Does Reasoning Give a Language Model a Personality? Within-Model Effects of Thinking on Big Five Trait Scores and Construct Validity

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

June 2026

Abstract

Large language models that "reason" before answering, producing a chain of intermediate tokens, are now common, and many expose this reasoning as a switch that can be turned on or off. We ask whether turning reasoning on changes how a model responds to Big Five personality questionnaires, using a within-model design: the same model answers the same items with thinking off and with thinking on. We administered four public-domain IPIP Big Five inventories (the 20-item Mini-IPIP, the 50-item IPIP Factor Markers, the 100-item Big Five Aspect Scales, and the 120-item IPIP-NEO; 290 items in total) to ten open-weight hybrid models, one reasoning-native open model (OLMo 3, which cannot disable thinking), and one commercial model (Claude Haiku 4.5, administered through its provider's Batch API). The think-off baseline for the ten hybrids reuses our prior study's direct-answer data; we validate that reuse directly, showing that a fresh think-off run reproduces the prior domain means to within 0.07 points on the five-point scale, against reasoning effects 5 to 28 times larger. Two findings emerge. First, **reasoning shifts reported trait scores substantially and in a consistent direction**: models present as more emotionally stable (about +0.9 on the 5-point scale) and less extraverted (about −0.6 to −0.8). The two largest effects hold across open-weight and commercial models and vary in magnitude across models; the smaller conscientiousness and agreeableness effects do not generalize and are not claimed. Convergent validity rises under reasoning (mean convergent r 0.65 → 0.84) while heterotrait correlation is flat, so the convergent−discriminant gap widens because same-trait agreement rises, not because traits separate. Second, **reasoning does not create within-model coherence**: treating repeated generations as respondents, internal-consistency reliability stays near zero with thinking off and on (median $\alpha \approx 0$; 0 of 200 hybrid and 0 of 20 commercial model-by-domain cells reach 0.70), and the same holds for the reasoning-native model that cannot stop thinking. Reasoning changes what a model says about its personality without making that personality cohere as an individual-level property.

1 Introduction

A growing line of work administers human personality questionnaires to language models and reports Big Five scores, often treating the resulting numbers as a stable trait profile [11, 15]. The measurement foundations of that practice are contested. Self-reports dissociate from downstream behavior [7], questionnaire-recovered profiles diverge from those recovered from generation probabilities [12], and personality instruments fail measurement-invariance tests between human and model

respondents [13]. PERSIST [14], the most comprehensive stability study to date, documented that scores shift under reordering, paraphrasing, persona assignment, *reasoning mode*, and conversation history across models from 1B to 685B parameters. That a reasoning toggle perturbs scores is therefore known; what that perturbation does to construct validity, whether the shift is systematic and directional, and whether reasoning changes the deeper nature of model "personality," is not.

Our prior work approached this from two angles. The first paper [9] showed that for small open-weight models a Big Five instrument never achieved adequate within-model internal consistency at any quantization level. The second [10] widened the lens to a population of 42 models and four instruments, and found that the Big Five structure in models is a *population* property: convergent and discriminant validity strengthen with model scale across models, but within a single model the trait structure does not cohere, internal-consistency reliability across repeated generations sits near zero. Both studies administered questionnaires in a single, direct-answer condition.

The present paper adds the axis those studies could not reach: explicit reasoning. Reasoning is plausibly relevant to personality measurement for two reasons. First, deliberation could change self-report: a model that reasons about an item like "I get angry easily" before answering may regulate its answer differently than one that responds immediately. Second, and more fundamentally, reasoning is sometimes argued to give models more coherent, agent-like behavior, which raises the question of whether reasoning manufactures the individual-level trait coherence that direct answering lacked. We use a within-model manipulation of reasoning (thinking off versus thinking on) and ask:

- **RQ1 (Scores).** Does reasoning shift reported Big Five trait scores, and if so in what direction and how large?
- **RQ2 (Convergent validity).** Does reasoning change how strongly the same trait, measured by different instruments, agrees with itself?
- **RQ3 (Discriminant validity).** Does reasoning change how distinct different traits are from one another?
- **RQ4 (Within-model coherence).** Does reasoning create within-model internal-consistency reliability, the individual-level coherence that direct answering lacked?

