Is LLM Personality an Artifact of Deployment? Psychometric Stability of Big Five Self-Reports Across Quantization Levels

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

June 2026

Abstract

Psychometric questionnaires are now routinely administered to large language models (LLMs), and a growing critical literature questions what such instruments measure when the respondent is a model. One deployment-side moderator has so far gone unexamined: quantization. Virtually every real-world local deployment of an open-weight model runs a quantized artifact (typically a 4- or 8-bit GGUF served by Ollama or llama.cpp), yet published "LLM personality" estimates are almost always obtained from full-precision checkpoints. We administered the 50-item IPIP Big Five Factor Markers to nine model variants—three open-weight instruction-tuned models (Llama 3.1 8B, Qwen 2.5 7B, Llama 3.2 3B), each at three quantization levels (q4_K_M, q8_0, fp16)—under three response-option presentation conditions, with one greedy and twenty independently seeded sampled repetitions per item: 28,350 stateless, JSON-schema-constrained chat calls on fixed consumer hardware. Three findings emerge. First, 4-bit quantization materially shifted domain scores relative to the same checkpoint at fp16 (6 of 15 baseline contrasts with bootstrap CIs excluding zero; shifts up to 0.42 scale points, $|d|$ up to 3.5), in directions idiosyncratic to model family; 8-bit quantization was largely score-preserving (2 of 15; max $|d| = 0.76$). Under greedy decoding, q4 variants gave a different answer than their fp16 parent on 8–32% of items (q8: 2–12%). Second, internal consistency was inadequate everywhere: across all 135 variant $\times$ condition $\times$ domain cells, Cronbach's alpha never reached the conventional 0.70 threshold (range -2.07 to 0.47, median -0.10; 56% of cells negative), with no systematic precision gradient—the instrument's keyed structure failed to organize responses at every quantization level, including fp16. Third, response-option presentation alone (display order; letter vs. numeric labels) shifted scores more on average than quantization did (mean $|\Delta|$ up to 0.21 vs. 0.17 scale points; $|d|$ up to 4.2), at every precision. Together these results indicate that "the personality of model X" is underdetermined without the deployment configuration—and that for small open-weight models served in deployment-realistic conditions, questionnaire-based personality measurement fails basic psychometric standards regardless of precision. Quantization level, inference stack, sampler settings, and presentation format belong in the methods section of any LLM psychometrics study.

1 Introduction

LLM psychometrics has consolidated rapidly into a recognizable field, with systematic reviews now organizing its instruments, validation strategies, and open problems [9]. At the same time, the

field's measurement foundations remain contested. Sühr et al. [6] showed that the BFI-2 fails tests of measurement invariance between human and LLM respondents, meaning the same instrument does not measure the same construct in the two populations. The "Personality Illusion" line of work [4] found that LLM questionnaire self-reports dissociate from the models' actual behavior in downstream tasks. And questionnaire-based trait profiles diverge from profiles recovered from generation probabilities over the same items [5], suggesting that the administration format itself partly determines the measured "personality."

Within this critical line, PERSIST [7] is the most comprehensive stability study to date: across 25 models spanning 1B to 685B parameters, it documented persistent instability of personality scores under question reordering, paraphrasing, persona assignment, reasoning mode, and conversation history—reordering alone shifted scores by roughly 20% of the scale range—and found that model scale does not rescue stability. Two gaps remain, both acknowledged in that work. First, PERSIST evaluated full-precision models only. Second, it reported no formal reliability statistics (such as internal-consistency coefficients), a limitation its authors state explicitly.

This paper addresses the deployment gap directly. Open-weight models are overwhelmingly run, in practice, as quantized artifacts: weights compressed from 16-bit floating point to 4- or 8-bit representations (e.g., the GGUF q4_K_M and q8_0 formats used by llama.cpp and Ollama) so that they fit in consumer RAM. Quantization is a lossy transformation of the very parameters that produce questionnaire answers, applied *after* every property reported in a model card or research paper was measured. If the personality profile measured on a checkpoint does not survive quantization of that same checkpoint, then deployment configuration is an unexamined moderator of every published LLM personality result—and a study reporting "the personality of model X" without reporting quantization and sampler configuration may simply not replicate. No study to date has asked this question.

