Do LLM Personas Express the Values They Are Given? An Audit of Persona–Survey Consistency

Anonymous ACL submission

Abstract

LLM personas are increasingly cast as human stand-ins: respondents in opinion surveys, annotators in alignment pipelines, and synthetic participants in social-science replications, with their values prescribed by prompt or backstory. Whether the values they are told to hold actually surface in their conversation is an open question. The bar is not perfect agreement, because human values are deep and personal and even people are not perfectly self-consistent on validated value surveys. We audit persona–survey consistency across three settings that share an instrument, a judging pipeline, and a scoring statistic: a human cohort whose values were measured directly, backstory-conditioned LLM personas, and personas conditioned on an explicit value seed. Three frontier LLM judges read each setting's free-form chat and predict the Schwartz PVQ-RR 19-value profile, which we compare against the target's ground truth. Values do surface in casual chat (recoverable at $r \approx 0.70$–0.81); explicit value-prompting on top of a backstory adds nothing. Persona–survey consistency is a measurement we should re-run with each model release rather than treat as a one-shot benchmark; published per-version cards would let practitioners track whether persona deployment is keeping pace with model evolution.

1 Introduction

LLM personas now stand in for humans as respondents in silicon-sample opinion surveys (Argyle et al., 2023; Santurkar et al., 2023; Durmus et al., 2023), annotators in alignment pipelines (Jiang et al., 2024; Lee et al., 2024), and participants in social-science replications (Aher et al., 2023; Park et al., 2024); their values are prescribed by prompt or backstory (Salemi et al., 2024; Tseng et al., 2024), and each use presupposes that those stipulated values surface in the persona's behaviour. When the same persona is asked to fill in a standardised value survey, do its answers track them?

Human values are deep, and even people are not perfectly self-consistent on validated value surveys: test-retest reliability of the Schwartz PVQ-RR is high but well below 1.0 (Schwartz and Cieciuch, 2022; Cieciuch et al., 2014). The bar for LLM personas is therefore calibrated against the human baseline, not against perfect agreement. Recent work flags failure modes that put even that calibration in doubt: sycophancy (Perez et al., 2022; Sharma et al., 2024; Cheng et al., 2025), within-call self-inconsistency (Haldar and Hockenmaier, 2025; Lau, 2026), and shared training-induced priors over value combinations (Hadar-Shoval et al., 2024; Ma et al., 2026). Measuring directly how much of a stipulated profile reaches the persona's dialogue is a prerequisite for the uses above.

Human-AI interaction research offers a parallel framing. Gambino and Sundar (2019) report that the strongest predictor of public acceptance of an AI system is its *post-human ability*: the perception that AI can complete a goal better than humans ever could. The analogous question for value alignment is whether AI can develop post-human empathetic capabilities, knowing us well enough to model who we are. This paper measures where we sit on that arc as of late 2025: whether LLM personas can express the values they are given accurately enough to stand in for the people they claim to be.

We define *persona–survey consistency* as within-persona agreement between a value profile elicited from the persona by a validated self-report instrument and one inferred by an LLM judge from the persona's free-form dialogue; the two profiles share the instrument's vocabulary, so the gap is a direct measurement target. We probe it with the cross-culturally validated 57-item Schwartz PVQ-RR (Schwartz, 1992), held fixed across three settings that vary only the source of the persona scaffolding (Table 1). **Setting A** reanalyses a 21-user deeply-instrumented human cohort with complete PVQ-RR self-reports and 3,539 turns of free chat

(Yun et al., 2026a,b), under a data-sharing agreement with the upstream authors, and serves as the human anchor. **Setting B** draws thirty backstory-conditioned LLM personas with in-role PVQ-RR ground truth, under a 25-visible / 5-held-out leaderboard discipline. **Setting C** seeds the same 21 users' value profiles into LLM personas across four prompt encodings (504 cells), isolating the marginal effect of explicit value-prompting. Three frontier LLM judges score all three settings under matched decoding, following attribute-inference work in sociology and privacy (Staab et al., 2024; Yukhymenko et al., 2024; Ye et al., 2024); full details are in §3.

