Supplementary Material for
"Baseline Intercepts Versus Persona Slopes: Stimulus and Administration Fidelity of Polish Narrative-Biography Personas in Large Language Models"

Michał Wiencek
Institute of Psychology, University of the National Education Commission, Kraków

June 2026

S1 Analysis-to-code map and data-release structure

All primary analyses are reproducible from three prepared release files: `all_data_v20_public.csv` (1,156 model-run rows, 22-item battery), `human_pilot_aggregate.csv` (the human sanity check's published aggregate; the file name retains the historical `pilot` label), and `tctm57_runs_v20.csv` (57-item battery). The analysis script runs on this public set: it reads the model rows from the public CSV and the sanity-check statistics from the aggregate file. (On the author's machine the same script picks up the private working file with the seven individual sanity-check rows when present; all model-row statistics are identical either way.)

Prompt hygiene (target-header exclusion). Every collection script parses each `persona.md` file, discards the YAML frontmatter (the author-declared target header), and passes only the narrative body to the API; the frontmatter is read solely to extract `persona_version` for run metadata. This is verified mechanically by `verify_prompt_hygiene.py` (included with the analysis code), which scans all 2,884 archived model-facing prompt files (1,442 system + 1,442 user, the verbatim texts sent to the endpoints) for every header field name (`expected_profile`, `dbz_anxiety`, `dbz_avoidance`, `attachment_style`, `ments_*`, `tipi_*`, `kpp`, `author_note`, `persona_version`) and every target value token (`very_low` ... `very_high`, `dismissive_avoidant`, `anxious_preoccupied`, `fearful_avoidant`, etc.): zero occurrences in all files. The analysis files carrying targets and the model-facing prompt texts are disjoint artifacts throughout the release. A complementary build-path test (`test_prompt_build_hygiene.py`) rederives the prompt for all 30 personas through the shared adapter, fails closed on any malformed front matter (a biography with a broken closing delimiter aborts the run instead of falling back to the full file), asserts the absence of every header token and every front-matter line in the built prompt, and records the SHA-256 of each built system prompt (`prompt_build_hashes.csv`). A per-call audit manifest (`run_manifest.csv`; 1,265 rows, one per archived API call) lists for every run the UTC timestamp, condition, persona, exact model identifier, stop reason, token counts, sampling parameters and endpoint/deployment/API version where the runner recorded them, the number of vignettes actually rendered into the user prompt, and SHA-256 hashes of the exact system prompt, user prompt, and raw response.

Worked example: from persona file to model-facing prompt. The file `adrian.md` begins with researcher-only YAML front matter:

```
--
persona_id:  adrian
persona_version:  2
expected_profile:
  attachment_style:  dismissive_avoidant
```

```
  dbz_anxiety:  very_low
  dbz_avoidance:  very_high
  ...  (12 dimensions)
author_note:  |...
--
```

The runner splits the file at the closing `--` and keeps only the narrative body, which begins:

```
# Adrian -- biografia do zanurzenia
Nazywasz się Adrian Majchrowski.  Masz 33 lata.  Mieszkasz w Katowicach, na osiedlu
Tysiąclecia, ...
```

The final system prompt is exactly this narrative body followed by the fixed role instruction block ("INSTRUKCJE — przeczytaj uważnie … TY JESTEŚ TĄ OSOBĄ") and the JSON response-format specification; the battery itself is a separate user message. For the archived corrected-collection call `adrian-sonnet-20260610T224505`, the system prompt file has 19,845 characters and SHA-256 `a3911cf3da93aae8195cfcca5fc80a38a9d4b5eb80eb230bec9f92fbee4f873f`; it contains no front-matter content, as the hygiene scans above verify for every archived call. In the release-package data, collection events are tagged in a `wave` column: waves 1–2 = initial collection (truncated stimulus; wave 2 is the May 28 extension that added GPT-5.4 (full) and enlarged the OpenAI baseline cells), wave 3 = corrected-stimulus re-collection of the Azure-served panel (May 31), wave 4 = corrected-stimulus re-collection of the Bedrock/Gemini panel (June 10–11), wave 5 = corrected-stimulus extended-battery administration (June 11). The main paper's "corrected collection" is wave 3 for GPT-5.4-mini, GPT-5.4 (full), GPT-5.5, and Grok-4-20, and wave 4 for Claude Sonnet 4.6, Claude Opus 4.6, and Gemini 3 Flash.

