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, Poland

Correspondence: seido1337@gmail.com

June 2026

Abstract

Large language models (LLMs) are increasingly used as substitutes for human respondents, but prior work reports strong social-desirability bias and poor individual-level psychometric fidelity under brief demographic prompts. This study tests a denser regime: 30 Polish second-person narrative biographies (1,489–2,861 words), each encoding a predetermined 12-dimension profile, administered to seven LLMs from four vendors under persona, baseline, and zero-prompt conditions with five instruments. Four findings emerge. First, persona-conditioned responses agree with author-declared targets (attachment-style Cohen's $\kappa = .69$–$.96$; seven-model Fleiss $\kappa = .85$). Because the targets and the TCTM-22 key are author-defined and not independently adjudicated, these are stimulus-fidelity quantities, not validated-accuracy estimates. Second, responding separates into fragile *baseline intercepts* (levels) and portable *persona slopes* (relative profile mappings and gains): model baselines differ by up to

1.7 *SD* on Polish norms, yet all 21 between-model Pearson-profile medians exceed .94. Third, deployment-window response locks, a descriptively bimodal post-hoc cell ($N = 31$), and a multi-*SD* baseline shift occur on the level axis, while persona mappings replicate at median $r \geq .92$. Fourth, a corrected stimulus-rendering defect and a confounded battery-composition comparison show item-level key agreement changing by up to 98 percentage points while aggregate scores remain comparatively stable. These results motivate reporting stimulus and administration fidelity alongside LLM-as-respondent findings. A release package containing scored data, stimuli, prompts, and code is prepared for archival deposition.

Keywords: synthetic respondents; narrative personas; large language models; attachment; mentalization; Polish psychometrics

1 Introduction

A growing literature evaluates large language models (LLMs) as substitutes for human respondents in survey and psychometric research. Its results split along two axes. The first is *population-level fidelity*: Argyle et al. (2023) showed that GPT-3, conditioned on thousands of demographic backstories from real survey participants, produces "silicon samples" whose response distributions match human subgroups, a property they termed algorithmic fidelity; Aher et al. (2023) replicated several classic behavioral experiments with LLM participants. The second axis is *individual-level psychometric fidelity*: when a model is asked to answer as one specific person, do its responses form a coherent, target-consistent profile? Serapio-García et al. (2025), in a framework evaluated on 18 LLMs, reported that personality scores from some instruction-tuned models show reliability and validity evidence and can be shaped by prompting, but Petrov et al. (2024) found that under brief demographic prompts both GPT-3.5 and GPT-4 yield poor individual-level psychometric properties. In parallel, Salecha et al. (2024) documented a large social-desirability bias in LLM survey responses (around $d = 1.0$–1.2 on human *SD* scale) that grows with model recency, and Cui et al. (2024) found that scenario-based replications reproduce direction far more often than effect size. The individual-level news, in short,

has been mixed, and it appears sensitive to how the persona is specified.

The present study examines a dense conditioning regime: the *narrative biography*—a 1,489–2,861-word, second-person Polish text with named characters, embedded relational history, and idiomatic register, encoding a predetermined profile across twelve psychometric dimensions. Narrative biographies differ from brief demographic prompts in the number of behavioral constraints they impose per page, and from aggregated demographic backstories in that they specify one individual rather than a demographic class. Rich-profile narrative conditioning is an active line of work: Wigler et al. (2026) generated first-person life stories from the psychometric profiles of 290 real participants and had independent LLMs recover trait scores from the narratives alone (mean $r = .750$); Kang et al. (2025) used long synthetic backstories to bind virtual personas deeply enough to reproduce in-group partisan misperceptions; Bai et al. (2025) report that persona-simulation quality scales with profile detail and realism; and Han et al. (2025) document a dissociation between LLM self-reports and behavior that cautions against reading questionnaire outputs as traits. Earlier role-play personality evaluation is surveyed in Wang et al. (2024); Lee et al. (2025); Li et al. (2024). Against this neighborhood the present design differs in combining (a) long culturally-embedded biographies in a non-English language, (b) a multi-instrument *self-administered* battery normed in that language, (c) a multi-vendor model panel with repeated administrations, and (d) a re-collection of the full design on a corrected stimulus that doubles as a longitudinal replication—with the stimulus-fidelity and administration-fidelity results (Sections 3.5–3.6) and the deployment-window instabilities (Section 3.3) as the contributions no neighboring study reports. Computational-psychometric work in Polish remains sparse (e.g., Plisiecki & Sobieszek, 2024), which further motivates evaluating the measurement battery in that language.

Beyond asking *whether* dense narrative conditioning recovers individual-level signal, the design permits a decomposition that, the paper argues, is the more useful lens for LLM-as-respondent work. Any persona-conditioned response is produced by a model that also has a *baseline*: what it answers when asked to respond as itself. The model-specific baseline position is termed the *intercept* here, and the mapping from biography to persona-

conditioned profile the *slope*. The study was structured so that three different potential sources of instability—differences between models, differences between deployment windows of the same model, and changes within single models between collections days to weeks apart—can each be located on one of these two axes.

Finally, the study contributes an unplanned methodological result: a prompt-serialization defect, discovered mid-study and then corrected, silently truncated the rendered conversation in three of the 22 vignettes of the author-keyed test, and a stem wording issue affected a fourth. The full design was re-collected on the corrected rendering, which converts an embarrassment into an informative (though uncontrolled) before–after comparison on stimulus fidelity; a follow-up administration of an extended 57-vignette battery converts a quality-control anomaly into a structured but confounded comparison of administration context. Both show that item-level results in LLM-as-respondent designs can be artifacts of the rendered stimulus or the administration protocol even when aggregate scores are robust.

The paper reports the design (Section 2), the results (Section 3), and the implications (Section 4), with limitations (Section 5) and follow-up directions (Section 6). Throughout, the five-instrument battery is treated as a *measurement spine* that records what each model produces under each condition, not as a validation battery that certifies an underlying construct; all agreement results are framed as stimulus fidelity, not construct validity, because the author-declared targets have not been validated by an independent rater panel and the seven model "raters" are not statistically independent.

2 Method

2.1 Personas

Thirty Polish-language biographies were written in the second person (1,489–2,861 words of narrative body), each with a structured header declaring the author-intended profile on 12 dimensions: attachment anxiety and avoidance, three mentalization subscales (self, other, motivation), need for cognition, the Big Five (E, A, C, ES, O), and a categorical

attachment style (secure, anxious-preoccupied, dismissive-avoidant, fearful-avoidant). The structured target header was used exclusively as analysis metadata and was never included in any model-facing prompt: every collection script parses the biography file, discards the header, and passes only the narrative body to the API. This is verified mechanically—an exhaustive scan of all 2,884 archived model-facing prompt files (1,442 system + 1,442 user) finds zero occurrences of any header field name or target value token (Supplement S1). The biographies were authored iteratively by the author with Claude Opus 4.6 as a writing assistant, then reviewed and edited before any model run; the resulting generator dependency is discussed in Section 5. The set is a *curated demonstration corpus*, not a sampled or externally validated stimulus set.

The biographies are deliberately culturally embedded: characters live in named Polish cities, attend specific Polish universities, reference Polish workplace and healthcare contexts, and write in idiomatic Polish registers (lowercase chat style for some characters, formal punctuation for others). Nine biographies additionally contain explicit trait vocabulary in the body text (the "explicit" subset); the remaining 21 convey the profile through narrative only, which allows a check that cross-model agreement does not reduce to trait-keyword detection (Section 3.1).

One persona was designed as a discriminative paradox: `ola` encodes secure attachment combined with low mentalization on all three subscales—a pattern rare in human samples, where security and mentalization correlate positively. If models simply reproduced the human-typical covariance structure, they would fail to reproduce Ola's author-intended dissociation.

2.2 Instruments

Every run in every condition administers the same five-instrument battery in Polish.

DBZ-R (Lubiewska et al., 2016), the Polish adaptation of the Experiences in Close Relationships–Revised: 36 items, 7-point Likert, two scales (attachment anxiety, items 1–18; avoidance, items 19–36). Polish norms: anxiety $M = 3.30$, $SD = 1.21$; avoidance $M = 3.04$, $SD = 0.96$. Categorical attachment style is derived from the two scales with a

threshold of 4 on each—the midpoint of the 7-point response scale—(secure: both low; anxious-preoccupied: high anxiety only; dismissive-avoidant: high avoidance only; fearful-avoidant: both high); the dimensional analyses do not depend on this categorization. The labels *disorganized* and *fearful-avoidant* are scored as equivalent throughout: the four-category self-report model has no disorganized cell, and its fearful-avoidant quadrant (high anxiety with high avoidance; Bartholomew & Horowitz, 1991) is the conventional self-report counterpart of disorganized phenomenology. This is an operational scoring decision; Supplement S6 reports both a stricter coding and a sensitivity analysis that drops the four disorganized-labeled biographies entirely.

MentS-PL (Jańczak, 2021), the Polish adaptation of the Mentalization Scale: 28 items, 5-point Likert, three subscales (self-mentalization, other-mentalization, motivation to mentalize). Polish norm for the total: $M = 105$, $SD = 13.7$.

KPP (Matusz et al., 2011), the Polish Need for Cognition Questionnaire: 36 items, 5-point Likert, single scale; norm $M = 3.74$, $SD = 0.49$ (mean per item). KPP and TIPI-PL Openness are treated as separate dimensions in all analyses.

TIPI-PL (Sorokowska et al., 2014), the Polish Ten-Item Personality Inventory: 10 items, 7-point Likert, five Big Five subscales of two items each, with published Polish norms.

TCTM-22 is an author-keyed vignette test of *text-mediated subtext recognition*: 22 chat-conversation vignettes (workplace, romantic, family, and friendship contexts across messaging platforms), each followed by four response options, one author-keyed correct. The three distractors per item are tagged with categories adapted from the MASC error taxonomy (Dziobek et al., 2006): undermentalizing (literal-comprehension readings, DOS), overmentalizing (over-attributive readings, NAD), and no mentalization (BK). A central caveat applies throughout: the key is author-defined and has not been adjudicated by an independent panel, so "TCTM-22 agreement" always means agreement with the author key—a stimulus-fidelity quantity, not a validated mentalization measurement. The full vignette text and distractor commentary are included in the release package (Wiencek, 2026a). The 22 vignettes are drawn from a larger 57-vignette pool; an extended-battery

administration of the full pool is used in Section 3.6.

All scale scores are expressed both on their raw metrics and as z-scores against the Polish norms above; nine z-dimensions are used in profile-level analyses (DBZ-R anxiety and avoidance, MentS-PL total, KPP, and the five TIPI-PL scales).