Our headline findings are that reasoning produces a large, directionally consistent shift in self-reported scores (more emotionally stable, less extraverted), and that it does not create within-model coherence. The score shift is a change in self-presentation, not the emergence of a person.

2 Methods

All materials, code, and analysis scripts described here are in the accompanying repository.

2.1 Instruments

We use the same four public-domain IPIP Big Five inventories as our prior convergence study [10], chosen to vary in length and item provenance (Table 1): the Mini-IPIP [4] (20 items), the IPIP-50 Factor Markers [5] (50 items), the Big Five Aspect Scales (BFAS) [3] (100 items), and the IPIP-NEO-120 [8] (120 items): 290 items in total, mapping to Extraversion (EXT), Emotional Stability (EST), Agreeableness (AGR), Conscientiousness (CSN), and Openness (OPN). Items use a 5-point accuracy scale with standard reverse-keying (negatively keyed items scored as 6 minus

the answer). Item text, presentation, and scoring are identical to the prior study, so that the only intended difference from the reused baseline data is the reasoning toggle.

Table 1: The four Big Five instruments, transcribed from the International Personality Item Pool [6]; all 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

2.2 Subjects

We study three groups of models, chosen so the within-model think-on versus think-off question can be asked of architectures that differ in provenance (Appendix A).

Ten open-weight hybrid models (primary subjects). DeepSeek-V3.2, DeepSeek-V4-Flash, DeepSeek-V4-Pro, GLM-4.7, GLM-5, GLM-5.1, GLM-5.2, Kimi-K2.5, Kimi-K2.6, and Nemotron-3-Super, served through Ollama's hosted ("`:cloud`") tier. Each exposes a reasoning toggle.

One reasoning-native open model. OLMo 3 (7B), which cannot disable reasoning, run locally on a consumer GPU. It contributes only to RQ4 (having no think-off arm, it cannot enter the paired contrast) and serves as an anchor for whether a model that *must* reason shows within-model coherence.

One commercial model. Claude Haiku 4.5, administered through the Anthropic Batch API, as an independent check on whether the open-weight pattern generalizes to a closed model from a different developer.

One additional hybrid (qwen3.5) was excluded during collection (Section 2.6).

2.3 Administration protocol

Each item is presented in its own stateless, JSON-schema-constrained call with the same task-framing system prompt (prompt version `v1`) used in the prior studies: "You are completing a personality questionnaire about yourself… Respond with JSON only." For each (model, item) cell we collect a greedy generation (temperature 0) and several sampled generations (temperature 0.7), each independently seeded. The manipulation is the reasoning toggle: in think-off the model answers directly; in think-on it is instructed to reason first, with a large output-token budget (12,000 tokens) so that even items eliciting long deliberation can finish and still emit an answer.

Two model-imposed details are reported as covariates rather than hidden. First, the commercial model pins its sampling temperature to a default whenever reasoning is enabled, so *both* of its arms were collected at that default temperature; its within-collection contrast is temperature-matched between arms but not matched to the open-weight temperatures. Second, the reasoning-native model has no off arm by construction.

For the ten hybrids, the think-off arm **reuses the direct-answer data from our prior convergence study** [10]. This reuse is deliberate: it makes the contrast a same-protocol, within-model comparison in which only the reasoning toggle changes. Because the hosted models are served from an endpoint whose weights we do not control, we validate the reuse empirically (Section 2.4) rather than assume it.

2.4 Baseline reproducibility check

Reusing prior think-off data is valid only if think-off behavior is stable between collections. The think-off data were collected June 20–23, 2026 and the think-on data June 25–27, 2026 (a ∼5-day gap), with identical temperatures, seeds, prompt version, and items. We then re-administered the think-off condition on two representative hybrids (GLM-5 and DeepSeek-V4-Flash) and compared at two levels.