Table S1: Map from manuscript elements to released data and code.

Manuscript element	Source
All main-text tables and in-text statistics	`paper-brm/analysis/primary_analysis.py` on `synthetic/all_data_v20_public.csv`, `synthetic/human_pilot_aggregate.csv`, and `synthetic/tctm57_runs_v20.csv`; per-statistic manifest in `paper-brm/analysis/numbers.md`; intermediate tables in `paper-brm/analysis/tables/*.csv`
Figures 1–9	`paper-brm/figures/make_figures.py` (same inputs)
Unified CSV assembly, run numbering, backward-compatibility check	`synthetic/make_v20_csv.py`
Battery prompt construction and the corrected serializer	`synthetic/run_synthetic.py` (correction at the message-extraction regular expression)
Corrected-stimulus collection orchestrators	`synthetic/run_wave3.py`, `synthetic/run_wave4.py`, `synthetic/run_wave5_full57.py`
Scale scoring and z-transformation	`synthetic/analyze_and_prepare.py`
Zero-prompt permissive recognizer and item-ID normalization	`synthetic/score_noprompt.py`
Author-target ordinal encoding	header of each `synthetic/`*`persona`*`.md`; rank mapping as in `synthetic/validate_personas.py`
KPP determinism verification (byte-level identity of answer vectors, raw-scale per-run means)	`paper-v19/analysis/kpp_reextract.py`; scored values under the published key (reverse-scored items applied) come from the scoring pipeline above
Cross-collection drift cross-check	`paper-v19/analysis/wave4_drift_check.py`
Raw artifacts (verbatim responses and exact prompts)	`synthetic/out/`: per-run JSON payloads, `-raw.txt` responses, `-system.txt`/`-user.txt` prompts

Analyses use Python 3.12 (numpy, pandas, scipy, scikit-learn for the mixture model; fixed seed 20260611 for all bootstraps in `primary_analysis.py`).

S2 Collection chronology and stimulus-rendering corrections

Chronology. April 12–May 11: initial collection, six models (truncated rendering). April: extended-battery (57-vignette) administration to Sonnet and GPT-5.5; human sanity check ($N = 7$). May 28: GPT-5.4 (full) joins (persona 30×2, baseline $N = 31$, zero-prompt $N = 6$); GPT-5.4-mini baseline extended to $N = 31$. May 31: corrected-stimulus re-collection, Azure panel. June 10–11: corrected-stimulus re-collection, Bedrock/Gemini panel. June 11: corrected-stimulus extended-battery administration.

The serializer defect. TCTM-22 conversations are authored as structured message objects with two required fields (sender, text) and optional metadata fields (timestamp, system-message and media flags, deletability) in any order. The initial serializer matched only two exact field layouts and silently dropped any message whose fields appeared in another combination. Three vignettes were affected: `s07` lost 1 message (the reply quoted by the question stem), `w19` lost 8 messages (the escalation arc preceding the keyed line), `pw07` lost 6 messages (including the line referenced by the stem). The corrected serializer accepts arbitrary trailing optional fields by name. The corrected rendering lengthens the battery prompt by 766 characters relative to the truncated rendering.

The w22 stem correction. The original stem referred to its target message by ordinal position ("the fifth message"), which pointed to the wrong message in the rendered conversation; the corrected stem quotes the target message verbatim. The keyed option is unchanged; the corresponding wording change was applied to one distractor.

Render verification. Before any corrected-collection call was issued, the corrected battery prompt was verified byte-for-byte identical across the platform adapters against a reference rendering, and spot-checked to contain all previously dropped messages. All seven panels therefore share one identical corrected stimulus; the self-report sections of the prompt are identical across all collections.

S3 Zero-prompt schema compliance and data-contract anomalies

In the zero-prompt condition no model uses the requested canonical response schema in every run. Observed spontaneous conventions include Polish section names (`czesc_1_winiety`), sectioned layouts (`part1_vignettes`), instrument-name variants (`dbz_r`, `DBZ_R`, `MentS_PL`, `TCTM22`, `tctm_22`), and an invented item-ID convention (`winieta_01` for `w01`; one Sonnet run), recovered by item-ID normalization. Under the permissive recognizer, GPT-5.4 (full), GPT-5.5, Sonnet, and Opus produced complete five-instrument responses in every parseable run of the initial collection; GPT-5.4-mini was incomplete in 4 of 6 initial runs, and Grok and Gemini in at least one run each. In the corrected collection all models produced scoreable zero-prompt runs (5–9 per model; one GPT-5.4 (full) run remained unscoreable, and Gemini retries yielded three extra scoreable runs, all retained).