2.3 Models, platforms, and conditions

Seven instruction-tuned LLMs from four vendors were administered the battery: Claude Sonnet 4.6 and Claude Opus 4.6 (Anthropic; served via AWS Bedrock), GPT-5.4-mini and GPT-5.5 (OpenAI; Azure OpenAI Chat Completions), GPT-5.4 (full) (OpenAI; Azure AI Foundry Responses endpoint), Grok-4-20-reasoning (xAI; Azure AI Foundry), and Gemini 3 Flash (Google; Gemini API, an endpoint explicitly labeled *preview*). The OpenAI trio gives a within-vendor structure matched on vendor-designated size class (parameter counts are undisclosed): GPT-5.4-mini is marketed as a small-class model, GPT-5.4 (full) and GPT-5.5 as flagship-class consecutive deployments. The panel is a convenience sample of frontier vendor offerings within budget; no sampling frame over "LLMs in general" is claimed.

Three conditions were run per model. In the *persona* condition the narrative body of the biography (never the target header; Section 2.1) is placed in the system prompt with an instruction to answer as the person described; each persona is administered twice per collection, giving a test–retest pair. In the *baseline* condition the system prompt instructs the model to answer honestly as itself (an AI language model), without roleplay. In the *zero-prompt* condition there is no system prompt at all; only the battery appears as the user message. Models respond in a structured JSON format; responses are scored deterministically by the accompanying code, with a permissive recognizer for the spontaneous key-naming variants that models produce in the zero-prompt condition.

Sampling parameters were left at vendor defaults in all collections, except a generous output-token ceiling (16,000). For Azure-served models in the initial collection the request code additionally passed `temperature` = 1.0, `top_p` = 1.0, a fixed seed, and medium reasoning effort where the endpoint accepted them; by the corrected-stimulus collection the

Foundry endpoint rejected sampling parameters outright (HTTP 400), so those runs are vendor-default by construction. Bedrock and Gemini requests passed only the token ceiling in all collections. The deliberate choice of vendor defaults reflects the ecological realism of default deployment use rather than cross-vendor comparability (different vendors' defaults differ), at the cost of not controlling the decoder—a trade-off revisited in Sections 3.3 and 5.

An informal human sanity check ($N = 7$; Polish friends-and-family convenience sample) completed the full five-instrument battery in April, i.e. on the *pre-correction (truncated) rendering*—the same battery version as the initial collection; its comprehensibility conclusion therefore covers that version, and results on the four correction-affected items are interpreted accordingly (Section 3.5). Its sole purpose was to verify that the vignettes are comprehensible and answerable for live human respondents—it is not a pilot study in the psychometric sense and does not establish a comparative norm. Only the TCTM-22 scores are used, descriptively, as a small human reference point; the sample enters no inferential comparison, and its data are released in aggregate form only (summary statistics and per-item agreement counts).

2.4 Stimulus rendering: verification and correction

TCTM-22 vignettes are authored as structured conversation objects and serialized into the prompt text by a rendering pipeline. During the study I discovered that an early version of the serializer silently dropped conversation messages whose metadata fields appeared in non-canonical order. Three vignettes were affected: in `s07` the dropped message was precisely the reply that the question stem asks about; in `w19` the dropped block contained eight escalation messages preceding the keyed line; in `pw07` the dropped messages again included the phrase referenced by the stem. Independently, the stem of `w22` referred to its target message by ordinal position ("the fifth message") while the rendered conversation placed that message tenth; the stem was corrected to reference the message by content, leaving the keyed option unchanged. The vignettes themselves were sound; the defect lay in the rendering pipeline. Throughout, s07, w19, and pw07 are called *rendering-affected*

Table 1: Data-collection events and scored run counts. The corrected-stimulus re-collection is the primary dataset for all analyses; the initial collection enters only the before–after correction comparison (Section 3.5), the longitudinal comparison (Section 3.4), and the deployment-window determinism analysis (Section 3.3).

Collection	Dates (2026)	Stimulus	Scored runs
Initial	Apr 12–May 11; May 28	truncated	persona 426; baseline 149 (incl. expanded $N = 31$ GPT-5.4-mini, $N = 31$ GPT-5.4 (full), $N = 50$ GPT-5.5); zero-prompt 43
Corrected re-collection	May 31 (Azure); Jun 10–11 (non-Azure)	corrected	persona 424 (30 personas × 2 runs × 7 models, plus retained retries); baseline 70 (10 × 7); zero-prompt 44
Extended battery (57 vignettes; Sonnet, GPT-5.5)	Apr (initial); Jun 11 (corrected)	both versions	61 initial; 62 corrected
Human sanity check	Apr	—	7

items and w22 a *stem-corrected* item.

After the correction, the entire design was re-collected on the corrected rendering, and the corrected prompts were verified byte-for-byte identical across platforms before any call was issued. The corrections touched only the TCTM-22 vignette section of the prompt; the self-report battery text is identical across collections, which makes the initial-versus-corrected self-report comparison a deployment-drift probe rather than a correction effect (Section 3.4).

2.5 Data collections

Data were collected between April 12 and June 11, 2026, in three logical phases (Table 1): an *initial collection* on the truncated rendering (April 12–May 11 for six models; May 28 for GPT-5.4 (full), which joined the panel late, with concurrently expanded baseline samples for the OpenAI models); a *corrected-stimulus re-collection* (May 31 for the four Azure-served models; June 10–11 for the Bedrock and Gemini models); and an *extended-battery administration* of the full 57-vignette pool to Claude Sonnet 4.6 and GPT-5.5 (initial version April; corrected version June 11). In the release-package data every run is tagged with its collection event.

Unless explicitly stated otherwise, every statistic in this paper is computed on the

corrected-stimulus re-collection. Within it, persona-level scores are the mean of a persona's administrations within the collection; analyses that require one row per persona (the classification matrix, Cochran's Q) use the first administration; item-level proportions pool all persona administrations (typically $n = 60$–61 per model). Retained retry runs give one persona a third administration in the four Azure panels, so pooled proportions weight runs rather than personas; persona-level analyses average within persona first, and run-weighted versus persona-weighted TCTM-22 means differ by at most 0.015 items. The initial collection is never mixed into primary cells.

2.6 Pre-specified versus exploratory analyses

The study was not preregistered. Four hypotheses were recorded in working notes before data collection: (H1) DBZ-R-derived attachment-style classifications will agree with the author-declared persona labels at $\kappa \geq .50$ for most models; (H2) TCTM-22 test–retest will reach $r \geq .50$ for most models; (H3) cross-vendor agreement on persona-conditioned z-profiles will be high (median $r \geq .70$); (H4) the paradoxical persona `ola` will degrade TCTM-22 agreement in all models, because models adapt to the biography rather than acting as answer oracles. The intercept/slope decomposition, the determinism and drift analyses, the before–after correction comparison, and the administration-context comparison are exploratory; the latter two were prompted by mid-study quality control rather than planned.

2.7 Statistical approach

Proportions are reported with Wilson 95% CIs; classification agreement with Cohen's κ (per model versus the author labels; percentile bootstrap CIs over personas, $B = 5{,}000$) and Fleiss κ (across the seven-model panel; bootstrap over personas, $B = 2{,}000$). Because the seven models are not independent raters (shared training corpora, shared vendor lineages), Fleiss κ is reported as *descriptive panel agreement*, not as chance-corrected agreement of an independent rater pool. All bootstrap intervals quantify resampling stability over the curated stimulus corpus and should not be interpreted as population-sampling intervals

for arbitrary persons, biographies, or models. Profile-level agreement is quantified three ways, because they answer different questions: Pearson correlations over the 30 personas (linear profile agreement, insensitive to additive level and positive multiplicative gain), Lin's concordance correlation coefficient (CCC; penalizes level and gain differences), and ordinary least-squares regressions of each model's per-persona profile on the leave-one-out consensus of the remaining six models, which estimate an explicit intercept (a, the model's persona-conditioned level offset) and slope (b, its gain on the shared persona signal) per dimension. Cross-model item heterogeneity uses Cochran's Q with Benjamini–Hochberg FDR correction across items. Two-group contrasts report Hedges' g with bootstrap CIs; where one cell shows zero within-group variance, the pooled standardizer reduces to the other group's spread and a standardized effect would be misleading, so the raw contrast is reported instead. At $N = 30$ the design detects (at $\alpha = .05$, power $= .80$) approximately $|r| \geq .49$ against zero; no closed-form power statement is attempted for κ, whose power depends on category marginals, and bootstrap CIs are reported instead; cells with $n < 15$ (zero-prompt, human sanity check) are descriptive only, and the baseline condition ($n = 10$ per model in the corrected collection) is treated as estimation-only: effect sizes with bootstrap intervals are reported, but no hypothesis tests are attached to baseline contrasts. The classical power statement above is a heuristic; because the corpus is curated rather than sampled, the reported bootstrap intervals—which quantify resampling stability of this stimulus set—are the primary uncertainty summaries. Given the study's descriptive aims, estimation is emphasized over null-hypothesis testing throughout. All analyses are reproducible from the release-package data files by the accompanying scripts (Supplement S1 lists the files and maps every table and figure to its generating code).

3 Results

3.1 Stimulus fidelity: seven models agree with the author-declared targets

Table 2 summarizes per-model agreement with the author-declared targets in the persona condition. Attachment-style classification matches the author label for 23–29 of 30 personas per model (Cohen's κ = .69–.96); Fleiss κ across the seven-model panel is .853 (bootstrap 95% CI [.75, .94]), and excluding any single model changes the value only in the second decimal. These figures use the disorganized-equals-fearful-avoidant scoring convention (Section 2.2); four biographies carry the disorganized label, and under strict four-class coding—in which the threshold-derived model classifications can never match that label—matches drop to 21–26 of 30, and excluding those four biographies outright gives 21–26 of 26 (κ = .74–1.00; Supplement S6). The classification is also robust to the operational scale-midpoint cut: re-deriving styles at thresholds of 3.5–4.5 on the 1–7 anxiety and avoidance means keeps every model's matches within 20–29 of 30 and κ within .56–.96 (at the extreme cut of 4.5; $\kappa \geq .69$ for cuts through 4.25; table `style_threshold_sensitivity.csv` in the release). The median per-dimension correlation between author-declared ordinal targets and model-produced z-scores is .79–.85, with every model inside that band; because the ordinal levels were mapped to numbers for this analysis, Spearman versions were computed as a coding-free check and fall in the same range (ρ = .74–.83). Twenty-one of 30 personas are classified in agreement with the author label unanimously by all seven models (Figure 1). Under the demographic-prompt regime, Petrov et al. (2024) report poor individual-level fidelity; under dense narrative conditioning, all seven models—including the small-class GPT-5.4-mini—exceed the pre-specified κ = .50 threshold. H1 is supported on the corrected data; the verdict on H3 (cross-vendor agreement) is deferred to Section 3.2, where the full cross-model agreement results are presented.