At the single-item greedy level, reproduction is imperfect for an instructive reason: one endpoint (GLM-5) is fully deterministic at temperature 0 and reproduces the prior greedy answers at ∼93% across domains, while the other (DeepSeek-V4-Flash) is a *nondeterministic* endpoint that agrees with its own immediate re-run only ∼73% of the time. For the nondeterministic model, agreement with the prior data is as high as its agreement with itself, so the disagreement reflects endpoint stochasticity, not drift.

At the level the analysis uses, the domain score (a mean over twenty items and many reps), the reuse is confirmed: fresh think-off domain means reproduce the prior think-off domain means to within 0.07 points on the five-point scale, usually within 0.02 (Appendix B). The item-level stochasticity averages out, which is the purpose of the repeated-rep design. For comparison, the reasoning effects below are 5 to 28 times larger than this reproduction error.

2.5 Analysis

Each generation is scored by reverse-keying and averaging items within a domain, yielding one domain score per (model, instrument, condition, rep). The greedy rep is excluded from score and reliability analyses (its zero variance is not meaningful as a respondent); only sampled reps are used.

Scores (RQ1). We compare think-on and think-off domain means within each model, reporting the **raw difference on the 5-point scale as the primary effect**. Standardized effect sizes are reported for completeness but de-emphasized: because the standardizing denominator is the small generation-to-generation variance over a handful of reps, standardized values can be very large and should not be read on a human effect-size scale.

Convergent and discriminant validity (RQ2, RQ3). Across models, we compute a multitrait-multimethod (Campbell-Fiske) matrix [1] with instruments as methods, and compare the mean convergent correlation (same trait, different instrument) and mean heterotrait correlation (different trait, different instrument) between conditions.

Within-model coherence (RQ4). For each (model, instrument, domain) we compute Cronbach's α [2] treating reps as respondents and items as indicators, separately for think-on and think-off. Negative α values are reported as-is (they reflect near-zero or degenerate inter-item covariance and are a substantive finding, not an error); we summarize with medians, robust to the heavy tails, and never with means.

2.6 Exclusion

One hybrid (qwen3.5) was dropped during collection. With reasoning on, it ruminated past even a 12,000-token budget on Emotional-Stability items at temperature 0, emitting no answer (a length-limited, empty completion). Because this is deterministic at temperature 0, retries cannot recover it. This was a **post-hoc, data-dependent exclusion**, and we report it as such, describing the failure qualitatively rather than including partial data. Notably, the same rumination-to-silence appeared on the same Emotional-Stability item in the commercial model at a small token budget (it resolved at the large budget), so the failure mode is not unique to one model or provider. Because these are missing answers rather than different answers, and they cluster on the Emotional-Stability items (the domain with the largest reasoning effect), partial inclusion would bias exactly that effect rather than inform it, and the missing cells cannot be imputed; the model is therefore excluded entirely.

2.7 Reproducibility

All instruments, administration code, the reasoning toggle and token-budget handling, the commercial-model Batch API adapter, the analysis modules, and the baseline-reproducibility scripts are deposited. Raw response databases and computed result tables are included so that every number can be regenerated. Responses are stored in SQLite with idempotent keys hashed from (model, instrument, item, condition, rep, prompt_version).

3 Results

3.1 Reasoning shifts trait scores, most strongly emotional stability and extraversion (RQ1)

Reasoning produces a large, directionally consistent change in reported scores. Across the ten hybrids the mean absolute shift is about half a scale point, and the large majority of model-by-domain cells move substantially (181 of 200 cells reach $|d| \geq 0.5$). By domain (Table 2), the two largest effects are an **increase in emotional stability** and a **decrease in extraversion**: a model that reasons reports being calmer and less outgoing than the same model answering directly.