We therefore administer a classic Big Five instrument across a 3 (model family) $\times$ 3 (quantization level) grid, holding the checkpoint, serving stack, hardware, prompts, and seeds fixed, and ask:

- **RQ1 (H1).** Do Big Five domain scores shift across quantization levels of the same checkpoint?
- **RQ2 (H2).** Does internal consistency—Cronbach's alpha computed on reverse-keyed item scores—change across quantization levels?
- **RQ3 (H3).** Does sensitivity to response-option presentation (display order; numeric vs. letter labels) interact with quantization?

We additionally run a determinism probe: the exact-agreement rate of greedy-decoded (temperature 0) answers between quantization levels of the same checkpoint, which bounds how much of any score shift is attributable to changed argmax behavior rather than changed sampling distributions.

The question matters in both directions. If measurement is quantization-invariant, the field gains a useful robustness result—though configuration reporting would still be warranted. If it is not, then methods papers must treat quantization and sampler settings as first-class experimental parameters, and practitioners persona-tuning or "personality-auditing" q4 local models are reading tea leaves unless their measurement protocol is validated at the precision they deploy.

2 Methods

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

2.1 Instrument

We use the 50-item IPIP Big Five Factor Markers [2], a public-domain instrument [3] with ten items per domain: Extraversion (EXT), Emotional Stability (EST), Agreeableness (AGR), Conscientiousness (CSN), and Intellect/Openness (OPN). Items are short first-person statements (e.g., "Am the life of the party.") rated on a 5-point accuracy scale (1 = Very Inaccurate … 5 = Very Accurate). Standard reverse-keying applies: for negatively keyed items, the scored value is 6 minus the answer. The full instrument, including per-item domain and key, is versioned in `data/ipip50.yaml` (see Appendix A).

2.2 Subjects and materials

The "subjects" are nine model variants: three open-weight instruction-tuned checkpoints—Llama 3.1 8B Instruct, Qwen 2.5 7B Instruct, and Llama 3.2 3B Instruct—each served at three quantization levels: q4_K_M (4-bit, k-quant medium), q8_0 (8-bit), and fp16 (16-bit floating point, the unquantized reference for this study). For readers outside machine learning: quantization stores each model weight in fewer bits, trading a small, structured amount of numerical precision for a several-fold reduction in memory; q4_K_M is the de facto default for local deployment.

All variants are served by Ollama (a llama.cpp-based local inference server) on a single fixed consumer machine: AMD Ryzen 7 CPU (integrated Radeon 780M graphics), 32 GB RAM, Ubuntu Linux. We regard the consumer-hardware, Ollama-served setting as a feature rather than a compromise: it is ecologically valid for exactly the deployment population whose measurement validity is in question. The exact model artifacts are pinned by their Ollama digests (Table 1).

Table 1: Model variants and Ollama digests.

Variant tag	Digest	Size
llama3.1:8b-instruct-q4_K_M	46e0c10c039e	4.9 GB
llama3.1:8b-instruct-q8_0	b158ded76fa0	8.5 GB
llama3.1:8b-instruct-fp16	4aacac419454	16 GB
qwen2.5:7b-instruct-q4_K_M	845dbda0ea48	4.7 GB
qwen2.5:7b-instruct-q8_0	2d9500c94841	8.1 GB
qwen2.5:7b-instruct-fp16	59805ce4a404	15 GB
llama3.2:3b-instruct-q4_K_M	a80c4f17acd5	2.0 GB
llama3.2:3b-instruct-q8_0	e410b836fe61	3.4 GB
llama3.2:3b-instruct-fp16	195a8c01d91e	6.4 GB

2.3 Administration protocol

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).

Presentation conditions. Each item is administered under three conditions: *baseline* (options displayed 1→5), *reversed* (options displayed 5→1; only the display order changes—the canonical mapping 1 = Very Inaccurate … 5 = Very Accurate is fixed), and *letters* (options labeled A–E in ascending order, mapped back to 1–5 at parse time).

Constrained decoding. Every call uses Ollama's structured-output (JSON-schema-constrained) decoding, with the answer field restricted to an enum of the legal option tokens for the condition, so every response is a valid option token by construction; no free-text answer coding is required.

System prompt. The system prompt is task framing only, with no persona. Verbatim (from `src/llmpsy/prompts.py`, 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 and seeding. For each (variant, condition, item) cell, repetition 0 is greedy (temperature 0) and repetitions 1–20 are sampled at temperature 0.7. Each sampled call receives an independent seed derived as the first four bytes of sha256(item_id | condition | rep). This per-call seeding avoids a common-random-numbers artifact: if all items within a repetition shared one seed, sampling noise would be correlated across items within that pseudo-respondent, mechanically inflating inter-item correlations—precisely the quantity our reliability analysis measures. (When a structurally malformed answer triggers a parse retry, the seed is deterministically bumped and the seed that actually produced the stored answer is persisted for provenance.)

