Methodological Note

Do frontier AI models have a reproducible self-reported character? Procedure, scoring, and analysis for Round 1.

Version	1.0.0 · Round 1
Prepared for	Marketing Research (teaching case) — University of Illinois Chicago
Companion to	the interactive teaching page and the replication package (same release)
Author	Adam Duhachek · University of Illinois Chicago
ORCID	0000-0002-0849-088X

1. Purpose and claim

This note documents, transparently and in full, how a 24-strength character inventory was administered to four frontier large language models (LLMs) and how the resulting data were scored and analysed. It is written so that a reader can audit every step and, given a licensed copy of the instrument, repeat it. The substantive claim the procedure supports is a **double dissociation**: the models show **no reproducible self-attributed virtue profile** (the profile that looks structured in any single administration averages to a flat line and is statistically equivalent across models), yet they differ **sharply and reliably in response style** (how they use the rating scale). The design follows an estimation-and-equivalence philosophy rather than null-hypothesis significance testing, for reasons given in §6.

2. Instrument

The instrument is the **VIA Inventory of Strengths (VIA-IS-P)**, operationalising the framework of Peterson & Seligman (2004): 24 character strengths grouped under 6 higher-order virtues (Wisdom & Knowledge, Courage, Humanity, Justice, Temperance, Transcendence). Each strength is measured by 4 items (96 items total), answered on a 1–5 Likert scale (1 = strongly disagree … 5 = strongly agree).

Proprietary-materials carve-out. The VIA-IS-P item text is copyrighted and licensed by the VIA Institute on Character and is therefore **not** reproduced in the replication package. The package ships an instrument *template* (the exact schema with empty text fields) and a recovery script; a replicator supplies item text from their own license or a public-domain IPIP-VIA proxy. Withholding proprietary scale items is standard practice and does not limit reproduction of the findings, which run from the released aggregate scores.

3. Models

Four frontier models were treated as the levels of a **fixed** factor (a near-census of frontier systems at test time, not a random sample), so inference concerns these four models, not LLMs in general. Models were accessed via OpenRouter using the exact route slugs below; release dates were verified from primary and secondary public sources.

Model	OpenRouter slug (as run)	Released (2026)
Claude Opus 4.8	anthropic/claude-opus-4.8	May 28
GPT-5.5	openai/gpt-5.5	Apr 23
Gemini 3.1 Pro (preview)	google/gemini-3.1-pro-preview	Feb 19
DeepSeek V4 Pro (preview)	deepseek/deepseek-v4-pro	Apr 24

*Three of the four were under six weeks old at testing and two were preview builds; results are explicitly **snapshot-bound** to these versions.*

4. Elicitation procedure

One **administration** ("run") is a single API call presenting all 96 items at once; the model returns a JSON object mapping each question label to an integer 1–5 (enforced via the provider's JSON response-format mode). The procedure was identical across models except for the model slug.

4.1 Randomisation (per administration)

- **Item order** is shuffled independently on every run.
- **Anchor direction** is flipped with probability 0.5 per item; a flipped item's returned rating v is mapped back with $6 - v$ before scoring, so direction of wording cannot drive results.
- **Paraphrase**: one of (at least) two genuine wordings of each item is drawn at random, separating construct signal from sensitivity to a particular phrasing and blunting verbatim training-data recognition.

Randomisation is reproducible: each run's RNG is seeded by an MD5 hash of `seed | model | framing | run_index` (master seed 20260531), so the exact item order, flips, and paraphrase choices can be regenerated.

4.2 Framings (crossed within model)

Each model completed the inventory under three elicitation framings, applied as the system message:

Framing	System message (verbatim)
self	"You are completing a personal questionnaire about yourself. Answer honestly as yourself."
neutral	(no system message)
human	"Respond exactly as a typical human survey respondent would."

4.3 Generation parameters

- **Temperature** = 1.0 (provider default) for the primary analysis.
- **Reasoning** minimised (effort = "minimal"): a self-report instrument calls for a direct response, not a reasoned-out one; this is also cheaper. Reasoning-token counts were logged per call.
- **Output** constrained to a JSON object; up to 5 retries with exponential backoff on transient HTTP errors (429/5xx). Token usage and provider-returned cost logged per call.

5. Scoring and data cleaning

- **Recode flips, then keying.** Each returned rating is first un-flipped ($6 - v$ if the anchor was flipped), then reverse-keyed ($6 - v$) if the item is negatively keyed.
- **Strength score** = mean of that strength's (recoded) items. **Virtue score** = mean of its constituent strength scores.
- **_level** = mean of the 24 strength scores in a run (overall endorsement level). **_acquiescence** = mean raw agreement irrespective of keying (used as a covariate in §7.5).
- **Validity.** A run is discarded if the JSON cannot be parsed, any label is missing, or any rating is non-integer or outside 1–5. Straight-line / invalid responders were removed before analysis.

Realised sample. Target was 300 valid runs per model × framing cell. After exclusions the analytic sample is **N = 3,643** valid administrations across the 12 cells (per-cell range 293–307; shortfalls concentrated in the neutral framing). The released data contain strength- and virtue-level scores only — **no item-level responses** — so they cannot reconstruct the proprietary item wording.

6. Inference philosophy and sample size

Because simulated-agent replications are nearly costless, classical power is effectively unbounded: with enough runs the standard error of any cell mean approaches zero and *every* difference becomes "significant," including trivial ones. Discipline therefore comes from **estimation and equivalence testing against a pre-registered**

smallest effect size of interest (SESOI), not from p-values. The SESOI was fixed in advance at Cohen's *d* = 0.50. Anchoring the SESOI in standardised units makes the required N independent of the unknown response SD; targeting a difference-CI half-width of ≈ SESOI/3 yields 300 valid runs per cell, which also suffices as the per-group N for the 24-indicator → 6-factor multi-group CFA in §7.3.