Where the panel disagrees with the author, it disagrees coherently. Three personas account for most mismatches, and in each case the seven models converge on the *same*

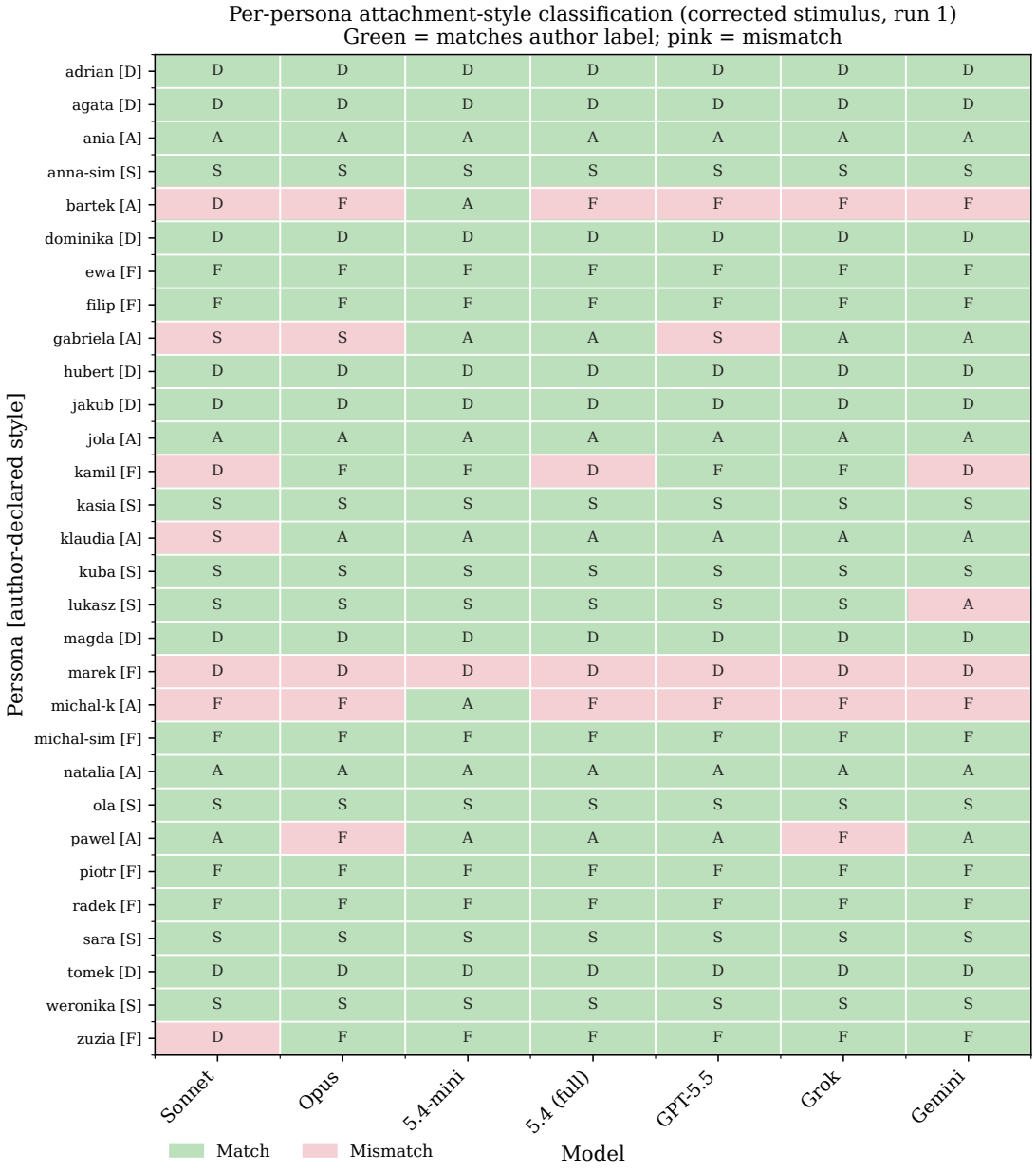

Figure 1: Per-persona attachment-style classification (corrected collection, first administration). Green cells match the author-declared style; pink cells do not. Letters give the produced category (S = secure, A = anxious-preoccupied, D = dismissive-avoidant, F = fearful-avoidant).

Table 2: Stimulus-fidelity summary, persona condition, corrected collection. Style = personas whose first-administration classification matches the author label (of 30), with Wilson 95% CI; κ = Cohen's kappa versus author labels (bootstrap 95% CI); Author–model r = median across nine z-dimensions of the per-dimension Pearson correlation between author ordinal targets and per-persona mean z-scores ($n = 30$); TCTM-22 = mean ± SD of author-key agreement (of 22) across persona runs.

Model	Style (/30) [CI]	κ [CI]	Author–model r	TCTM-22
Claude Sonnet 4.6	23 [.58, .89]	.69 [.48, .87]	.84	20.97 ± 0.61
Claude Opus 4.6	25 [.66, .93]	.78 [.59, .95]	.85	19.15 ± 1.73
GPT-5.4-mini	29 [.83, .99]	.96 [.86, 1.00]	.79	18.26 ± 1.64
GPT-5.4 (full)	26 [.70, .95]	.82 [.64, .96]	.82	20.28 ± 1.42
GPT-5.5	26 [.70, .95]	.82 [.64, .96]	.82	20.79 ± 1.05
Grok-4-20	26 [.70, .95]	.82 [.64, .96]	.80	20.00 ± 2.74
Gemini 3 Flash	25 [.66, .93]	.78 [.59, .95]	.82	19.32 ± 2.07
Human sanity check ($N = 7$)	—	—	—	14.29 ± 1.38

alternative reading rather than scattering: `marek` (author: fearful-avoidant) is read as dismissive-avoidant by the panel (0/7 author matches)—the biography evokes avoidance without enough anxiety signal to cross the categorical threshold; `bartek` and `michal-k` (author: anxious-preoccupied; 1/7 each) are both read as fearful-avoidant—both biographies wrap an anxious core in a socially withdrawn surface, and the questionnaire responses track the surface. When several non-independent models converge on a category other than the intended one, this pattern suggests a biography-design diagnostic rather than joint model failure; Section 4.4 returns to this use of the panel.

The paradox persona behaves as designed (H4). All seven models classify `ola` as secure, and all seven simultaneously degrade on her TCTM-22 items relative to the mean of the other 29 personas (deltas −1.5 to −12.9 points of 22; e.g., Grok 7.5 versus 20.4 for the rest). The models adopt the biography's low-mentalization stance at the cost of test agreement instead of answering as capable oracles—particularly direct evidence that narrative conditioning controls responding rather than merely re-labeling it. This is the constructive counterpart of a capacity that Westwood (2025) frames as a survey-integrity threat: an autonomous LLM respondent conditioned on a persona passed 99.8% of 6,000 standard attention-check trials. Here, the same capacity is used diagnostically—the biography, rather than an answer key, governs the response.

Agreement does not reduce to trait-keyword detection. Restricting the cross-model

Table 3: Baseline-condition intercepts on the corrected collection ($n = 10$ runs per model): mean z-scores against Polish population norms (SD in parentheses). Boldface marks the extremes of the three dimensions with the largest between-model spread (avoidance, extraversion, agreeableness).

Dimension	Sonnet	Opus	5.4-mini	5.4 (full)	GPT-5.5	Grok	Gemini
DBZ-R Anx	−0.99 (.03)	−0.89 (.18)	−1.33 (.32)	−1.51 (.75)	−1.68 (.08)	−1.28 (.24)	−1.89 (.04)
DBZ-R Avo	**+1.10** (.28)	−0.26 (.19)	−0.02 (.33)	+0.53 (1.07)	+0.39 (.36)	**−0.64** (.47)	−0.28 (2.15)
MentS total	+1.23 (.11)	+1.75 (.37)	+1.90 (.34)	+1.31 (.28)	+1.66 (.13)	+1.93 (.26)	+2.26 (.34)
KPP mean	+2.04 (.10)	+1.92 (.08)	+2.50 (.13)	+2.41 (.09)	+2.34 (.11)	+2.30 (.21)	+2.56 (.04)
TIPI E	−1.18 (.40)	−0.41 (.00)	−1.18 (.64)	−1.18 (.78)	−0.49 (.42)	**−0.08** (.44)	**−1.62** (.76)
TIPI A	+0.02 (.22)	**−0.24** (.00)	+0.53 (.18)	+0.87 (.36)	+0.62 (.00)	+0.49 (.29)	**+1.26** (.36)
TIPI C	−0.06 (.16)	+0.50 (.18)	+0.67 (.26)	+1.02 (.26)	+0.88 (.17)	+0.81 (.25)	+1.27 (.22)
TIPI ES	+0.67 (.14)	+0.63 (.16)	+1.21 (.17)	+1.69 (.00)	+1.69 (.00)	+1.40 (.28)	+1.69 (.00)
TIPI O	+0.91 (.14)	+1.31 (.00)	+1.13 (.31)	+1.26 (.25)	+1.18 (.21)	+0.87 (.36)	+1.44 (.36)

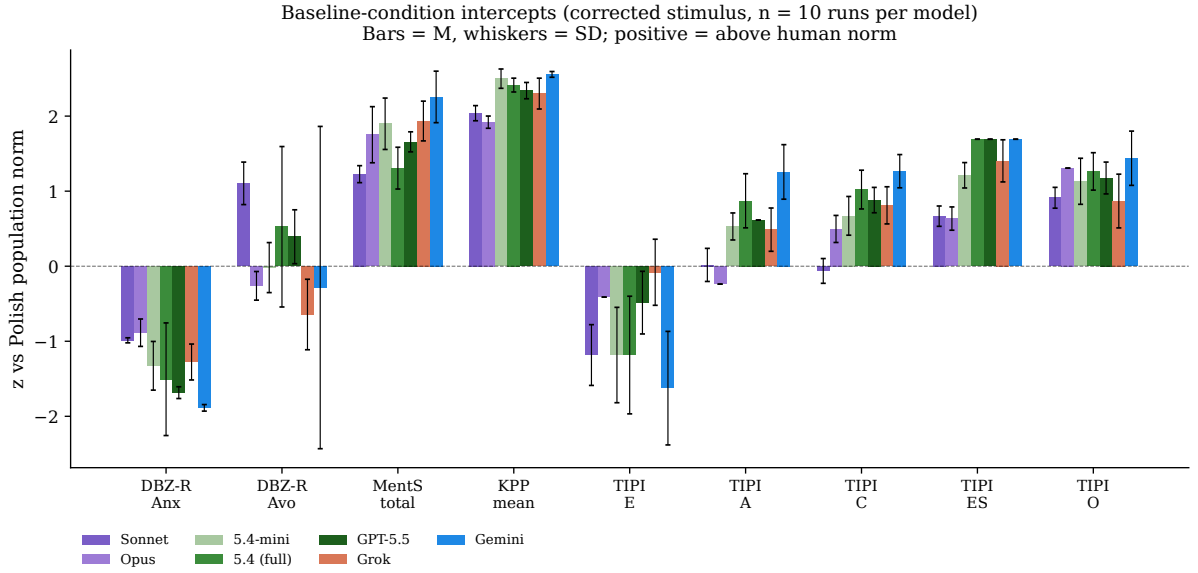

Figure 2: Baseline intercepts per model across the nine z-dimensions (corrected collection; bars = M, whiskers = SD, $n = 10$). All models share a low-anxiety, high-mentalization, high-need-for-cognition profile far from the human norm, with model-specific offsets of up to 1.7 SD.

agreement analysis (below) to the 21 narrative-only biographies yields a median pairwise correlation of .957, versus .972 on the nine explicit-label biographies; the flagship OpenAI pair reaches .988 versus .991. Dense narrative alone sustains the agreement.