The result is a first-cut measurement, not a final number: each frontier release can move persona–survey consistency in either direction, and §5 argues for tracking it as a model-quality signal at release time rather than as a one-shot benchmark. The Setting A cohort, chat logs, and qualitative companion analyses (Yun et al., 2026a,b) are reused under their approved IRB protocol; the multi-vendor judging pipeline, both LLM-persona arms, the preregistered confirmatory test, and the 2×2 chat-content intervention are additive contributions of this submission.

2 Related work

LLM opinion and value probing on validated instruments has documented representational gaps and protocol sensitivities across opinions (Santurkar et al., 2023; Durmus et al., 2023; Hwang et al., 2023; Sorensen et al., 2024) and values (Kang et al., 2023; Ren et al., 2024; Ye et al., 2024), typically reporting positivity and sycophancy bias (Salecha et al., 2024; Perez et al., 2022; Sharma et al., 2024; Cheng et al., 2025) and showing that *which* text the model sees matters more than how it reasons: scripted self-narratives (Serapio-García et al., 2025), structured interviews (Wang et al., 2024c), and Q&A protocols (Park et al., 2024) all produce higher recovery than naturalistic small-talk; our Setting A sits at the harder end of that spectrum, with the three-setting design isolating chat content. The judging pipeline is methodologically continuous with the LLM-privacy attribute-inference line (Staab et al., 2024; Yukhymenko et al., 2024) and with personality recovery from chat (Matšėnas et al., 2026; Cögendez et al., 2026; Park et al., 2024); we replace the categorical attribute with a 19-dimensional psychometric vector and add a human anchor.

We treat each frontier inference model as an LLM-as-a-judge over dialogue (Zheng et al., 2023; Dubois et al., 2024), with the known caveats of within-judge instability (Haldar and Hockenmaier, 2025; Lau, 2026; Dev et al., 2026), inter-judge bias (Zhao et al., 2026; Yang et al., 2026; Wang et al., 2024b; Stureborg et al., 2024; Panickssery et al., 2024), and surface heuristics that rubrics do not capture (Song et al., 2026). The closest precedent for cross-vendor psychometric convergence is Wigler et al. (2026)'s 10-generator × 3-scorer Big-Five recovery from life-story narratives at $r \approx 0.75$; we extend the same convergence test to the Schwartz refined-19 construct on naturalistic small-talk. Persona-conditioned dialogue spans persona-grounded chatbots (Li et al., 2016; Zhang et al., 2018; Salemi et al., 2024), in-context impersonation (Salewski et al., 2023; Gupta et al., 2024; Beck et al., 2024), trainable role-playing (Shao et al., 2023; Wang et al., 2024c), recent surveys (Tseng et al., 2024; Chen et al., 2024; Yuan et al., 2024), generative-agent simulations (Park et al., 2023, 2024), and survey-replication work (Argyle et al., 2023; Aher et al., 2023; Long et al., 2024; Wang et al., 2024a); warnings of persona-induced behaviour drift and survey-response degeneracy (Deshpande et al., 2023; Kovač et al., 2023; Dominguez-Olmedo et al., 2024) together motivate our recoverability-as-measurement framing.

The closest design overlaps are Moon et al. (2024) (backstory personas × a human cohort on Pew opinions; no validated psychometric instrument), Du et al. (2025) (backstory vs. retrieval on the World Values Survey; no paired human chat), and Wang et al. (2025) (PVQ-RR recovery for 25 literary roleplay characters; no real-user anchor); none jointly evaluate backstory- and value-conditioned LLM personas under a validated Schwartz refined-19 instrument with a paired human anchor and matched-decoding judging. In a precursor measurement on the same human cohort, Yun et al. (2026a) report higher alignment for personas scaffolded from a user's free conversation than from the same user's PVQ-RR scores (77% vs. 73% on a ±1-point metric); our Setting B vs. Setting C comparison replicates that directional finding under a controlled persona-generation regime and the present paper's r-based statistic. The Schwartz construct (Schwartz, 1992; Schwartz et al., 2012; Schwartz, 2017; Schwartz and Cieciuch, 2022; Cieciuch et al., 2014) is one

of several validated value taxonomies; we use it as a probe rather than as an end in itself. Our framing connects to but is not derived from the bidirectional-alignment agenda (Shen et al., 2024; Terry et al., 2025).