Volume. 50 items × 3 conditions × 21 repetitions = 3,150 calls per variant; 28,350 calls in total. All 28,350 calls completed and parsed successfully (the schema-constrained decoder produced a legal option token on every call; no rows were excluded).

2.4 Analysis

Internal consistency. For each (variant, condition, domain) we compute Cronbach's alpha [1] with a 95% confidence interval (pingouin [8]) on the repetitions × items matrix of reverse-keyed scored values, using sampled repetitions only (rep ≥ 1) and dropping incomplete repetitions. For readers outside psychometrics: alpha summarizes how consistently a set of items presumed to measure one construct covary across respondents; values near 1 indicate the items move together, values near 0 indicate they do not, and negative values typically indicate response sets (e.g., acquiescence) overpowering the keyed structure. We frame alpha here explicitly as *generation consistency over pseudo-respondent repetitions of a single model*—not as evidence about population trait structure, which would require a population of respondents.

Score stability. Mean and SD of domain scores (mean scored value over the domain's 10 items) across sampled repetitions, per (variant, condition, domain); only repetitions with complete item sets enter.

Quantization effects (RQ1). For each family, domain, and comparison (q4 vs. fp16, q8 vs. fp16), restricted to the baseline condition: the difference in mean domain score, with a 5,000-draw percentile bootstrap 95% CI and Cohen's d using the $(n-1)$-weighted pooled SD.

Condition effects (RQ3). The same estimator applied to reversed vs. baseline and letters vs. baseline within each variant, allowing condition sensitivity to be compared across quantization levels of a family.

Determinism probe. Percentage of items on which the greedy (rep 0) answers of two quantization levels of the same family agree exactly, per condition.

Inferential framing. All effects are reported as descriptive estimates with 95% CIs. We make no null-hypothesis "significance" claims: the grid yields roughly 120 comparisons (3 families $\times$ 2 quant comparisons $\times$ 5 domains, plus 9 variants $\times$ 2 condition comparisons $\times$ 5 domains), so about 6 CIs would be expected to exclude zero by chance alone at the 95% level even under global invariance. We interpret patterns—consistency across families and domains, dose-response from q8 to q4—not isolated intervals.

2.5 Reproducibility

All code, seeds, prompts, and instrument data are public. Responses are stored in SQLite with idempotent primary keys hashed from (model, item, condition, rep, prompt_version), so interrupted runs resume without duplication and prompt-wording changes cannot silently mix data; `prompt_version` is pinned at `v1` for all data in this paper. Model artifacts are pinned by the digests in Table 1.

3 Results

Complete per-cell tables are emitted by the analysis scripts as CSVs in `results/` (`alpha.csv`, `stability.csv`, `quant_effects.csv`, `condition_effects.csv`); this section summarizes them.

3.1 Internal consistency (RQ2)

Internal consistency was inadequate in every cell of the design. Across all 135 variant $\times$ condition $\times$ domain cells, Cronbach's alpha ranged from -2.07 (Llama 3.1 8B fp16, reversed condition, EST; 95% CI $[-4.56, -0.41]$) to 0.47 (Llama 3.2 3B q8_0, reversed condition, AGR; CI $[0.04, 0.76]$), with a median of -0.10. **No cell reached the conventional 0.70 adequacy threshold, and 75 of 135 cells (56%) were negative.** Negative alpha on reverse-keyed scores with near-uniform raw responding is the signature of a response set—answers tracking option position or a global agreement tendency rather than item content.

Critically for RQ2, low reliability is not something quantization inflicts on an otherwise-coherent instrument: the fp16 reference variants themselves produced deeply negative alphas (e.g., Llama 3.1 8B fp16 averaged -0.44 across baseline domains), and no monotone precision gradient is discernible—within a family, the sign and magnitude of alpha jump erratically between q4, q8, and fp16 and between presentation conditions. The instrument's keyed structure fails to organize these models' responses at every precision; what quantization changes is *which way* it fails (Section 3.4).

3.2 Score stability across repetitions

Between-repetition variability of domain scores was substantial relative to the effects of interest: the median SD across the 135 cells was 0.23 scale points (range 0.00–0.42; maximum: Llama 3.1 8B

q4_K_M, baseline, EXT, mean 2.96, SD 0.42). With 20 sampled repetitions, the standard error of a cell's mean domain score is roughly 0.05 scale points, so the cross-configuration differences reported below (0.2–0.4 points) are large relative to sampling noise—but a *single* administration of the questionnaire (the modal design in applied work) carries an expected error of a quarter scale point before any configuration effect is considered.