3.2 Baseline intercepts diverge; persona slopes converge

The shared baseline profile replicates social-desirability findings on Polish instruments. When answering as themselves, all seven models report attachment anxiety far below the Polish norm (z between −0.89 and −1.89), mentalization +1.2 to +2.3 SD above it, need for cognition +1.9 to +2.6, and elevated emotional stability and openness

(Table 3, Figure 2). The direction and size of these elevations are directionally consistent with the social-desirability bias Salecha et al. (2024) documented on English Big Five instruments, extending the pattern to Polish-normed instruments across four vendors (a conceptual, not direct, replication—instruments and prompts differ).

On top of the shared profile, models differ by up to 1.7 SD. The largest between-model spreads are on avoidance (Sonnet +1.10 versus Grok −0.64), extraversion (Grok −0.08 versus Gemini −1.62), and agreeableness (Opus −0.24 versus Gemini +1.26). Anyone treating an unconditioned model as a respondent inherits these offsets.

Persona orderings converge across all seven models. Correlating per-persona mean z-profiles between every model pair and taking the median across the nine dimensions yields the matrix in Figure 3: all 21 pairwise medians lie in [.947, .989]; the persona-bootstrap 95% CI for the panel minimum is [.91, .95]. The highest pair is the within-vendor flagship pair GPT-5.4 (full) ↔ GPT-5.5 at .989 (CI [.98, .99]; per-dimension range .94–.99). These medians summarize dimension-level correlations that individually range down to .75 (Opus–mini on the weakest dimension), and rank-based Spearman versions tell the same story (panel-minimum median $\rho = .91$; median $|\rho - r|$ across pairs = .03). Seven models that disagree by up to 1.7 *SD* about themselves produce highly similar relative orderings of 30 fictional people. H3 (cross-vendor pairwise agreement $r \geq .70$) is supported, comfortably, on the corrected data.

Estimating the slopes explicitly. Correlations are insensitive to level and gain, so the decomposition is additionally estimated as a regression: each model's per-persona profile, dimension by dimension, on the leave-one-out consensus of the other six models. The shared persona signal accounts for most variance in every model (median $R^2 =$ .94–.97), the fitted level offsets are small (median $|a| = .04$–$.20$ z), and the fitted gains cluster around unity but are not identical: median b runs from 0.84 (Sonnet, GPT-5.4-mini) through 0.91–1.03 (Opus, Grok) to 1.08–1.11 (GPT-5.4 (full), GPT-5.5, Gemini)—some models compress the persona signal, others amplify it. Consistently, absolute agreement is somewhat lower than linear profile agreement: pairwise Lin's CCC, which penalizes level and gain differences, has medians of .86–.98 (individual dimension-level values down

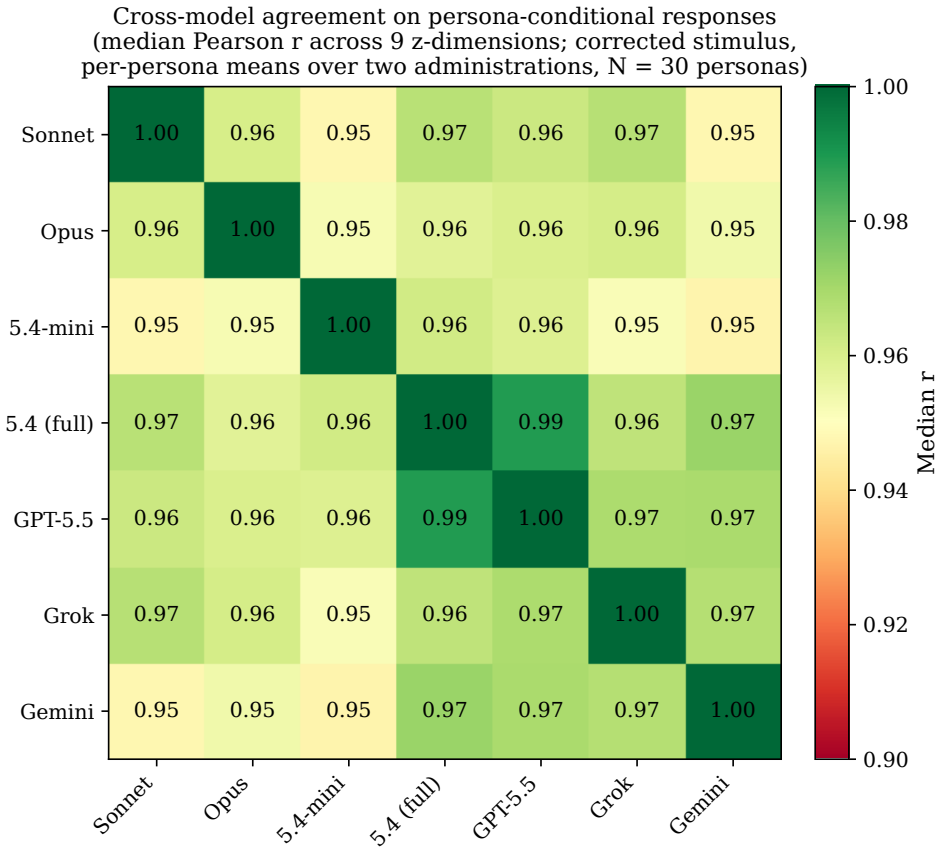

Figure 3: Persona-profile agreement matrix: pairwise median Pearson r across nine z-dimensions (corrected collection, per-persona means over two administrations, $N = 30$ personas). All 21 between-model values exceed .94.

to .62) against the Pearson .95–.99. The convergence claim is therefore precise: models produce highly similar relative *orderings* of personas, and only approximately similar absolute scale values.

This is the central decomposition of the paper: the *intercept* (the level component—where a model sits when answering as itself, and the residual level offset it adds under conditioning) is model-specific and, as the next two sections show, unstable in several distinct ways; the *slope* (the profile component—how the model orders and scales personas around that level) is shared across vendors, size classes, and—as Section 3.4 shows—time. One scope note: the decomposition is an organizing description with two operationalizations (the separate baseline condition measures the level directly; the consensus regression estimates level and gain within the persona condition), not a single fitted model linking baseline and persona responding—a hierarchical formalization is left to future work. A related caveat: the leave-one-out consensus predictor itself carries measurement error, which biases OLS gain estimates toward attenuation, so the reported gains are conservative summaries of alignment with the shared signal rather than error-free structural parameters. Within the OpenAI trio the decomposition has a within-vendor corollary: on the corrected collection the two flagship deployments show no detectable baseline contrast on avoidance (Hedges' $g = 0.16$, bootstrap 95% CI $[-0.71, +1.38]$—an interval too wide to claim equivalence at $n = 10$ per cell) and nearly identical slopes, and under persona conditioning the avoidance contrast is $g = 0.007$ (persona-clustered bootstrap 95% CI $[-0.51, +0.53]$). The dramatic within-vendor baseline contrasts that *did* appear in this study were bound to a deployment window, as the next section documents.

3.3 Intercept instability I: deterministic response locks within a deployment window

The initial collection contained a striking event. All 31 GPT-5.4 (full) baseline runs—collected on one day through the Foundry endpoint—returned byte-distinct raw responses (distinct hashes, varying token counts and latencies) that nevertheless scored *identically* on three of five summary metrics: anxiety ($M = 1.00$, $SD = 0$ across 18×31 Likert

responses), TCTM-22 total (19/22 in every run), and need for cognition (a byte-identical 36-item answer vector in all 31 runs, confirmed by independent re-extraction from the raw response payloads; $M = 4.917$, $SD = 0$, under the published scoring key with its 14 reverse-scored items). Avoidance, answered within the same instrument in the same responses, carried ample variance ($SD = 1.39$)—and its 31 values form two clusters separated by an empty interval ([3.50, 4.06]) that happens to straddle the avoidance threshold of the attachment-style classification rule (Figure 4): 13 runs classify the model's baseline as low-avoidance, 18 as high-avoidance. A two-component mixture is favored descriptively (BIC 108.8 versus 114.4; a parametric-bootstrap likelihood-ratio test is reported in the Supplement); at $N = 31$, and as one cell selected from a wide post-hoc search space, the test carries no confirmatory weight and the bimodality is treated as strictly descriptive.

Five mechanisms were tested against forensic evidence (request/response artifacts, live endpoint probes, a same-platform control): seed pinning (the endpoint rejects seeds), response caching (cache-buster probes returned identical option choices on fresh requests; all raw texts are byte-distinct), silent greedy decoding (free-text portions vary), scoring-code error (independent rescoring reproduces the locks), and platform-level effects (Grok, served from the same platform, shows variance on every metric). The remaining account is *response-level determinism on constrained items*: the deployment sampled freely over surface text while converging on identical option choices for most constrained psychometric items, with avoidance the single within-instrument channel where item-level uncertainty survived to the output. This is a behavioral characterization: variable free text shows that *some* sampling occurred, but API access cannot identify the decoding mechanism behind the convergent option choices. Supplement S4 reports the full evidence trail.