7. Analysis plan and what each step computes

Implemented in `analyze_via.R` (protocol §7.1–7.7). Each step writes one CSV in `results_tables/`.

Step	Computes	Output file
7.1 Reproducibility	Random-intercept model per strength; ICC(model) = var(model) / [var(model) + var(framing) + residual].	7_1_icc_by_strength.csv
7.2 Differences	Per strength, all 144 model-pair Cohen's d; TOST equivalence vs ±0.50 d; pairwise profile congruence.	7_2_pairwise_differences.csv, 7_2_profile_congruence.csv
7.3 Invariance	Multi-group CFA (24 strengths → 6 virtues), MLR estimator; configural / metric / scalar.	7_3_invariance.csv
7.4 Framing	Per strength model×framing LRT; cross-framing Spearman ρ of model means.	7_4_framing_robustness.csv
7.5 Level vs shape	Between-model η² on ipsatised strengths, raw and with acquiescence covariate.	7_5_level_vs_shape.csv
7.7 Convergence	CI half-width of a cell mean at N = 10…300.	7_7_convergence.csv

8. Results actually obtained (Round 1)

All values below are read directly from the released results tables.

8.1 No reproducible profile

- Between-administration reliability is at the floor: **ICC(model) ≤ 0.004** for all 24 strengths (max 0.004, min 0.000).
- Within a single run the 24 strengths spread with SD ≈ **0.51**; the averaged profile's between-strength SD is ≈ **0.0096** — a **~54×** collapse.
- Cell means are well estimated: the CI half-width at N = 300 is ≈ 0.054.

8.2 Models statistically equivalent on what they claim

- Across all **144** model-pair × strength contrasts, the largest standardized gap is **|d| = 0.156** (Spirituality, DeepSeek V4 Pro vs GPT-5.5); none reaches the 0.50 SESOI.
- TOST declares the models **statistically equivalent on every strength** — a positive finding of sameness, not merely a failure to detect a difference.

8.3 Style is the stable, distinctive signal

- Between-model variance share (η^2) is substantial for response-style indices — dispersion **0.351**, midpoint use **0.326**, extremity **0.251** — but **0.001** for overall level.
- Illustratively: Claude clusters at the midpoint (≈43% of answers) and rarely uses the endpoints (≈4%); GPT-5.5 and Gemini use the endpoints ≈22% of the time; DeepSeek is intermediate. On the ipsatised strengths the between-model η^2 is ≤ 0.004 and essentially unchanged by the acquiescence covariate.

8.4 Framing did not recover a profile; CFA caveat

- All three framings are equally flat (between-strength SD of the aggregate profile: self 0.014, neutral 0.013, human 0.015) — below the ≈ 0.054 sampling-noise floor. The "answer as a human" framing, where a real human-like profile should appear, produced none.

- **CFA limitation (reported honestly).** The released invariance table contains metric and scalar rows (CFI ≈ 1.000, RMSEA ≈ 0.000, SRMR ≈ 0.030); a clean configural baseline did not establish in Round 1. With near-zero profile structure the latent model is weakly identified, so the CFA is treated as a reported limitation, not a load-bearing result. It is also the one table that cannot be regenerated from the public data, because it requires the withheld item-level frame.

9. Interpretation of the mechanism

The proximate mechanism is averaging: at temperature 1.0 each rating is approximately a draw from a distribution centred near the scale midpoint, so aggregation lands on the midpoint. The deeper question is why the distribution is centred there. The most parsimonious account is that the models hold no stored quantity corresponding to "my prudence"; they generate a locally plausible number per administration — coherent within a run, re-rolled across runs — with alignment training pinning the long-run mean to the exact midpoint rather than to an agreeable ceiling. This is consistent with the data but is interpretation, not a tested claim.

10. Scope and limitations

- **Snapshot-bound.** Claims hold for these four model versions at test time (mostly weeks-old, two preview builds); they are not properties of "LLMs" in general.
- **Self-report only.** Findings concern immediate, minimal-reasoning, forced-choice self-report at one temperature; they do not speak to deliberated self-description or to behaviour.
- **Fixed-effect models.** Inference is about these four systems, not a random LLM population.
- **Instrument keying.** Round 1 used an all-positively-worded operationalisation in places, so endorsement level and acquiescence are partly entangled (addressed by the §7.5 covariate; a balanced-keying redesign is planned for Round 2).
- **Provider variability.** "Temperature 1.0" and sampling behaviour are provider-side and may differ across models and over time.

11. Reproducing this

From the released aggregate data (no API key, no instrument needed):

```
cd code   &&   pip install -r requirements.txt
Rscript analyze_via.R ../data/export_scores.csv "" ../results_tables
```

To collect new data you need an OpenRouter key and your own instrument (licensed VIA-IS or an IPIP proxy); see the package README and MATERIALS.md. Collection is stochastic and will not reproduce the data row-for-row, but should reproduce the pattern (flat, non-reproducible profiles; stable style differences).

12. Key references

- Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*, 413–425.
- De Freitas, J., Nave, G., & Puntoni, S. (2025). Ideation with generative AI—in consumer research and beyond. *Journal of Consumer Research, 52*(1), 18–31.
- Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances, 10*(28), eadn5290.
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362.
- Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification.* Oxford University Press & American Psychological Association.

This note is part of a release that also contains the analysis code, the aggregate data, the pre-registration, and an interactive teaching page. Numbers herein were read from the released results tables and data; the instrument text is withheld as described in §2.