3.3 Quantization effects on domain scores (RQ1)

Table 2: Baseline-condition domain-score shifts vs. fp16 with bootstrap 95% CIs excluding zero (8 of 30 contrasts; full table in `results/quant_effects.csv`).

Family	Domain · Comparison	Δ	95% CI	Cohen's d
llama3.1-8b	CSN · q4 vs fp16	+0.36	[0.17, 0.55]	1.17
llama3.1-8b	OPN · q4 vs fp16	+0.42	[0.20, 0.64]	1.14
qwen2.5-7b	AGR · q4 vs fp16	−0.34	[−0.40, −0.28]	−3.50
qwen2.5-7b	CSN · q4 vs fp16	−0.26	[−0.32, −0.19]	−2.36
qwen2.5-7b	EST · q4 vs fp16	−0.32	[−0.38, −0.25]	−2.96
llama3.2-3b	OPN · q4 vs fp16	−0.28	[−0.44, −0.12]	−1.03
qwen2.5-7b	AGR · q8 vs fp16	+0.06	[0.00, 0.12]	0.63
qwen2.5-7b	CSN · q8 vs fp16	−0.09	[−0.16, −0.02]	−0.76

Three patterns stand out (Table 2, Figure 1). First, a clear **dose-response**: q4 shifted scores in 6 of 15 baseline contrasts (mean $|\Delta| = 0.17$ scale points) versus 2 of 15 for q8 (mean $|\Delta| = 0.05$)—against roughly 0.75 contrasts expected by chance per comparison set. Every family showed at least one credible q4 shift; q8 was close to score-preserving everywhere. Second, the directions are **family-idiosyncratic**: 4-bit quantization made Llama 3.1 8B describe itself as *more* conscientious and open (+0.36, +0.42) while making Qwen 2.5 7B *less* agreeable, conscientious, and emotionally stable (−0.26 to −0.34) and Llama 3.2 3B less open (−0.28). Quantization does not nudge all models in a common direction; it perturbs each checkpoint's response tendencies in its own way, which is exactly the behavior that makes it impossible to correct for post hoc. Third, the standardized effects are large ($|d|$ 1.0–3.5 for credible q4 shifts) because between-repetition variance within a configuration is small relative to the shift—the configurations are *internally* consistent enough to be reliably different from each other.

3.4 Presentation-condition sensitivity × quantization (RQ3)

Presentation effects were pervasive and, on average, larger than quantization effects: 53 of 90 condition contrasts excluded zero (reversed vs. baseline mean $|\Delta| = 0.21$ scale points; letters vs. baseline 0.16; versus 0.17 for q4-vs-fp16). The largest single effect in the study was presentational: merely listing the options 5→1 instead of 1→5 raised Llama 3.2 3B fp16's Emotional Stability score by 0.73 scale points ($d = 2.78$), and relabeling options A–E moved Qwen 2.5 7B q8_0's EST by −0.38 ($d = -4.17$).

On RQ3's specific question—whether presentation sensitivity *grows* as precision falls—the answer is no: we found no monotone interaction. The largest presentation effects occurred at fp16 and q8, not q4. Presentation sensitivity is better described as large, ubiquitous, and configuration-idiosyncratic: each (checkpoint, precision) pair has its own profile of format-induced distortions. This replicates PERSIST's order-effect findings under stricter conditions (stateless single-item calls