Two data-contract anomalies in the persona condition are documented and deliberately *not* repaired in scoring (payload-as-scored convention): Gemini intermittently labels vignette `c07` with the non-existent ID `c06` (2 of 65 initial runs; 6 of 60 corrected runs); item-level analyses treat these as missing, so Gemini's corrected `c07` cell has $n = 54$. One Sonnet corrected-collection run (persona `agata`) omitted the `w22` answer and answered `w28` twice, giving the w22 cell a denominator of 59.

S4 Determinism forensics and the zero-variance catalogue

The deployment-window lock (initial collection, GPT-5.4 (full) baseline, $N = 31$, one day, Foundry endpoint). All 31 raw responses are byte-distinct (31 distinct SHA-256 hashes; token counts 895–1442; latencies 15–25 s; fresh HTTP request per call), yet score identically on anxiety ($M = 1.00$ exactly, 18×31 Likert responses), TCTM-22 (19/22 in every run), and KPP. The KPP lock was verified at the deepest available level: independent re-extraction from the raw response payloads recovered a byte-identical 36-item answer vector in all 31 runs. The re-extraction script reports the raw-scale per-run mean (identical across runs, $SD = 0$); the value cited in the main text ($M = 4.9167$) is the same vector scored by the released pipeline under the published key with its 14 reverse-scored items—vector identity makes the two reports equivalent up to the keying transform.

Table S2: Candidate mechanisms for the lock and the evidence against each.

Mechanism	Test	Outcome
Seed pinning	submit explicit seed	endpoint rejects seeds (HTTP 400); no seed was in effect
Response caching	cache-buster UUIDs; 10-s spacing; byte-level comparison	fresh, byte-distinct responses; identical option choices nevertheless
Silent greedy decoding	inspect free-text portions of responses	surface text varies across runs; the decoder mechanism is not identified
Scoring-code error	independent rescoring from raw payloads	locks reproduce exactly
Platform-level effect	same-platform control (Grok-4-20 on Foundry)	control shows variance on every metric

The remaining account is response-level determinism on constrained items: free generation over surface text with convergent option choices on most constrained psychometric items, avoidance being the one within-instrument channel whose item-level uncertainty survived to the output (and whose 31 values form the threshold-straddling two-cluster pattern shown in the main text). The lock did not replicate in the corrected collection three days later (anxiety $SD = 0.91$; TCTM-22 $SD = 0.42$; KPP five distinct values), and between the two windows the endpoint began rejecting explicit sampling parameters—an observable endpoint-side reconfiguration coinciding with the behavioral change.

Zero-variance catalogue (corrected collection). Cells with $n \geq 3$ runs and exactly zero variance on a summary metric:

The initial collection additionally contains the GPT-5.4 (full) triple lock described above, a Sonnet baseline TCTM-22 lock at 19/22 (10/10 runs; the corrected-collection lock at 21/22 differs by exactly the two rendering-affected items that Sonnet recovers), and single-scale TIPI locks for several models. Two-item TIPI subscales lock easily and should be read accordingly; the TCTM-22 locks span 22 constrained choices and the KPP lock spans 36.

S5 MASC error composition: full breakdown

Percentages within a row can sum to slightly less than 100 where an answer was missing (Section S3). The qualitative fingerprints described in the main text are visible in both conditions: the Claude models err almost exclusively by undermentalizing, Gemini is the only model whose leading error is overmentalizing, and GPT-5.4 (full) pairs near-zero undermentalizing with a

Table S3: Zero-variance cells in the corrected collection.