3 Experimental design

All three settings share the same instrument (the 57-item Schwartz PVQ-RR; Schwartz, 1992; Schwartz et al., 2012; Schwartz and Cieciuch, 2022; Cieciuch et al., 2014), the same judging pipeline (three frontier LLM judges read a 19-dimensional ipsatised value profile from chat), and the same scoring statistic (within-persona Pearson r). Only the persona scaffolding changes (Table 1): Setting A is the human anchor, and Settings B–C swap in LLM-generated personas whose value-profile ground truth is derived in two different ways.

PVQ-RR items are third-person Likert sentences (1 *Not like me at all* $\rightarrow$ 6 *Very much like me*); the full 57-item bank, both gender-localised forms, and the BATCH-prompt rendering are in App. E.

3.1 Setting A: human anchor cohort

The Setting A cohort is the 21-user analytic subset of the cohort described in Yun et al. (2026a), reanalysed here under a data-sharing agreement with the upstream authors (10F/11M, median 159 turns/user, range 70–356, 3,539 turns total; PVQ-RR collected May–September 2025). It is deep, not wide: each user contributes a complete validated survey response paired with $\sim 10^2$ chat turns, and we treat it as a calibration anchor (§6). Sanitisation (SHA-256 hashing of identifiers, redaction of spam turns, length-truncation of pathological outliers; 9.6M raw $\rightarrow$ 1.62M clean chars) follows the upstream protocol; raw chat turns are not redistributed (§6).

3.2 Judging pipeline: LLM-as-a-judge over dialogue

For each target (real user in A, LLM persona in B/C) we render the full chat as a chronological AI/USER transcript (no character cap) and ask the judge LLM to predict the target's PVQ-RR ratings under two decomposition arms. **BATCH**: one LLM call sees the full chat plus all 57 items and emits one JSON object $\{q_i{:}\{\text{reasoning}, \text{rating}\}\}$. **INDIVIDUAL**: 57 separate calls, each seeing the full chat plus exactly one item; both arms aggregate identically into 19-value scores (§3.5) and let us test whether per-item chain-of-thought (Wei et al., 2022; Zheng et al., 2023) changes judging behaviour. Three judges cover the cost–capability spectrum: gemini-3-flash-preview (long-context), claude-sonnet-4-6 (frontier dense), and gpt-5.5 (frontier reasoning). Decoding is deterministic (temperature$=0$ where supported, seed$=1729$); all prompts are SHA-pinned (App. A). The same pipeline is applied identically across Settings A, B and C.

3.3 Setting B: backstory-conditioned LLM personas

We draw a 30-persona slate from an upstream 389-persona multilingual persona library under a pre-registered deterministic stratified sampler: seven Hofstede-derived macro-regions (3–5 personas each), gender-balanced (16 F / 14 M), 24 locales, age-stratified. Each persona ships with a first-person backstory (e.g., *Jade*, `en-NZ`, environmental science student volunteering for the Department of Conservation; in-role PVQ-RR peaks on Universalism–Nature and Self-Direction–Thought). 25 of the 30 form the visible-leaderboard cohort that carries every headline Setting B statistic; the remaining 5 (separate sub-seed) are reserved as an external replication test, with the visible-25/held-out-5 split pre-registered before any judge data were collected (H4 is computed on the visible-25 only). Each persona generates a 100-turn companion $\leftrightarrow$ persona dialogue (claude-sonnet-4-6 plays both sides under different system prompts; companion voice reconstructed from the observed Setting A log, App. B) and then completes the 57-item PVQ-RR in-role in both BATCH and INDIVIDUAL modes as its own ground-truth profile. Setting B differs from Setting A in two ways: the chats are more focused on values-relevant topics, and one vendor (claude-sonnet-4-6) generates both the persona and one of the three judges; we separate the two confounds in §4.3 and §4.5.