Crucially, *the lock did not replicate three days later.* In the corrected-stimulus collection on the same endpoint—which by then had begun rejecting explicit sampling parameters, evidencing an endpoint-side reconfiguration—the same model's baseline carried variance on every metric (anxiety $SD = 0.91$; TCTM-22 $SD = 0.42$; KPP five distinct values, $SD = 0.045$), and its avoidance mean moved from 4.32 to 3.54. The determinism,

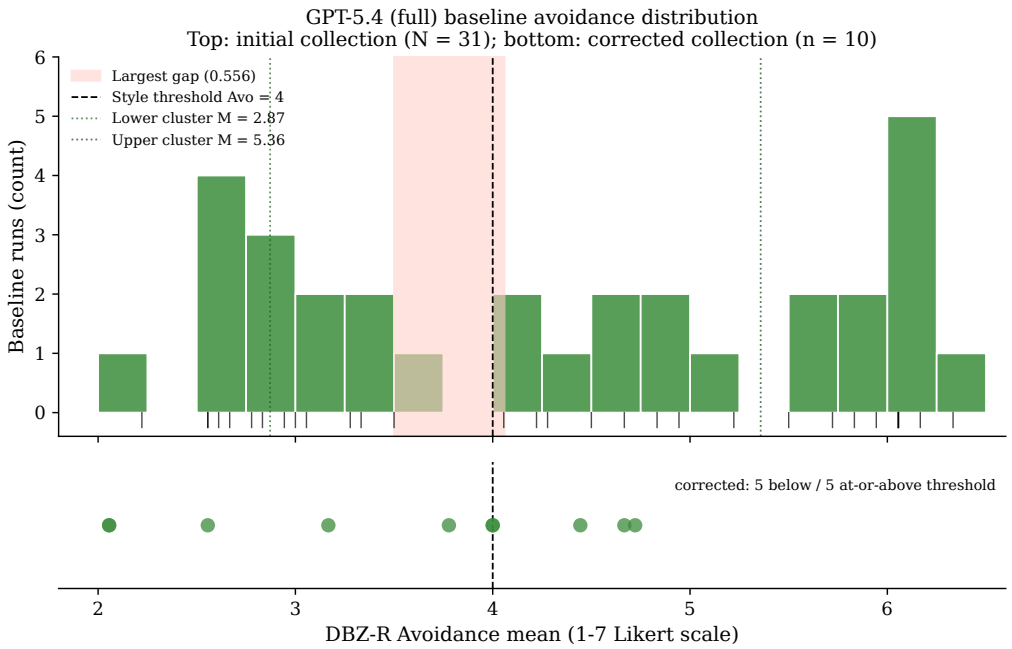

Figure 4: GPT-5.4 (full) baseline avoidance. Top: initial collection ($N = 31$); the distribution splits into two clusters around the attachment-classification threshold, while anxiety, TCTM-22, and need for cognition from the same responses are deterministic. Bottom: corrected collection three days later ($n = 10$); variance has returned to all metrics and the values continue to straddle the threshold (5 below, 5 at-or-above).

the bimodality, and the large within-vendor baseline contrast it produced (avoidance Hedges' $g = 1.14$ versus GPT-5.5, bootstrap 95% CI [0.65, 1.74], in the initial window—against $g = 0.16$ for the same contrast on the corrected collection, Section 3.2) were therefore observed in one deployment-window/request-configuration combination and did not reproduce in the subsequent collection; with API access alone the responsible element (model snapshot, routing, endpoint schema, rejected parameters, or implicit defaults) cannot be isolated. Lesser locks of the same kind recur across the panel and persist in the corrected collection: Claude Sonnet's baseline TCTM-22 total locks at 21/22 in 10 of 10 runs, Claude Opus at 20/22 in 10 of 10, Opus additionally locks its zero-prompt TCTM-22 at 21/22, GPT-5.5 answers the zero-prompt TCTM-22 perfectly (22/22) in all runs in both collections, and several models freeze single TIPI-PL scales at one value across all 10 baseline runs (e.g., three Opus scales; full catalogue in Supplement S4). Repeated administration of an unconditioned LLM does not generally yield an i.i.d. sample; it can yield one answer, repeated.

3.4 Intercept instability II: changes between collections land on the level axis

Because the full design was re-collected after the rendering correction, every model has two complete collections separated by three days (GPT-5.4 (full)), roughly seven weeks (the remaining Azure models), or eight weeks (the Bedrock and Gemini models)—with a byte-identical self-report battery throughout. Comparing them asks: what moved, the levels or the profiles?

Orderings: essentially nothing moved. Correlating per-persona means across the two collections, dimension by dimension, yields median r of .92–.99 per model (Sonnet .994, Opus .991, GPT-5.5 .994, GPT-5.4 (full) .988, Gemini .980, Grok .973, GPT-5.4-mini .916; Figure 5a). For six models the concordance version of the same comparison, which is additionally sensitive to level and gain changes, is just as high (median CCC .97–.99): for them the biography-to-profile mapping replicates essentially point-for-point up to two months apart—evidence that, for these models and this corpus, the profile-ordering results did not decay on the timescale of weeks (the levels did, as shown next; and with two time points per model the design can document change, not its trajectory).

Levels: two models moved, one of them within days. GPT-5.4-mini's entire baseline profile shifted—and the collection structure bounds *when*. Its 21 late-May baseline runs still show the April profile (anxiety $M = 4.69$ versus April's 4.68), while the corrected collection three days later shows a different model: anxiety down 3.0 raw Likert points (to 1.69; z +1.14 to −1.33—previously the only model with above-norm anxiety), mentalization up 28 points on the 28–140 scale, need for cognition up 0.92 on a 5-point scale, and emotional stability, openness, and conscientiousness up 0.9–1.6 points (Figure 5b). This is not gradual drift but an abrupt collection-window difference inside three days—the same late-May interval in which a sibling endpoint observably reconfigured (Section 3.3). The shift propagated to classification behavior, and in a way that sharpens the decomposition: before the change the model labeled almost every baseline run anxious-preoccupied and matched the author style on only 16 of 30 personas, the panel outlier; after it, 29 of 30—the best in the panel—so its persona-style agreement

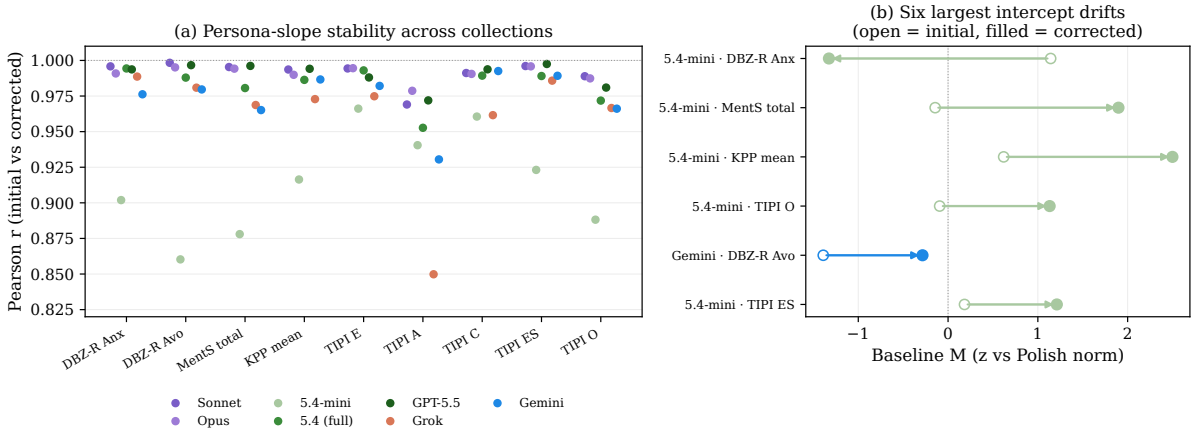

Figure 5: Initial versus corrected collection. (a) Per-dimension correlation of per-persona means across collections ($N = 30$ personas per point): orderings replicate at $r \geq .85$ everywhere, with per-model medians .92–.99. (b) The six largest baseline-level moves (of 63 model × dimension cells), in z-units: five belong to GPT-5.4-mini (bounded to a three-day window by its late-May baseline runs), the sixth is Gemini's avoidance, whose *SD* simultaneously tripled.

across collections is only 15/30, the panel's lowest. Those flips are level effects: the style rule cuts the anxiety and avoidance scales at fixed thresholds, so when the persona-conditioned levels moved with the baseline, classifications crossed the cutoffs even though the persona *ordering* held ($r = .92$ with its own earlier profiles, $\geq .94$ with every other model). The concordance diagnostic isolates the same dissociation: mini's cross-collection CCC drops to .56 on the dimension that moved most (avoidance) against r of .92 on the same comparison—ordering intact, level shifted. Gemini 3 Flash—the one panel member on an explicitly preview endpoint—moved on two dimensions across its eight-week gap (avoidance +1.06 raw points with *SD* inflating from 0.76 to 2.06; extraversion −1.30) while its orderings held at .980. The remaining five models' baselines replicated within noise, including across the three-day GPT-5.4 (full) window (largest move: MentS −5.8 on the 28–140 scale).

Strictly, initial-versus-corrected comparisons confound elapsed time and deployment updates with the rendering correction itself. Three observations make the correction an implausible cause of the baseline moves: the self-report battery text is identical across collections; the three-day GPT-5.4 (full) comparison brackets the correction and shows only small moves; and the large moves concentrate in one small-class model and one preview endpoint, with five models stable across the same prompt change.

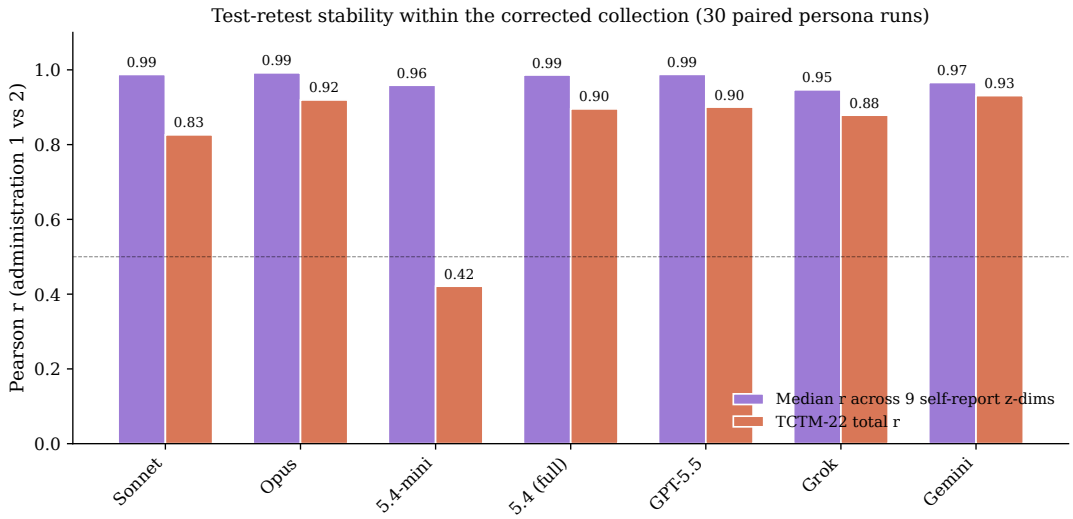

Figure 6: Test–retest within the corrected collection (administration 1 versus 2, 30 paired persona runs per model).

