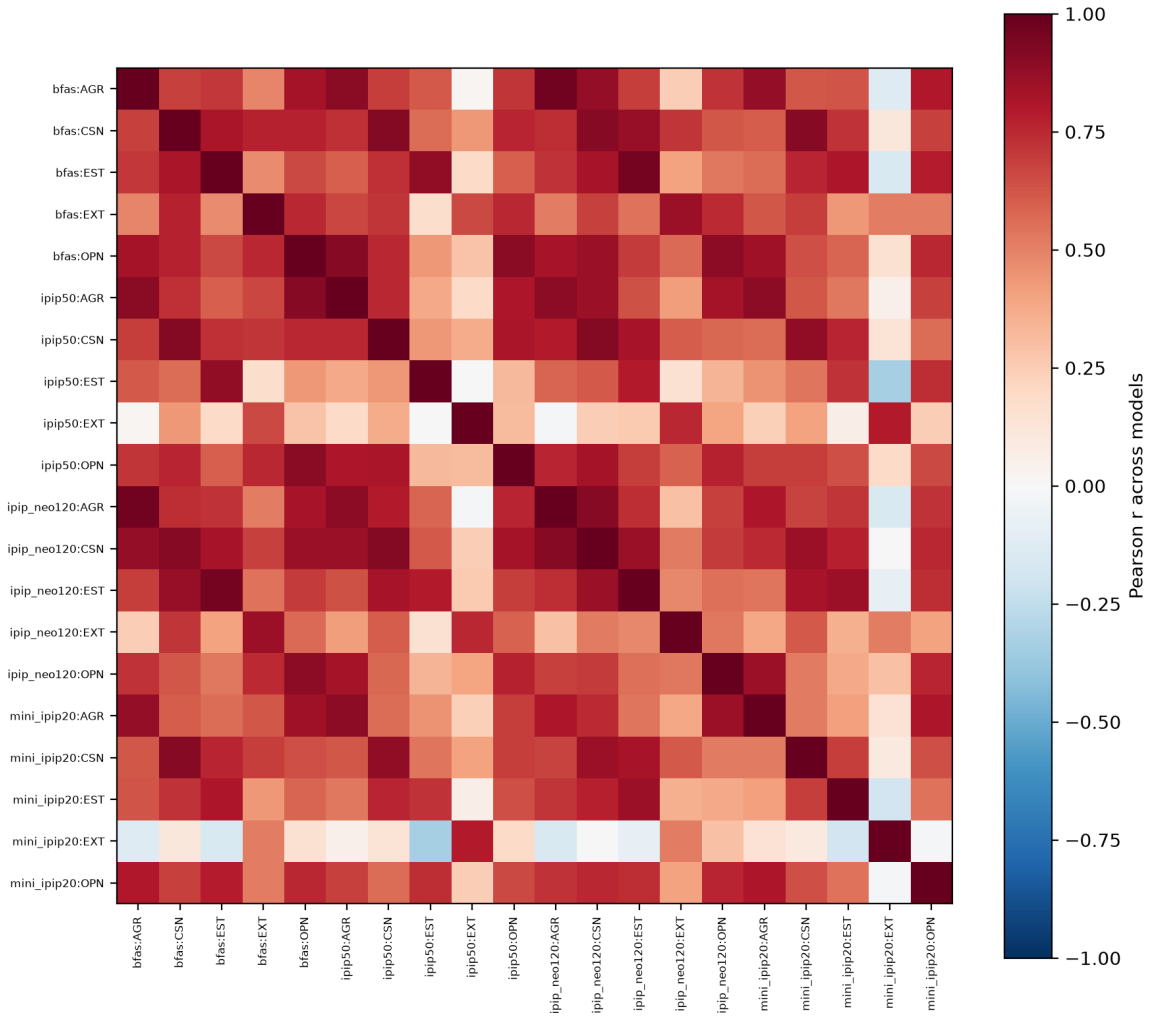

Figure 3: Full MTMM matrix: Pearson correlations across the 42 models among all 20 instrument×domain columns. The uniformly warm matrix is the population-level "everything correlates" pattern that the tiered analysis (Table 4) localizes to the smaller models; the Mini-IPIP Extraversion band is the conspicuous low-reliability exception.