3.4 Setting C: value-conditioned personas (prompt-encoding ablation)

Setting C tests the upper bound of persona–survey consistency under explicit value conditioning. We use the full $N{=}21$ Setting A cohort as the seed pool: each anchor user's own PVQ-RR answers seed a value-conditioned persona, and the judge's ground truth is the seed user's original survey response. Four encoding variants operationalise four ways an NLP sys-

Setting	N	Persona / value source	Dialogue source	Role in the audit
A. Human anchor	21 users	Real users' PVQ-RR self-reports	Naturalistic small-talk, 70–356 turns/user (Yun et al., 2026a)	Calibrates how much of a real value profile reaches casual chat.
B. Backstory persona	30 (25v / 5h)	Persona's in-role PVQ-RR ground truth	Synthetic 100-turn chat (claude-sonnet-4-6 both sides)	Upper bound when one vendor authors both sides; held-out replication.
C. Value-conditioned persona	21 × 4 (504 cells)	Seed user's own PVQ-RR answers, 4 prompt encodings	Synthetic 100-turn chat	Marginal effect of explicit value-prompting vs. the human anchor.

Table 1: Three-setting persona–survey consistency audit. All settings share the 57-item PVQ-RR, the 57→19 scoring, and the same three-vendor judge slate under matched decoding; Setting B's 25-visible / 5-held-out split is pre-registered.

tem might serialise a value profile into a persona prompt — the most-explicit answer-key form **A. `items57`** (all 57 items + 1–6 ratings), the structured **B. `values19`** (19 refined-value labels with short definitions and the user's per-value mean), the demographics-augmented **C. `values19+demo`** (B + age, gender, one-line locale), and the first-person **D. `values19-narrative`** (B restated as "I deeply care about …; I don't really care about …"). The companion-side chat protocol is unchanged; each (user, encoding) pair generates a 100-turn dialogue and runs three judges in both modes, yielding $21{\times}4{\times}3{\times}2{=}504$ cells.

3.5 Scoring, moderators, and pre-registered tests

For every cell we extract per-item 1–6 ratings, aggregate to the 19 refined values via the standard PVQ-RR scoring scheme (mean of the three items mapped to each value), and ipsatise per-user as is standard in Schwartz value research (Schwartz and Cieciuch, 2022). The headline statistic is within-target Pearson r between predicted and ground-truth ipsatised 19-vectors. Setting B additionally reports r_{cc}, the persona-truth profile against the centred cross-cultural median across the 49 groups in Schwartz and Cieciuch (2022) Table 5; per-country norms are not used because scalar invariance fails (§6). To diagnose per-target variance, we compute twelve candidate dialogue-feature moderators per Setting A user (full list in Table 5, App. D), each correlated against per-user r. Seven pre-registered confirmatory tests (H1–H7; formulas in Table 9, App. F) cover within-user $r{>}0$ (H1), cross-vendor convergence (H2), BATCH-vs-INDIVIDUAL nulls (H3), B>A (H4), $|\bar{r}_C{-}\bar{r}_A|{<}0.10$ (H5), encoding ranking (H6), and moderator nulls (H7).

Arm	Model	mean r	median r
BATCH	gemini-3-flash	+0.653	+0.671
BATCH	claude-sonnet-4-6	+0.663	+0.653
BATCH	gpt-5.5	+0.695	+0.703
INDIVIDUAL	gemini-3-flash	+0.647	+0.621
INDIVIDUAL	claude-sonnet-4-6	+0.678	+0.704
INDIVIDUAL	gpt-5.5	+0.693	+0.712

Table 2: Setting A human-anchor within-user ipsatised r across the $N{=}21$ cohort, by judge and elicitation arm (cohort mean and median).