Within-collection test–retest. Within the corrected collection, the two administrations per persona agree at median $r = .95$–$.99$ across the nine z-dimensions for every model; TCTM-22 totals retest at $r = .83$–$.93$ except GPT-5.4-mini (.42); style classifications agree on 28–30 of 30 personas per model (Figure 6). Because r tracks ordering only and is fragile near a ceiling, absolute-agreement companions are reported alongside: TCTM-22 CCC is .82–.92 (GPT-5.4-mini .42), the mean administration-1-versus-2 difference is at most 0.30 points of 22 for every model, and the mean absolute difference is 0.13–1.33 points—so even GPT-5.4-mini's low r reflects reshuffling within a 1–2-point band near its ceiling, not large score changes; median per-dimension z-score CCC is $\geq .95$ for all models. H2 is supported for six of seven models. Dimension-specific coefficients are reported in Supplement S9. Avoidance—the scale that was deterministic and threshold-straddling at baseline in Section 3.3—retests under persona conditioning at $r = .958$–$.998$ across the seven models. Thus, instability of one baseline level coexists with a nearly perfectly reproducible ordering of the 30 personas on the same scale.

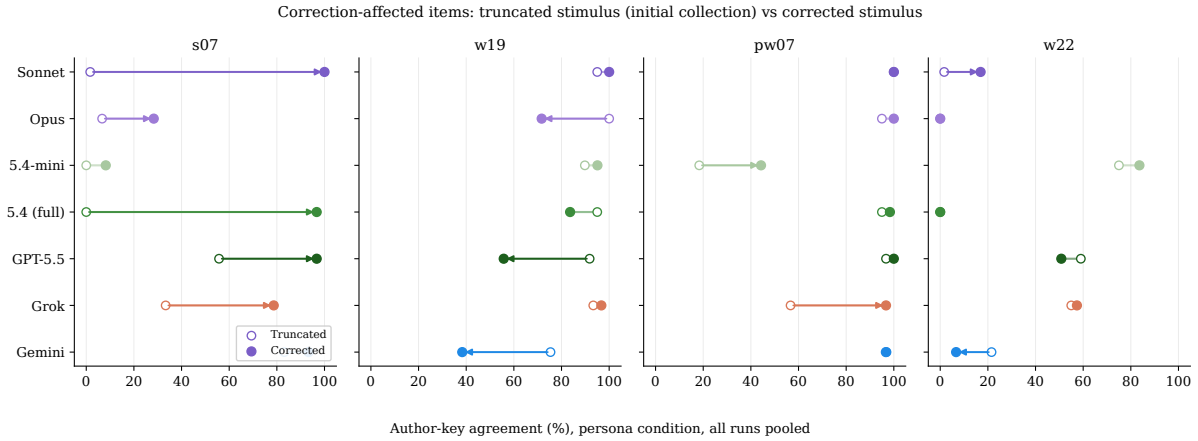

Figure 7: Author-key agreement on the four correction-affected items, initial (truncated; open markers) versus corrected (filled markers) collection, persona condition pooled over administrations. Across the correction, agreement rises by up to 98 points (s07), falls by up to 37 (w19), or stays at a deterministic floor (w22).

3.5 A before–after comparison across the correction: items swing, aggregates hold

Because the initial collection ran on the truncated rendering and the re-collection on the corrected one, the four affected items received a before/after comparison across all seven models (Figure 7). One caveat governs the whole section: this is an uncontrolled before/after design—elapsed time, endpoint configuration, and possible deployment updates co-vary with the correction—so the differences below are *associated with* the correction rather than experimentally attributed to it. Two observations bound the confounding: the 18 untouched items moved by a median of at most 2 percentage points per model across the same comparison (single-item tails up to 27 points; full distribution in the release-package tables), and the three-day GPT-5.4 (full) bracket shows that collections days apart agree closely on untouched material. GPT-5.4-mini is excluded from this attribution logic entirely, since its deployment changed wholesale between the collections (Section 3.4).

With the stem-referenced line restored, s07 recovers—for most models. On the truncated rendering, the s07 stem asked about a reply the model never saw; agreement was at or near floor for four models. On the corrected rendering, Claude Sonnet moves from 1.7% to 100% (+98.3 points, the largest single swing in the dataset),

GPT-5.4 (full) from 0% to 96.7%, Grok +45, GPT-5.5 +41. Two models stay low on the *identical* corrected stimulus: GPT-5.4-mini (8.2%) and Claude Opus (28.3%)—a residual literal-comprehension cluster spanning two vendors and two size classes, which rules out a pure size account. The item asks whether a polite Polish workplace hedge ("Pewnie tak, jak uznasz") signals disengagement; why exactly these two models read it charitably while five siblings do not is a question for cross-lingual follow-up (Section 6).

With context restored, w19 gets harder. Agreement *drops* on the corrected rendering for Gemini (−37 points), GPT-5.5 (−36), Opus (−28), and GPT-5.4 (full) (−11): with only the final six messages visible, the keyed reading (resignation after escalation) was the salient one; with the full fourteen-message arc restored, models distribute mass across competing plausible readings. Truncation had accidentally *simplified* the item—apparent accuracy can be an artifact of a stimulus defect in either direction.

The stem-corrected w22 remains contested. After the stem fix, two flagships from two vendors sit at a deterministic 0% (GPT-5.4 (full) 0/61; Opus 0/60—every call selecting the same non-keyed option), while GPT-5.4-mini reaches 83.6%; the human sanity-check sample, which answered the original pre-correction stem, reached 57%. Item w22 is treated as having a contested key pending independent adjudication (Section 3.6 shows the contest is administration-bound): no inferential claim in this paper rests on its item-level agreement, it contributes at most one point to the descriptive 22-item totals in which it remains, and the model ordering replicates with all four affected items removed (Supplement S6).

Aggregates barely notice any of this. Across the same before/after comparison, per-persona TCTM-22 totals move by at most +1.56 points of 22 (per-model paired means −0.27 to +1.56), and the profile-*ordering* results of Sections 3.1–3.4 replicate across collections; the documented exception is a level effect, not an ordering one (GPT-5.4-mini's threshold-crossing style classifications, Section 3.4). Item-level conclusions, by contrast, can invert. Pooled over the corrected collection the hardest items are w22 (31%), s07 (72%), and w19 (77%); 12 of the 20 testable items differ significantly across models (Cochran's Q, FDR-corrected; two items are at ceiling in all models), with s07 the most

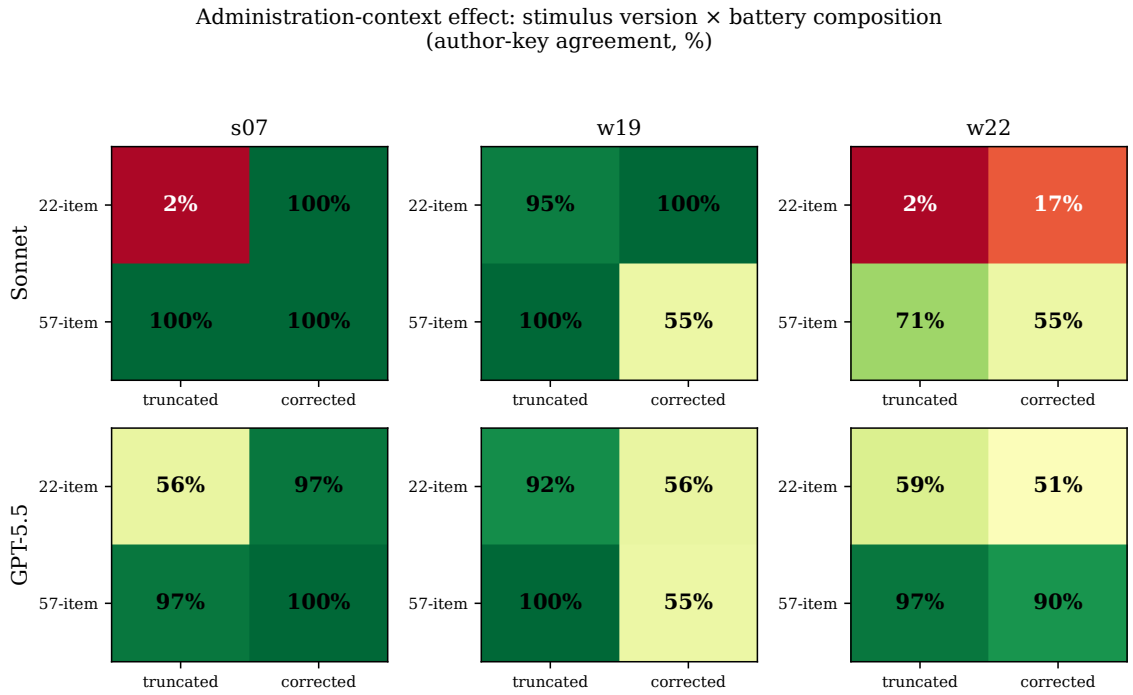

Figure 8: Administration-context effect: author-key agreement for the same model and item as a function of stimulus version and battery composition, for the three most affected items (columns) in Claude Sonnet 4.6 and GPT-5.5 (rows). 22-item cells pool persona administrations ($n = 59$–61); 57-item cells are single-run administrations ($n = 30$–31).

heterogeneous ($Q = 127.9$, $q = 7 \times 10^{-24}$).

3.6 The same item, the same stimulus, a different battery: an administration-context effect

The 22 vignettes are a subset of a 57-vignette pool that was separately administered—same scoring, same platforms—to Claude Sonnet 4.6 and GPT-5.5, in both stimulus versions. This yields a 2×2 comparison (truncated/corrected $\times$ 22-item/57-item) for every affected item (Figure 8).

Three observations. First, on s07 the battery context separates the cells with the stimulus held fixed: Sonnet answers it at 100% inside the 57-item battery *on the truncated stimulus*—the same stimulus, key, model, and (in the initial collection) the same day on which it scored 1.7% inside the 22-item battery. Neither the missing line (decisive in the short battery, irrelevant in the long one) nor its restoration explains the item without reference to the administration—though battery composition, item position,

prompt length, and administration count change together here, so which ingredient of the administration carries the effect remains open. Second, on the corrected stimulus w19 is hard in *both* batteries—both models land on an identical 17/31 (54.8%) in the corrected 57-item cell—confirming that its high truncated-stimulus scores were accidental simplification rather than a battery artifact. Third, the w22 key that fails in the 22-item administration shows far higher author-key agreement in the 57-item administration for both models and both stimulus versions (54.8–96.7%): the key's failure is administration-specific, not absolute. A plausible mechanism is context composition—the 35 additional vignettes, weighted toward manipulation and bond-decay scenarios, plausibly shift the model's reading posture toward the warier interpretations the key encodes—but position and composition are confounded in this design and cannot be separated post hoc.