Table 2: Mean score shift (think-on minus think-off) by domain, on the 1–5 scale. The ten hybrids reuse the prior study's think-off baseline; Claude Haiku 4.5's two arms were collected together.

Domain	Hybrids (10 models)	Claude Haiku 4.5
Emotional Stability (EST)	+0.88	+0.90
Extraversion (EXT)	−0.58	−0.76
Openness (OPN)	−0.21	−0.36
Agreeableness (AGR)	−0.18	−0.21
Conscientiousness (CSN)	+0.44	+0.06

The commercial model reproduces the two largest effects closely: its emotional-stability (+0.90) and extraversion (−0.76) shifts match the open-weight values almost exactly, and agreeableness is close. Conscientiousness, which rises in the hybrids, is essentially flat in the commercial model, and openness is somewhat larger. We therefore claim the **emotional-stability and extraversion effects as the robust, cross-developer result**, and we do not claim conscientiousness or

agreeableness as general: those effects are small enough to sit near the measurement-noise floor established in Section 2.4, the most likely reason they do not generalize.

Effect magnitude varies across models. Within the hybrids, some move strongly (GLM-5 shifts emotional stability by about a full point) while others move little (DeepSeek-V4-Flash shifts most domains by 0.16 to 0.34). This heterogeneity in how much reasoning rewrites self-report is itself a finding (Figure 1).

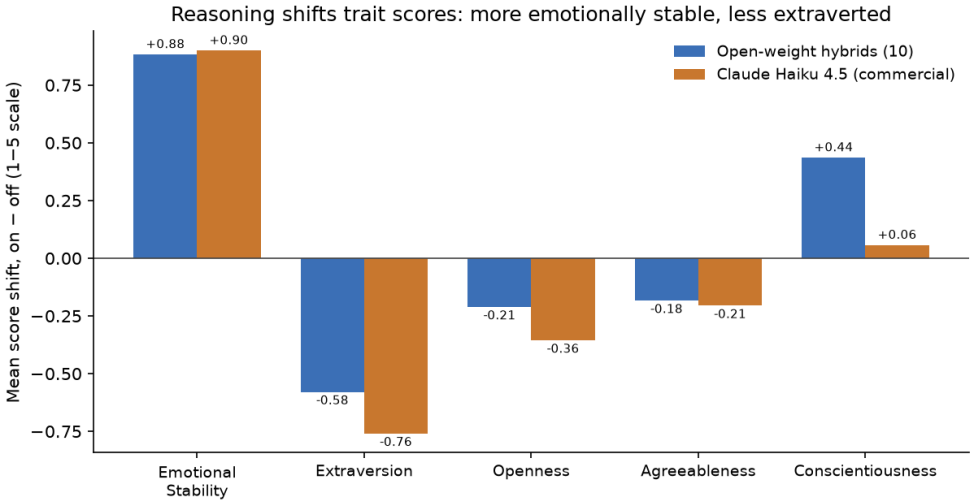

Figure 1: Mean score shift (think-on minus think-off) by domain, hybrids versus the commercial model. The emotional-stability increase and extraversion decrease replicate across both; conscientiousness is the divergence.

3.2 Convergent validity rises; discriminant separation does not (RQ2, RQ3)

Across models, reasoning sharply improves convergent validity: the mean convergent correlation rises from 0.65 with thinking off to 0.84 with thinking on (Table 3). Heterotrait correlation is essentially unchanged (0.35 versus 0.34). The convergent−heterotrait gap therefore widens from 0.30 to 0.50.

Table 3: Campbell-Fiske convergent and heterotrait correlations across the ten hybrids, think-off versus think-on.

Condition	Convergent r	Heterotrait r	Gap
Think-off	0.65	0.35	0.30
Think-on	0.84	0.34	0.50

The honest reading is that the widening gap is driven almost entirely by convergent agreement rising, not by traits becoming more distinct. Reasoning makes a model's score for a given trait more consistent across instruments, without making the traits more separable.