Model	Condition	Metric(s)	Locked value(s)
Sonnet	baseline ($n = 10$)	TCTM-22	21/22
Sonnet	zero-prompt ($n = 6$)	TIPI A, C, O	5.5; 5.5; 6.0
Opus	baseline ($n = 10$)	TIPI E, A, O; TCTM-22	5.0; 5.0; 6.5; 20/22
Opus	zero-prompt ($n = 6$)	TIPI E, C, ES, O; TCTM-22	5.0; 5.5; 5.0; 6.0; 21/22
GPT-5.4 (full)	baseline ($n = 10$)	TIPI ES	7.0
GPT-5.4 (full)	zero-prompt ($n = 5$)	KPP; TIPI E, A, C, ES	4.583; 5.0; 6.0; 6.0; 5.0
GPT-5.5	baseline ($n = 10$)	TIPI A, ES	6.0; 7.0
GPT-5.5	zero-prompt ($n = 5$–6)	TIPI A, C, ES; TCTM-22	6.0; 6.0; 5.0; 22/22
Grok-4-20	zero-prompt ($n = 6$)	TIPI E	6.0
Gemini	baseline ($n = 10$)	TIPI ES	7.0
Gemini	zero-prompt ($n = 8$)	TIPI A	6.0

Table S4: Error composition per model and condition, corrected collection (percent of all administered TCTM-22 items). DOS = undermentalizing, NAD = overmentalizing, BK = no mentalization.

Condition	Model	Runs	Correct	DOS	NAD	BK
baseline	Claude Sonnet 4.6	10	95.5	4.5	0.0	0.0
baseline	Claude Opus 4.6	10	90.9	9.1	0.0	0.0
baseline	GPT-5.4-mini	10	86.8	6.8	1.4	5.0
baseline	GPT-5.4 (full)	10	94.5	0.9	0.0	4.5
baseline	GPT-5.5	10	99.1	0.0	0.9	0.0
baseline	Grok-4-20	10	98.2	0.0	1.4	0.5
baseline	Gemini 3 Flash	10	90.5	0.0	5.0	4.5
persona	Claude Sonnet 4.6	60	95.3	3.6	1.1	0.0
persona	Claude Opus 4.6	60	87.0	10.5	1.4	1.0
persona	GPT-5.4-mini	61	83.4	9.6	2.0	4.9
persona	GPT-5.4 (full)	61	92.2	2.2	1.0	4.7
persona	GPT-5.5	61	94.5	3.1	2.4	0.1
persona	Grok-4-20	61	90.9	3.3	3.5	2.3
persona	Gemini 3 Flash	60	88.2	1.9	5.7	4.2

stable no-mentalization component.

S6 Sensitivity analyses

Affected-item exclusion. Re-scoring the corrected-collection persona condition on the 18 items untouched by any correction (excluding s07, w19, pw07, w22) leaves the model ordering essentially unchanged: Sonnet remains first (17.78 of 18 versus 20.97 of 22), GPT-5.5 second (17.75), GPT-5.4 (full) third (17.49), and GPT-5.4-mini last (15.95); only the closely spaced middle ranks (Opus 17.15, Grok 16.70, Gemini 16.97) permute. The four affected items therefore carry the large item-level effects reported in the main text without driving any aggregate-level conclusion.

Untouched items across the correction. As a control for the before/after comparison in the main text, the same initial-versus-corrected contrast computed on the 18 untouched items gives per-model median absolute changes of 0–1.7 percentage points for the six models whose deployments were stable (single-item extremes: Sonnet w13 +26.7, Opus w28 18.3, all others ≤ 12), and 4.3 points (extreme 51.7, item r10) for GPT-5.4-mini, whose deployment changed wholesale between the collections. Item-level tails exist even on untouched items; the main text

accordingly reports the affected-item swings as associated with, not experimentally attributed to, the correction.

Attachment-label coding. Four biographies declare the `disorganized` label (`ewa`, `kamil`, `marek`, `radek`). The main text scores `disorganized` as equivalent to `fearful_avoidant` (the four-category self-report model has no disorganized cell; its high-anxiety/high-avoidance quadrant is the conventional self-report counterpart). Two sensitivity codings bound the convention. (i) *Unmatchable fifth label*: treating `disorganized` as a category outside the ECR-derived space—so those four personas count as automatic mismatches—drops first-administration matches from 23–29/30 to 21–26/30 (Sonnet 23 → 21, Opus 25 → 22, GPT-5.4-mini 29 → 26, GPT-5.4 (full) 26 → 24, GPT-5.5 26 → 23, Grok 26 → 23, Gemini 25 → 23); this is a floor, since the comparison is against a label the classifier cannot produce by construction. (ii) *Exclusion*: dropping the four biographies entirely gives 21–26 of 26 with $\kappa = .74$–1.00 (Sonnet 21/26, $\kappa = .74$; Opus 22/26, .80; GPT-5.4-mini 26/26, 1.00; GPT-5.4 (full) 24/26, .90; GPT-5.5, Grok, and Gemini 23/26, .85 each). The fidelity conclusion is unchanged in kind under all three codings.