The methodological consequence parallels the rendering lesson: item-level key agreement in LLM-as-respondent designs is a property of the (item, rendered stimulus, administration protocol) triple, not of the item alone. Reports should state battery composition and item position, exactly as human-testing standards require, and adjudication panels for contested items must specify the administration under which the key is judged.

3.7 Distractor-category fingerprints and the zero-prompt condition

Because every TCTM-22 distractor carries an author-assigned MASC-style category tag, the *composition* of non-keyed selections can be compared across models (Figure 9); like the key itself, the tags are authorial and unadjudicated, so what follows describes which author-categorized options models select, not directly diagnosed mentalization failures. On the corrected collection the baseline fingerprints are model-specific and qualitatively distinct: both Claude models' non-keyed selections fall exclusively in the author-tagged undermentalizing category (Sonnet 4.5%, Opus 9.1% DOS; zero overmentalizing); Gemini is the only model whose non-keyed selections fall primarily in the over-attribution category (5.0% NAD, plus 4.5% no-mentalization); GPT-5.5 is near ceiling with almost no non-keyed selections of any kind; and GPT-5.4-mini spreads them across all three categories. Models

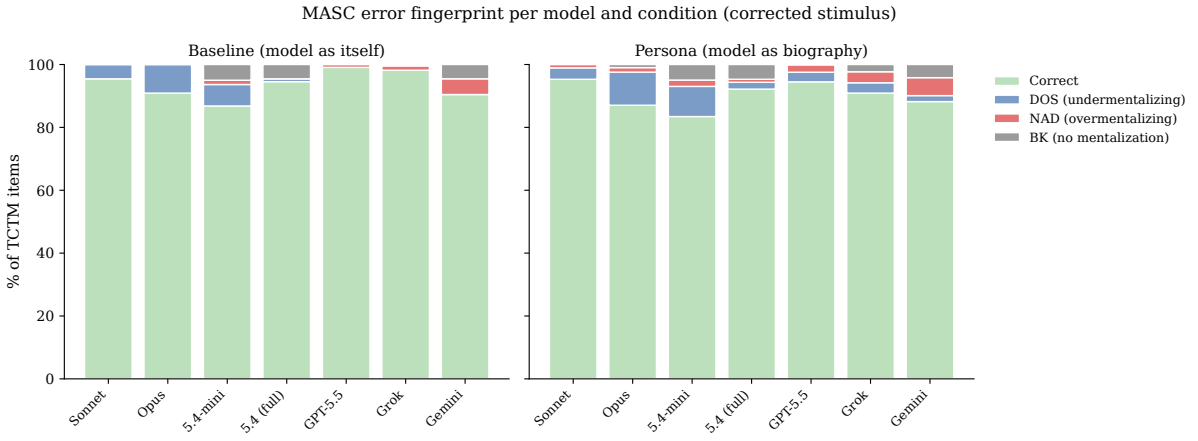

Figure 9: Composition of author-tagged MASC-style categories per model (corrected collection): author-key matches and the three author-assigned distractor categories as percentages of all TCTM-22 items, in the baseline (left) and persona (right) conditions.

with similar totals can thus diverge from the key in opposite directions—an argument against treating any single agreement total as a unidimensional benchmark, and a reusable idea for LLM evaluation generally: tag distractors, not just keys.

In the zero-prompt condition (no system prompt at all), every model produces at least one scoreable battery, but no model uses the requested canonical response schema in every run. Models spontaneously rename instruments and sections (Polish key names, sectioned layouts, and invented item-ID conventions). The separate intermittent `c07→c06` mislabel occurs in Gemini's persona-condition payloads, not in this zero-prompt result (Supplement S3). Scored under a permissive recognizer (all variants documented in Supplement S3), zero-prompt TCTM-22 totals are high and tightly clustered (e.g., GPT-5.5 at a deterministic 22/22; Claude Opus locked at 21/22), and zero-prompt profiles resemble each model's baseline rather than any persona. The condition's chief lesson is operational: unprompted LLM administration measures data-contract reliability as much as psychometrics, and recognizer coverage is part of the measurement instrument.

4 Discussion

4.1 A third regime for LLM respondents

The literature's two dominant conditioning regimes bracket the present one from below and above: brief demographic prompts yield poor individual-level fidelity (Petrov et al., 2024), while thousands of aggregated backstories yield good population-level fidelity (Argyle et al., 2023). Dense single-person narrative biographies—hundreds of behavioral constraints about one individual—occupy the space between and, on the present evidence, yield strong *individual-level stimulus fidelity*: seven models from four vendors recover author-declared categorical targets at $\kappa = .69$–$.96$, track author-declared dimensional targets at $r \approx .8$, agree with each other at median $r > .94$ per pair, reproduce a deliberately atypical trait dissociation at the cost of test scores (the paradox persona), and reproduce the persona orderings essentially point-for-point up to eight weeks later. Two qualifications bound this claim. First, conditioning density was not experimentally manipulated: all thirty personas received a full-length biography, so the study demonstrates strong target alignment *under* one dense narrative-conditioning regime but does not establish density as the causal source of the observed fidelity—the contrast with brief-prompt results is a between-study comparison that also differs in models, language, instruments, and scoring. Second, within this regime, model choice mattered far less than the regime itself: an ablation that varies prompt density on the same personas (full biography, summary, demographic sketch, bare trait list) is the natural next experiment (Section 6).

4.2 Levels are the fragile axis; profiles are the portable one

The decomposition organizes otherwise disparate observations. Between models, baselines differ by up to 1.7 *SD* while persona orderings agree at median $r > .94$ (with gains of 0.84–1.11 on the shared signal—highly similar ordering, approximately similar scale). Within one model, a deployment window produced deterministic response locks and a threshold-straddling bimodal avoidance distribution that vanished three days later behind a documented endpoint reconfiguration. Between two collections three days apart, one

model's baseline profile moved by up to 3 raw Likert points—enough to flip it from the panel's classification outlier to its best performer, and enough to flip half its persona classifications across fixed scoring thresholds—and across eight weeks a preview-endpoint model's avoidance variance tripled; both models' persona orderings stayed at $r \geq .92$ throughout. The largest instabilities the study could observe landed on the level axis; no comparable instability was observed on the ordering axis.

For practice, this yields three concrete recommendations. First, treat baseline ("as yourself") LLM measurements as deployment-window-specific snapshots: date them, and do not pool them across windows without a stability check. Second, where the scientific claim permits, prefer slope-style designs (differences or rankings across conditioned personas within one collection) to intercept-style designs (absolute scores of unconditioned models), because the former carried all of the replicable structure in this dataset. Third, repeated administration is not automatically replication: under response-level determinism, N calls can contain one observation, so variance itself is a diagnostic that must be reported, not assumed.

4.3 Stimulus fidelity and administration fidelity as reporting layers

The before–after comparison and the battery comparison jointly motivate a reporting practice. A rendering defect moved single-item agreement by up to 98 points in one direction and 37 in the other while aggregate totals moved by at most 1.6 points of 22—so aggregate robustness cannot certify item-level claims. And an item's key can fail in a 22-item administration while succeeding in a 57-item administration of the same stimulus to the same model on the same day—so item difficulty is not separable from the administration protocol. LLM-as-respondent studies should report (a) *stimulus fidelity*: verification that the prompt actually rendered contains every element the items reference (the defect reported here was invisible at the authoring layer and obvious at the rendered-prompt layer), and (b) *administration fidelity*: battery composition, item position, and administration protocol for any item-level claim. Both checks are cheap;

both would have changed published numbers in this study had it stopped at the initial collection.

4.4 What TCTM-22 measures, and the panel as a design diagnostic

TCTM-22 adapts constructs from a video-based social-cognition tradition (Dziobek et al., 2006) into text-only chat vignettes. Text-mediated and live mentalization are different cognitive subtasks: an LLM trained on billions of chat interactions is closer to a native speaker of the former than any human, while humans operating on text-only stimuli are outside their primary modality. LLM–human comparisons on such tests should accordingly not be read as comparisons of mentalization capacity (cf. Cronbach & Meehl, 1955); here the test serves exclusively as a keyed stimulus-fidelity probe, and the human sanity check ($M = 14.3$ of 22) exclusively as a small descriptive human reference, with no claim that seven convenience respondents bound the test's difficulty.

Separately, the persona-classification matrix illustrates a reusable diagnostic: when most of a multi-vendor panel converges on the *same* unintended category for a stimulus (as for `marek`, `bartek`, and `michal-k`), the stimulus, not the panel, is the likely cause—in each of the three cases here the biography demonstrably foregrounds the surface presentation the models reported. Used before human data collection, a multi-model panel can thus serve as an inexpensive pre-pilot for stimulus design, with the explicit caveat that shared training corpora make convergence weaker evidence than independent raters would provide (Zheng et al., 2023; Panickssery et al., 2024).

5 Limitations

Author-keyed test and author-declared targets. The TCTM-22 key and the biography target profiles are both author-defined, with no independent adjudication panel and no blind clinical coding of the biographies; all agreement results are therefore stimulus fidelity, not construct validity (Cronbach & Meehl, 1955; Borsboom et al., 2004). Cross-

model heterogeneity on 12 of 20 testable items, and the administration-bound behavior of w22, are themselves evidence that parts of the key do not capture a single shared construct. The designed target space is also partially redundant: over the 30 personas the median absolute correlation between the 11 continuous target dimensions is .28 (maximum .86), the first principal component carries 35% of target variance, and the participation-ratio effective dimensionality is 4.6 of 11 (Supplement S1, `target_correlation_matrix.csv`); the single paradox persona (`ola`) probes one direction of covariance decoupling and is not a full orthogonality test. Cross-model agreement is not an artifact of retest averaging: restricted to the first administration only, all 21 pairwise medians remain in [.92, .98] (`first_admin_pairwise.csv`).

Generator dependency. The biographies were drafted with Claude Opus 4.6, one of the seven evaluated models, so the stimuli could be subtly tuned to its reading (cf. LLM self-preference, Panickssery et al., 2024). Descriptively, the two Claude models show no large family advantage: across all persona runs their mean TCTM-22 score is 19.66 versus 19.44 for the five non-Claude models (Hedges' $g = 0.10$), while first-administration style agreement is .80 versus .88. In the corrected collection the corresponding TCTM-22 means are 20.06 and 19.73. These are descriptive contrasts, not independent-sample tests: runs are nested within 30 shared personas and only two Claude deployments are observed. Sonnet–Opus profile agreement ($r = .96$) is also close to mean Claude–non-Claude agreement (.957). Collectively, the results show no large descriptive generator advantage, but they cannot eliminate shared method variance. Independently authored biographies with hidden targets remain the definitive control; no multitrait–multimethod evidence (Campbell & Fiske, 1959) is available to separate trait variance from method/generator variance.