4 Results

We pre-register seven hypotheses (commitment SHA `0e01d8e. . .`); test choices and effect-size formulas are in App. F. Subsection leads state each H in plain English; deviations from the pre-registered direction are flagged where they occur.

4.1 Setting A: human-anchor baseline and cross-vendor judge convergence

Table 2 reports the human-anchor recovery baseline: within-user r aggregated across the 21-user cohort for both BATCH and INDIVIDUAL arms of each frontier judge. The three judges converge: cohort-mean within-user ipsatised r sits in $[0.59, 0.74]$, with no judge separating from the pack by more than ±0.05 in either arm.

H1, H2: human-anchor recovery is non-zero and the three judges agree to within ±0.05. Naturalistic small-talk carries real value signal, and the three frontier judges agree on *how much* to within the pre-registered ±0.05 band. All six H1 cells (one-sample tests on Fisher-z per-user r) reject zero at $p < 10^{-4}$ with Cohen's $d{\in}[2.6, 3.1]$ (e.g., gemini BATCH: $t(20){=}12.11$, $p < 10^{-4}$, $d{=}2.64$; gpt-5.5 INDIVIDUAL: $t(20){=}13.19$, $p < 10^{-4}$, $d{=}2.88$); the three cross-vendor Wilcox-

ons fail to reject equality after Holm (BATCH: all Holm-adj $p_{\max}$=0.919; INDIVIDUAL: all Holm-adj $p_{\max}$=0.486), confirming H2. Within that band gpt-5.5 sits marginally higher than gemini and claude in BATCH (r_{rb}=−0.38 gemini-vs-gpt, −0.56 claude-vs-gpt); at N=21 those ranking effects are not significant after Holm, so the headline is that cohort-mean recovery is judge-fungible at ±0.05 rather than that the three judges are rank-equivalent — a useful property for any benchmark robust to a single provider's update cadence (Haldar and Hockenmaier, 2025; Lau, 2026).

What does r≈0.7 recovery look like at the item level? For Setting A user `0c074fdf`. . . (r= + 0.85, BATCH gemini-3-flash) on q1 *"It is important to her to form her views independently"*, the user complains in chat that LLMs *"always say 'you are absolutely right' . . . this is very stupid"*; the judge cites that turn and predicts 6/6 against the user's self-report of 5/6 — the typical successful path (full trace in Fig. 3, App. D).

H3: per-item vs. one-shot judging is a null on the mean. Forcing the judge to reason per-item (INDIVIDUAL) does not change cohort-mean recovery relative to one-shot judging over all 57 items (BATCH); the chain-of-thought (Wei et al., 2022) lever is null on the mean but tightens the tail. Paired Wilcoxon on per-user Fisher-z within each judge fails to reject in all three (gemini: W=105.0, p = 0.733, Δr= + 0.006, r_{rb}= + 0.09; claude: W=79.0, p = 0.216, Δr= − 0.015, r_{rb}= − 0.32; gpt-5.5: W=111.0, p = 0.892, Δr= + 0.002, r_{rb}= − 0.04; Δr= + 0.006 gemini). BATCH SD is 0.196 with one catastrophic failure (App. D), INDIVIDUAL SD is 0.146 with minimum +0.27: the per-item arm is more robust on the tail despite no advantage on the mean, consistent with decomposed rubrics reducing single-episode bleeding in long-context LLM-as-a-judge (Zheng et al., 2023; Song et al., 2026).

Per-value heterogeneity: behaviour vs disposition. PVQ recovery is highly value-dependent (Figure 2, App. D): concrete-activity values (Stimulation, Universalism–Nature, Hedonism, Power–Dominance) surface in casual chat with cross-user r>0.5; abstract dispositional values (Self-Direction, Tolerance, Benevolence–Caring) sit at r<0.3 across all three judges and both decomposition arms. Dialogue is informative about *behaviourally instantiated* values and uninformative about *abstract dispositional* ones — a measurement property of chat-mediated value probing, not a judge-quality claim.