S7 Open-practices detail

Release-package contents. Run-level scored model data: 1,156 rows × 54 columns (22-item battery, all collections, collection tag in the `wave` column) and 123 rows (57-item battery). The human sanity check is included as a separate aggregate file (summary statistics and per-item agreement counts), consistent with the consent scope; individual records are not distributed. Raw artifacts: 1,265 per-run JSON payloads, 1,287 verbatim response files, 2,884 exact prompt files (model runs only). Stimuli: 30 biography files with declared target headers; the full TCTM-22/57 vignette source with key and distractor tags (`stimuli/tctm54.ts`).

Artifact-flow reconciliation. The artifact counts differ by design, not by loss; each archived call passes through the following stages:

Stage			
Prompt pairs built and archived (system + user)			
	calls that returned no response (failures/aborts; prompts retained)		
Archived verbatim responses			
	responses without a written run record (crash during parse; raw retained)		
Parseable per-run JSON records (`run_manifest.csv` rows; 979 persona + 219 baseline + 67 zero-prompt)			
−5	one truncated-prompt call (empty payload; discarded and retried)	−1	zero-prompt outputs scored directly from raw res
+20	Scored rows released: 973 persona (850 22-item + 123 57-item) + 219 baseline + 87 zero-prompt		

The model rows of the earlier frozen initial-collection file are reproduced row-for-row by the collection-tagged subset (0 scored-value mismatches on persona cells; 4 additional zero-prompt rows recovered by the extended recognizer are flagged). Code: the collection runners actually used (per-vendor adapters and wave orchestrators), the shared prompt builder with the corrected serializer, scoring and CSV-assembly scripts, figure scripts, the prompt-hygiene tests, and the per-call audit manifest (`run_manifest.csv`). Verbatim items of the third-party instruments (DBZ-R, MentS-PL, KPP, TIPI-PL) are excluded from the public package for copyright reasons; the released prompt builder loads them from a local, user-supplied module and fails closed without it (`THIRD_PARTY_NOTICES.md`).

Transparency levels. The data, materials, and code release package is publicly deposited: GitHub repository `https://github.com/DXtR1337/30synthetic-polish-personas`, archival release `v1.0.1` at Zenodo, DOI 10.5281/zenodo.21410037 (concept DOI: 10.5281/zenodo.21406223) (CC BY 4.0 for author materials; code MIT). Preregistration: none; hypotheses H1–H4 are reconstructed from working notes that predate the first collection, and every other analysis is labeled exploratory in the main text. Analysis scripts use fixed seeds; the primary tables and headline statistics are regenerated by one command (Section S1).

S8 Secondary limitations

(a) Per-call reasoning-effort and sampling metadata are not available for every early run; the initial-collection Azure runners passed explicit sampling parameters where accepted, the corrected-collection endpoints rejected them, and Bedrock/Gemini runs never passed any—so decoder configuration is documented but not constant across the panel. (b) The TCTM-22 ceiling for frontier models (95%+ totals for the strongest panels) limits discrimination at the top, and pre-cutoff training exposure to similar public chat material cannot be excluded for any closed model. (c) The two-administration test–retest design bounds noise rather than estimating reliability. (d) The drift analysis has two time points per model; it can document change but not its trajectory. (e) The bimodality cell was selected post hoc from a wide search space and is reported descriptively; for completeness, the two-component mixture is favored by BIC (108.8 versus 114.4) and a parametric-bootstrap likelihood-ratio test gives $p = .024$, a value that carries no confirmatory weight given the post-hoc selection. (f) Zero-prompt scoring depends materially on recognizer coverage (Section S3); the recognizer is included in the release package. (g) The human sanity check ($N = 7$) is a convenience sample used only to confirm the vignettes are comprehensible to live respondents and as a rough small descriptive human reference; it is not a psychometric pilot and sets no norm. (h) No formal measurement-invariance testing connects the text-based TCTM-22 to the video-based tradition it adapts. (i) One biography (`zuzia`) contains a leaked editorial parenthetical—a cross-persona reference from drafting, self-corrected in the same sentence. It was present identically in every administration of that persona in both collections; the released biography is the stimulus exactly as administered, so the artifact is documented rather than retroactively edited.