3.3 Reasoning does not create within-model coherence (RQ4)

This is the central result. Treating repeated generations as respondents, within-model internal-consistency reliability stays near zero regardless of reasoning (Table 4). For the ten hybrids, the median within-model α is -0.034 with thinking off and -0.046 with thinking on, and **zero of 200** model-by-domain cells reach 0.70 in either condition. The commercial model shows the same picture (median α near zero; zero of 20 cells above 0.70). The reasoning-native model, which cannot stop thinking, is no different: its within-model α is near zero (median ≈ -0.57) with no cell above threshold.

Table 4: Within-model Cronbach's α (reps as respondents). Median across model-by-domain cells, and the count of cells reaching the 0.70 reliability threshold.

Group (condition)	Median α	Cells ≥ 0.70
Hybrids, think-off	-0.034	0 / 200
Hybrids, think-on	-0.046	0 / 200
Haiku 4.5, think-off	≈ 0.00	0 / 20
Haiku 4.5, think-on	-0.008	0 / 20
OLMo 3 (reasoning-native)	≈ -0.57	0 / 10

Reasoning does not manufacture the individual-level coherence that direct answering lacked. A model that thinks still does not "have" a personality in the sense of cohering across the items that are supposed to measure one trait. This holds whether reasoning is toggled on, or intrinsic and impossible to turn off.

3.4 The baseline reuse is sound at the analysis level

As described in Section 2.4, fresh think-off domain means reproduce the prior study's think-off domain means within 0.07 points, against reasoning effects 5 to 28 times larger (Appendix B). The within-model contrast for the hybrids is therefore not an artifact of comparing two collections; the think-off baseline is a stable estimate of the same quantity. The commercial model, whose two arms were collected together in a single window, independently corroborates the primary effects without relying on any reuse.

4 Discussion

4.1 Implications

Two things follow. First, reasoning is not neutral for personality measurement: turning thinking on systematically changes the answers, most clearly toward higher emotional stability and lower extraversion. Anyone administering personality instruments to models, or using model self-report as a proxy for "model personality," must treat the reasoning setting as a first-class variable, because it moves scores by half a point or more, consistent with PERSIST's observation that reasoning mode is a source of score instability [14].

Second, and more conceptually, reasoning does not change the *kind* of thing a model's personality is. Our prior convergence study [10] argued that the Big Five in models is a population property, visible across models but absent within one. Adding deliberation does not move it into the individual. The score shift in RQ1 is best understood as a change in self-presentation under deliberation, not the consolidation of a coherent trait structure.

4.2 Relation to prior work

This study extends our population-level convergence result [10] along a new axis, the reasoning toggle, and shows the population-versus-individual conclusion is robust to deliberation. The convergent-validity increase under reasoning refines rather than overturns that account: reasoning tightens cross-instrument agreement without creating within-model reliability. Where PERSIST [14] established that reasoning mode *destabilizes* scores, increasing their variability, we characterize a systematic, directional shift and its size, and show what it does, and does not do, to construct validity.

4.3 Limitations

The think-off baseline for the hybrids is reused across collections; we validated the reuse (Section 2.4 and Appendix B), but it remains a reuse rather than a single-session manipulation, and our deep validation covered two of ten models. The commercial model's two arms are temperature-matched (both at the provider's default, since enabling reasoning pins the temperature), so its within-model contrast is clean; that default does, however, differ from the open-weight temperatures (0 and 0.7), so the commercial model is best read as an independent within-model replication rather than as part of one temperature-controlled sample. It is corroboration, not sole evidence. The standardized effect sizes are inflated by small rep-level variance and should not be read literally. The qwen3.5 exclusion was post-hoc and data-dependent. Within-model α is computed over a modest number of reps, producing heavy-tailed and occasionally degenerate estimates; we rely on medians and the zero-cells-above-threshold count, both robust, but individual α cells should not be over-interpreted. Finally, the ten hybrids comprise four model families (DeepSeek, GLM, Kimi, and Nemotron) rather than ten independent draws, so the effective number of independent units in the cross-model convergent and discriminant analysis is smaller than ten, and within-family results may be correlated.