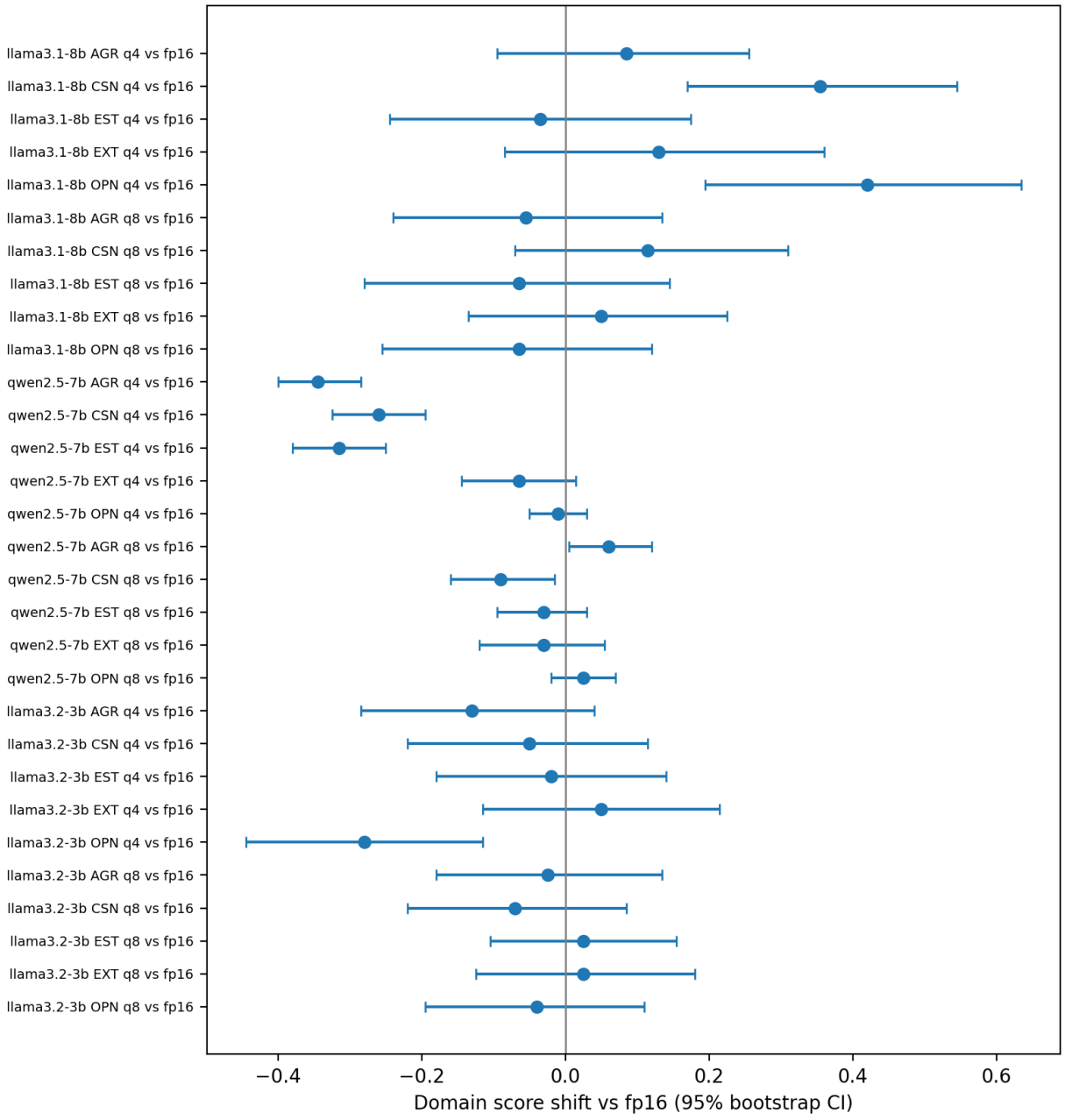

Figure 1: Domain-score shifts vs. fp16 (baseline condition) with 95% bootstrap CIs, per family, domain, and quantization comparison.

with schema-constrained decoding, where no conversational or formatting degrees of freedom remain) and extends them: even the *response options' visual arrangement*—with item content, scale semantics, and decoding constraints held fixed—moves scores by amounts comparable to a full quantization step.

3.5 Greedy determinism across quantization levels

Table 3: Exact-agreement of greedy (temperature 0) answers with the same family's fp16 variant (% of 50 items).

Family	Condition	q4 vs fp16	q8 vs fp16
llama3.2-3b	baseline	72%	88%
llama3.2-3b	reversed	76%	98%
llama3.2-3b	letters	86%	96%
qwen2.5-7b	baseline	68%	90%
qwen2.5-7b	reversed	74%	94%
qwen2.5-7b	letters	92%	96%
llama3.1-8b	baseline	76%	96%
llama3.1-8b	reversed	74%	94%
llama3.1-8b	letters	78%	94%

Under greedy decoding—the setting practitioners reach for when they want "the model's answer"—the 4-bit variant disagreed with its own fp16 parent on 8–32% of items (median across the nine cells: 24%), and the 8-bit variant on 2–12% (Table 3). Quantization changes the argmax answer to roughly one personality item in four at q4; the sampled-score shifts above are therefore not a tail-probability artifact but reflect changed modal behavior.