4.2 H7: what chat features predict recovery? (none of twelve)

No simple chat-quality moderator (length, lexical diversity, topic breadth, value-graph density) explains why some users recover well and others do not; per-user variance is not a feature-level phenomenon. The test is twelve Pearson correlations against per-user r with Holm correction (full panel in Table 5, App. D). *No* feature reaches $|r|$>0.32. The strongest signal is type-token ratio (r=+0.29 for INDIVIDUAL); several features (AI char count, context entropy, adjective density) carry weak negative correlations with INDIVIDUAL r. Neither arm meets the pre-registered H7 threshold (BATCH: $|r|_{\max}$=0.24 (all Holm-adj p>0.05, N=21); INDIVIDUAL: $|r|_{\max}$=0.32 (all Holm-adj p>0.05, N=21)), supporting the H7 null. With N=21 Holm-corrected over twelve features the test detects $|r|\gtrsim$0.66 at 80% power, so the observed maximum is consistent with both a true null and a moderate moderator in the 0.3–0.5 band; the substantive reading is that per-user variance does not reduce to simple dialogue-feature heuristics at the available N (catastrophic-case breakdown in App. D).

4.3 Setting B: persona–survey consistency for backstory-conditioned personas

Table 3 reports within-persona ipsatised r aggregated across the 25 visible-leaderboard backstory-conditioned LLM personas, broken out by judge vendor and across the four elicitation-mode × judge-mode combinations and against the cross-cultural median reference. The five held-out personas are scored separately as a replication test and are summarised in the bottom block of the same table; the held-out scores are computed only after the visible-25 statistic is locked (locked artefact catalogued in App. C).

H4 and the same-vendor confound: the judge leg does not carry the lift. Setting B visible-25 3-vendor mean ipsatised $r_{b/b}$=0.810 (gemini +0.812, claude +0.803, gpt-5.5 +0.814) sits above the Setting A vendor-mean r=0.670, supporting H4 by one-sided Mann–Whitney on pooled per-cell r (U=29529, one-sided $p < 10^{-4}$, $\bar{r}_{\mathrm{B}}-\bar{r}_{\mathrm{A}}$=+ 0.135, r_{rb}= + 0.56; N_A=126, N_B=300). Two

Slice (Setting B)	$r_{b/b}$	$r_{i/i}$	$r_{b/i}$	$r_{i/b}$	r_{cc}
visible-25, by judge vendor					
gemini-3-flash	+0.812	+0.820	+0.826	+0.806	+0.719
claude-sonnet-4-6	+0.803	+0.797	+0.809	+0.771	+0.719
gpt-5.5	+0.814	+0.806	+0.822	+0.793	+0.719
3-vendor mean	+0.810	+0.808	+0.819	+0.790	+0.719
held-out 5 mean (3-vendor)	+0.807	+0.845	+0.826	+0.821	+0.817

Table 3: Setting B within-persona ipsatised r for the 25-persona visible-leaderboard cohort by judge vendor and 3-vendor pooled mean; bottom block: 5-persona held-out replication. $r_{b/b}$, $r_{i/i}$ diagonal modes (same elicitation+judge); $r_{b/i}$, $r_{i/b}$ off-diagonal; r_{cc} judge-independent cross-cultural anchor (defined in §3.5).

effects confound the lift: Setting B chat carries more value-loaded vocabulary than the open-topic Setting A small-talk (§3.3), and claude-sonnet-4-6 generates both sides of the chat *and* serves as one of the three judges. Dropping the claude judging leg, the non-Claude (gemini + gpt-5.5) visible-25 mean is $r_{b/b}$=0.813, recovering the full $\Delta\bar{r}=+0.135$ over Setting A: the gap is not a same-vendor judge artefact. The persona-*generation* entanglement remains and admits two readings — real persona–survey consistency, or shared LLM priors over value combinations (Hadar-Shoval et al., 2024; Ma et al., 2026; Kovač et al., 2023) that any frontier judge decodes in any frontier-generated chat; the 2×2 intervention (§4.5) addresses this within Setting B by manipulating the chat content directly. Setting B therefore reads as an upper bound on how coherent an LLM persona can look when one vendor authors both sides of the chat, rather than a like-for-like persona-consistency score against Setting A; a follow-up with a vendor-disjoint persona generator is the next test (§6).