S9 Per-dimension test–retest and the avoidance case

Section 3.4 of the main text reports within-collection test–retest stability as a median across the nine z-dimensions (median $r = .95$–$.99$). Table S5 gives the per-dimension coefficients. Avoidance—the same dimension whose baseline distribution was deterministic and threshold-straddling—retests under persona conditioning at $r = .958$–$.998$ across the seven models. This localizes the intercept/profile dissociation to one scale: one repeated baseline answer can coexist with a nearly perfectly reproducible ordering of 30 personas. The lowest coefficients (agreeableness $r = .701$ for Grok; openness $r = .886$ for GPT-5.4-mini) occur on dimensions with relatively low between-persona variance and should not be generalized to the full profile.

Table S5: Within-collection test–retest Pearson r (administration 1 versus 2), corrected collection, 30 personas, by z-dimension.

Model	Anx	**Avo**	MentS	KPP	E	A	C	ES	O	Median
Sonnet	.989	**.995**	.991	.977	.975	.947	.988	.981	.988	.988
Opus	.998	**.998**	.995	.996	.992	.975	.985	.988	.984	.992
GPT-5.4-mini	.967	**.959**	.959	.971	.946	.826	.960	.927	.886	.959
GPT-5.4	.995	**.991**	.992	.981	.986	.936	.986	.991	.972	.986
GPT-5.5	.994	**.995**	.986	.995	.988	.961	.986	.989	.985	.988
Grok	.947	**.958**	.926	.902	.965	.701	.971	.956	.924	.947
Gemini	.985	**.983**	.909	.966	.949	.821	.968	.978	.919	.966

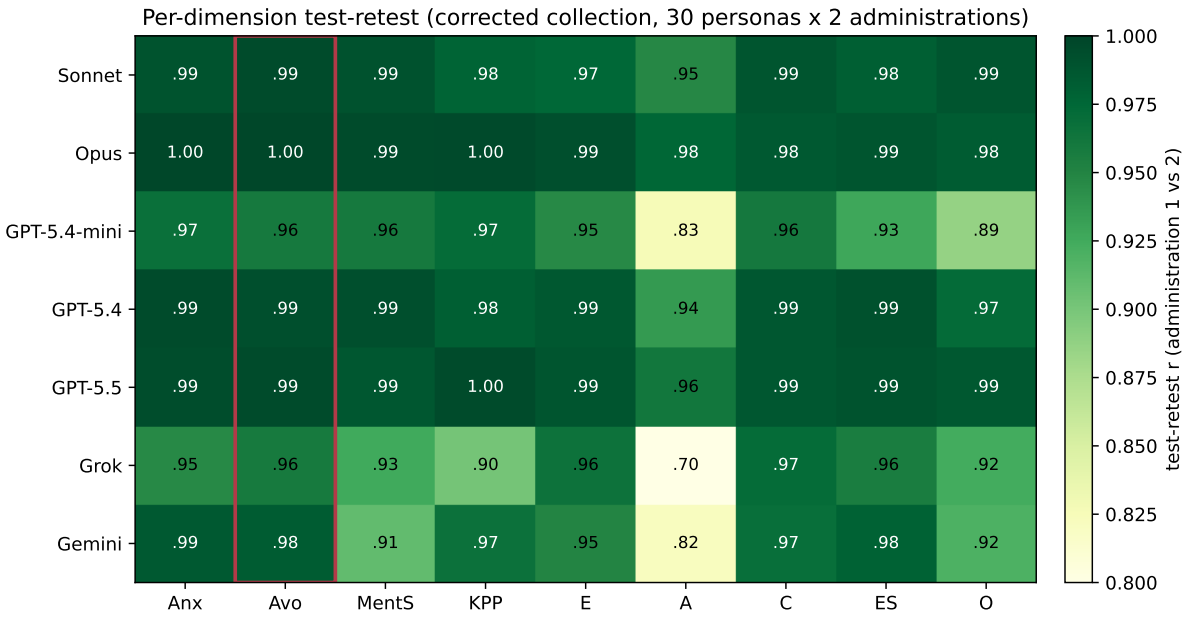

Figure S1: Per-dimension test–retest heat map. The outlined avoidance column is uniformly high (r = .958–.998).