4 Discussion

4.1 Implications

The results land on the sharper side of the design's two branches. Domain scores did shift materially under 4-bit quantization, in family-specific directions, with greedy answers flipping on a quarter of items; meanwhile the measurement instrument itself never achieved adequate internal consistency in any configuration, and the visual arrangement of response options moved scores as much as or more than quantization did. Three implications follow. (1) **"The personality of model X" is underdetermined without the deployment configuration.** A profile measured at fp16 does not characterize the q4 artifact most users actually run; replication attempts across inference stacks or precisions are not replications of the same measurement. (2) **Configuration reporting is the cheap, immediate fix**: quantization format, inference stack and version, sampler settings, temperature, and option-presentation format belong in the methods section of every LLM psychometrics study, exactly as administration conditions do in human testing. q8_0 appears close to score-preserving in our grid and may be an acceptable stand-in for full precision; q4_K_M is not. (3) **For small open-weight models, the deeper problem is that the instrument fails before quantization is even considered.** With zero of 135 cells reaching $\alpha = 0.70$ and most cells negative, domain means computed from these questionnaires do not behave like trait scores under any configuration we tested; persona-auditing or "AI personality" claims built

on single-administration questionnaires of small local models are unsupported at the measurement level.

4.2 Relation to prior work

Relative to PERSIST [7], we add the deployment axis it did not test and the formal reliability statistics it did not report, on a deliberately narrower model grid. Our alpha-over-repetitions framing also speaks to the validity-critique line [6, 5]: with internal consistency low or negative across 135 cells—strongly negative under reverse-display conditions in several variants—the instrument's keyed structure is not organizing the model's responses, and interpreting domain means as trait scores is unwarranted for these configurations. The negative-alpha signature is consistent with acquiescent or position-driven response sets dominating item content. We emphasize that this is a finding about the fragility of questionnaire administration to LLMs, not an artifact to be corrected away: a respondent whose answers track option position rather than item content does not have a measurable Big Five profile under that protocol. Our results thereby furnish the formal psychometric grounding for PERSIST's instability observations—and locate a new source of instability (precision of the deployed weights) that survives even our strictest administration controls.

4.3 Limitations

First, our alpha is generation consistency over pseudo-respondent repetitions of one model; it is not interchangeable with population-sample alpha, and we make no claims about latent trait structure. Second, all results are specific to one inference stack (Ollama/llama.cpp and its GGUF k-quant formats); quantization effects may differ under vLLM, ExLlama, AWQ, or GPTQ pipelines. Third, the grid spans three model families at $\leq$ 8B parameters; larger models may behave differently, though PERSIST gives no reason to expect scale to confer stability. Fourth, we use a single instrument (IPIP-50); the item-wording recognition confound—models may recognize verbatim questionnaire items from pretraining (cf. [5])—applies to it in full. Fifth, administration is English-only. Sixth, sampled repetitions use a single temperature (0.7); temperature itself is another deployment parameter, examined here only at its greedy limit via the determinism probe.

4.4 Future work

Phase B will extend the same 9-variant grid to behavioral tasks in the Personality Illusion paradigm [4], asking whether the self-report/behavior dissociation is itself moderated by quantization—that is, whether deployment configuration changes not only what models say about themselves but whether those statements predict anything.

5 Reproducibility statement

The repository contains the instrument (`data/ipip50.yaml`), administration code (`src/llmpsy/`), analysis code (`analysis/`), the grid driver (`scripts/run_grid.sh`), and a 52-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 database read-only and regenerate every table and figure in this section from it. Model artifacts are pinned by Ollama digest (Table 1) and prompts by `PROMPT_VERSION = "v1"`. The complete raw dataset (28,350 rows) ships as `results/runs.csv`.

Author note

Funding. This research received no external funding; all computation was performed on the author's own hardware.

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

Ethics. This study involved no human or animal subjects — the respondents are publicly available open-weight 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] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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 Instrument

The complete IPIP-50 Big Five Factor Markers—all 50 items with domain assignment (EXT/EST/AGR/CSN/OPN keying direction, and the 5-point accuracy scale—are versioned in machine-readable form in `data/ipip50.yaml` in the repository, which is the exact file loaded at administration time (`src/llmpsy/instrument.py` validates the 50-item count at load). The instrument is public domain [2, 3]. Representative items: "Am the life of the party." (EXT, +); "Get stressed out easily." (EST, −); "Sympathize with others' feelings." (AGR, +); "Leave my belongings around." (CSN, −); "Have a vivid imagination." (OPN, +).