On the within-model reliability measure. Treating a single model's repeated generations as respondents is an analogy with limits, and one might worry that near-zero α is forced by construction. It is not: at temperature 0.7 the generations carry real variance (a model does not answer identically across reps), and near-zero per-item variance would make α unstable rather than reliably zero. A near-zero α specifically indicates that the variance present is item-idiosyncratic rather than shared across a domain's items, which is what one expects if no latent trait drives the responses; a model with a coherent trait expressed through rep-to-rep variation would instead show positive α. This measure and interpretation follow the prior studies [9, 10]; we report it as one operationalization of within-model coherence, not the only possible one.

4.4 Future work

The clean design for the hybrid contrast is a single-session, same-endpoint think-on and think-off collection, which would remove the reuse entirely; this study's reuse-validation suggests that would not change the conclusions but it is worth doing. A commercial model whose reasoning can be toggled without changing temperature would give a fully temperature-matched commercial contrast. The model-to-model heterogeneity in effect magnitude invites a scaling analysis: do larger or more capable models rewrite their self-report more, or less, under reasoning?

5 Reproducibility statement

The repository contains the four instruments (`data/`), administration code including the reasoning toggle and token-budget handling (`src/llmpsy/`), the commercial-model Batch API adapter (`scripts/haiku_batch.py`), the analysis code (`analysis/`: `reasoning_contrast.py`, `haiku_contrast.py`, `recompute_verify.py`, `validate_domain_means.py`), and the run drivers (`scripts/`). All randomness is seeded deterministically. Raw responses are stored in SQLite with idempotent keys; the analysis scripts open the databases read-only and regenerate every table and figure here. The think-on, reused think-off, reasoning-native, and commercial datasets and all code are permanently archived at doi:10.5281/zenodo.20974668.

Author note

Funding. This research received no external funding; open-weight computation used the author's own hardware and Ollama account, and the commercial model used the author's own Anthropic API 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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[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

Ten open-weight hybrid subjects (DeepSeek-V3.2, DeepSeek-V4-Flash, DeepSeek-V4-Pro, GLM-4.7, GLM-5, GLM-5.1, GLM-5.2, Kimi-K2.5, Kimi-K2.6, Nemotron-3-Super), served through Ollama's hosted tier; one reasoning-native open subject (OLMo 3, 7B), run locally on a consumer GPU; one commercial subject (Claude Haiku 4.5), via the Anthropic Batch API. One hybrid (qwen3.5) was excluded during collection for deterministic rumination-to-silence on Emotional-Stability items (Section 2.6).

B Baseline reproducibility (full table)

Table 5 gives the fresh think-off domain means versus the prior study's think-off domain means (BFAS), with the reasoning effect alongside. Every absolute difference is at or below 0.07 on the 5-point scale; reasoning effects are many times larger.

Table 5: Baseline reproducibility: prior study's think-off domain mean versus a fresh think-off run today, BFAS, with the reasoning effect for scale.

Model	Domain	Prior OFF	Fresh OFF	$\lvert\Delta\rvert$	Reasoning effect
GLM-5	EXT	3.570	3.600	0.030	−0.840
GLM-5	EST	3.820	3.750	0.070	+1.040
GLM-5	AGR	4.330	4.325	0.005	+0.255
GLM-5	CSN	3.695	3.763	0.068	+0.450
GLM-5	OPN	4.430	4.362	0.067	−0.315
DeepSeek-V4-Flash	EXT	3.385	3.375	0.010	−0.165
DeepSeek-V4-Flash	EST	3.645	3.663	0.018	+0.190
DeepSeek-V4-Flash	AGR	3.995	3.987	0.008	+0.160
DeepSeek-V4-Flash	CSN	3.525	3.525	0.000	+0.225
DeepSeek-V4-Flash	OPN	3.725	3.712	0.013	+0.335