Sample sizes and post-hoc cells. $N = 30$ personas yields wide CIs on κ and r; baseline cells are $n = 10$ per collection; the bimodality observation is descriptive ($N = 31$, one cell from a wide search space, no familywise control; Simmons et al., 2011). The test–retest analysis uses two administrations and bounds noise rather than estimating reliability.

Confounded longitudinal contrasts. Initial-versus-corrected comparisons confound elapsed time, deployment updates, and the rendering correction; the three-day flagship bracket and the byte-identical self-report battery constrain but do not eliminate the confound. The within-vendor deployment comparison additionally confounds endpoint, request schema, and sampling defaults; all such contrasts are reported as *deployment-window* signatures, never as model capabilities.

Fixed administration order. Within each battery the instruments and items were administered in one fixed order for every run; the 22- versus 57-item comparison therefore cannot separate battery composition, item position, prompt length, and order effects (they change together, Section 3.6), and no order-counterbalanced administration exists in the dataset.

Closed models, one language, one culture. None of the seven models expose weights or internals, so mechanisms behind intercept offsets, locks, and drift are uninspectable; replication on open-weight models is the necessary complement. All stimuli are Polish and culturally embedded; the WEIRD critique inverts here (Henrich et al., 2010)—Anglocentric-trained models are tested on Polish pragmatics (Wierzbicka, 1991)—and English replication could move scores in either direction. The human sanity check is a seven-person convenience sample and carries no normative weight; a properly powered human validation study remains future work.

6 Future directions

Six follow-ups are specified by the present results. (1) *Cross-lingual s07*: semantically matched and culturally adapted English variants administered to the residual literal-reading pair (GPT-5.4-mini, Opus) with family siblings as controls, holding battery composition fixed. (2) *Independent adjudication of contested items*, w22 first, by a panel of ≥ 3 judges with $\mathrm{ICC}(2, k) \geq .75$, with the administration context fixed and reported—the present data show the key's behavior is administration-dependent. (3) *Independently generated biographies* (different LLM or human authors) to bound the

generator dependency. (4) *Deployment-window monitoring*: scheduled small re-collections that separate elapsed time from endpoint reconfigurations, turning the drift observations into a measured time series; plus a confirmatory $N \geq 100$ replication of the avoidance bimodality on a second endpoint. (5) *A human comparison sample* ($N \geq 100$) on TCTM-22 with battery-composition manipulated, enabling administration-fidelity checks on human respondents—if battery composition moves human item agreement the way it moves model agreement, the effect is a property of the test, not of LLMs. (6) *Conditioning-density ablation*: the same personas administered as full biography, 150–250-word summary, brief demographic sketch, and bare trait list, on at least two model families, separating the contributions of length, explicit labels, behavioral detail, and narrative form—the manipulation the present between-study contrast cannot substitute for.

Declarations

Funding. No funding was received for conducting this study.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Ethics approval. The study analyzes LLM outputs. The only human component was an informal comprehensibility check of the battery: seven adult volunteers personally known to the author completed the questionnaire once, anonymously (no names or other identifying information collected), to confirm that the vignettes and instructions were understandable. This was a minimal-risk, non-interventional exercise, not a human-subjects investigation, and its results enter the paper descriptively and in aggregate form only. Formal ethics-committee approval was not required: under the Polish statutory framework, mandatory research-ethics review covers medical experiments and does not extend to anonymous, non-interventional questionnaire checks of this kind; accordingly, no institutional committee decision exists for this component.

Consent to participate / for publication. Sanity-check participants provided written informed consent covering anonymous participation and publication of aggregate

data; the public release follows that scope. No other human data are involved.

Availability of data and materials. The scored data (run-level CSV files tagged by collection event), the 30 persona biographies with their declared target profiles, the complete TCTM-22/57 vignette source with key and distractor tags (`stimuli/tctm54.ts`; Wiencek, 2026a,b), the collection and analysis code, the per-call audit manifest, and documentation are available in the public GitHub repository. An immutable `v1.0.1` release of the scored data, stimuli, and code is archived at Zenodo under DOI 10.5281/zenodo.21410037 (concept DOI for all versions: 10.5281/zenodo.21406223). Exact model-facing prompts and raw per-run responses are deposited separately under restricted access because they embed copyrighted third-party instrument items. Analyses reported in this manuscript correspond to release `v1.0.1`, commit `72c104c`. Author-created materials are released under CC BY 4.0 and code under MIT; verbatim items of the published third-party instruments are excluded from these licenses (see `THIRD_PARTY_NOTICES.md` in the repository).

Code availability. All collection, scoring, analysis, and figure code will be deposited in the same archive; Supplement S1 maps every table and figure in this paper to its generating script. Analyses used Python 3.12 with fixed seeds.

Use of AI assistance. The persona biographies were drafted with Claude Opus 4.6 as a writing assistant (disclosed as a design limitation in Section 5); analysis code and manuscript drafting were assisted by Claude (Anthropic), and the present revision was additionally assisted by OpenAI Codex. The author reviewed all generated content and takes full responsibility for the manuscript.

Authors' contributions. Sole author: conceptualization, methodology, software, investigation, formal analysis, data curation, visualization, writing.

Open Practices Statement

All data, stimulus materials, and collection and analysis code are publicly available in the project's GitHub repository (`https://github.com/DXtR1337/30synthetic-polish-personas`);

the immutable release `v1.0.1` is archived at Zenodo under DOI 10.5281/zenodo.21410037, and the raw per-run artifacts are deposited as a separate restricted-access record. The study was not preregistered; H1–H4 are a priori expectations documented in working notes that predate data collection—not a formal preregistration—and all other analyses are labeled exploratory in the text.

Acknowledgments

This work emerged from the author's MA studies at the Institute of Psychology, University of the National Education Commission, Kraków. The author thanks the sanity-check participants.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning*.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Bai, Y., Huang, T., Sun, K., & Chen, Y. (2025). Scaling law in LLM simulated personality: More detailed and realistic persona profile is all you need. *arXiv preprint* arXiv:2510.11734.

Bartholomew, K., & Horowitz, L. M. (1991). Attachment styles among young adults: A test of a four-category model. *Journal of Personality and Social Psychology*, *61*(2), 226–244.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cui, Z., Li, N., & Zhou, H. (2024). Can large language models replace human subjects? A large-scale replication of scenario-based experiments in psychology and management. *arXiv*, 2409.00128.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J., Wolf, O., & Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*(5), 623–636.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

Han, P., Kocielnik, R., Song, P., Debnath, R., Mobbs, D., Anandkumar, A., & Alvarez, R. M. (2025). The personality illusion: Revealing dissociation between self-reports and behavior in LLMs. *arXiv preprint* arXiv:2509.03730.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Jańczak, M. O. (2021). Polish adaptation and validation of the Mentalization Scale (MentS)—a self-report measure of mentalizing. *Psychiatria Polska*, *55*(6), 1257–1274. https://doi.org/10.12740/PP/125383

Kang, M., Moon, S., Lee, S. H., Raj, A., Suh, J., Chan, D. M., & Canny, J. (2025). Deep binding of language model virtual personas: A study on approximating political partisan misperceptions. *arXiv preprint* arXiv:2504.11673.

Lee, S., Lim, S., Han, S., Oh, G., Chae, H., Chung, J., Kim, M., Kwak, B.-w., Lee, Y., Lee, D., Yeo, J., & Yu, Y. (2025). Do LLMs have distinct and consistent personality?

TRAIT: Personality testset designed for LLMs with psychometrics. *Findings of NAACL 2025*, 8412–8452. https://doi.org/10.18653/v1/2025.findings-naacl.469

Li, W., Liu, J., Liu, A., Zhou, X., Diab, M., & Sap, M. (2024). BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. *arXiv*, 2410.16491.

Lubiewska, K., Głogowska, K., Mickiewicz, K., Wojtynkiewicz, E., Izdebski, P., & Wiśniewski, C. (2016). The Experiences in Close Relationships–Revised questionnaire: Factorial structure, reliability, and a short version of the scale in a Polish sample. *Psychologia Rozwojowa*, *21*(1), 49–63. https://doi.org/10.4467/20843879PR.16.004.4793

Matusz, P. J., Traczyk, J., & Gąsiorowska, A. (2011). Kwestionariusz Potrzeby Poznania—konstrukcja i weryfikacja empiryczna narzędzia mierzącego motywację poznawczą [Need for Cognition Questionnaire: Construction and empirical verification of a measure of cognitive motivation]. *Psychologia Społeczna*, *6*(2), 113–128.

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, *37*.

Petrov, N. B., Serapio-García, G., & Rentfrow, J. (2024). Limited ability of LLMs to simulate human psychological behaviours: A psychometric analysis. *arXiv*, 2405.07248.

Plisiecki, H., & Sobieszek, A. (2024). Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection. *Behavior Research Methods*, *56*, 4716–4731. https://doi.org/10.3758/s13428-023-02212-3

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, *3*(12), pgae533.

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, *7*, 1954–1968. https://doi.org/10.1038/s42256-025-01115-6

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Sorokowska, A., Słowińska, A., Zbieg, A., & Sorokowski, P. (2014). *Polska adaptacja testu Ten Item Personality Inventory (TIPI)—TIPI-PL—wersja standardowa i internetowa* [Polish adaptation of the Ten Item Personality Inventory (TIPI)—TIPI-PL—standard and online versions]. WrocLab.

Wang, X., Xiao, Y., Huang, J.-t., Yuan, S., Xu, R., Guo, H., Tu, Q., Fei, Y., Leng, Z., Wang, W., Chen, J., Li, C., & Xiao, Y. (2024). InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1840–1873.

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, *122*(47), e2518075122. https://doi.org/10.1073/pnas.2518075122

Wiencek, M. (2026a). *TCTM-22: Full text of 22 chat-conversation vignettes with author key and distractor taxonomy* [Stimulus appendix]. Prepared for the study archive.

Wiencek, M. (2026b). *Thirty Polish narrative biographies for LLM-as-respondent research* [Data set]. Prepared for the study archive.

Wierzbicka, A. (1991). *Cross-cultural pragmatics: The semantics of human interaction*. Mouton de Gruyter.

Wigler, B., Tsfasman, M., & Matej Hrkalovic, T. (2026). Stories of your life as others: A round-trip evaluation of LLM-generated life stories conditioned on rich psychometric profiles. *arXiv preprint* arXiv:2604.06071.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, *36*.