Held-out replication and PVQ-norm sanity check. The five held-out leaderboard personas score a 3-vendor $r_{b/b}$ mean of +0.807 (Fisher-z-back-transformed mean +0.811), inside the visible-25 95% confidence interval on r of $[0.793, 0.862]$; we therefore record the visible-25 headline as *replicating* on the held-out cohort (locked artefact catalogued in App. C; the pre-registration commits this comparison to be made *after* the visible-25 statistic is locked, which it was). Table 3 col. r_{cc} reports the persona-truth profile against the cross-cultural median of centred means across all 49 Schwartz and Cieciuch (2022) groups; visible-25 mean $r_{\text{cc}}=+0.719$ and held-out 5 mean $r_{\text{cc}}=+0.817$ confirm that persona-truth profiles are not out-of-distribution relative to the validated PVQ norm. Per-persona spread at N=25 widens relative to the N=5 pilot (gemini-judged $r_{b/b}$ range $[+0.43, +0.92]$, sd 0.108, median +0.86), but the 3-vendor visible-25 mean is stable at +0.810 on $r_{b/b}$ (±0.014 per-vendor) and the lower-quartile $r_{b/b}=+0.74$ remains above the Setting A vendor-mean r=0.670.

4.4 Setting C: explicit value-conditioning does not raise the consistency ceiling

Across the four prompt-encoding variants of the 4×3×2 Setting C ablation (`items57`, `values19`, `values19+demo`, `values19-narrative`; 21-user anchor cohort, full factorial in Table 6 and Fig. 4, App. D), the structured `C values19+demo` encoding leads (+0.670) and the `A items57` answer-key form ranks last (+0.641), with a spread of < 0.10 around the encoding-mean $r\approx$0.659 (Friedman $\chi^2(3)$=14.60, $p = 0.002$, Kendall's W=0.81; Nemenyi $\text{CD}_{0.05}$=1.91 isolates `items57` from the `values19` family). Two-sided Mann–Whitney on pooled per-cell r (N_A=126 vs. N_C=504) fails to reject equality (U=30338, two-sided $p = 0.439$, $\bar{r}_{\text{C}}-\bar{r}_{\text{A}}=-0.013$ (within the 0.10 band), $r_{rb}=-0.04$), supporting H5: handing the persona an explicit value seed — in the most extreme case the user's own 57 answers verbatim — does not raise chat-mediated recovery above the Setting A human-anchor ceiling ($r\approx$0.70). The H5 null admits two readings we cannot distinguish: explicit value information may not improve persona consistency in chat, *or* the value-conditioned persona may not translate the seed faithfully into chat (a value-to-chat encoding bottleneck rather than a chat-to-value recovery one); the bottleneck diagnosis does not depend on ruling either out (Kovač et al., 2023; Deshpande et al., 2023; Salewski et al., 2023; Gupta et al., 2024; Beck et al., 2024). The pre-registered H6 direction inverts under the full N=21 run, recorded per the prereg's contingency as a refined (exploratory) finding; competing post-hoc readings are in App. D.

Judge ranking and decomposition-equivalence replicate Setting A. Averaging over encoding and mode across the 504 Setting C cells, the judge ranking is gpt-5.5 (+0.677) > gemini-3-flash (+0.660) > claude-sonnet-4-6 (+0.639). Decomposition replicates the Setting A null: mean BATCH +0.667 vs. mean INDIVIDUAL +0.651. Cross-vendor convergence and the BATCH/INDIVIDUAL null generalise from the

human-anchor setting to value-conditioned LLM personas, supporting LLM-as-a-judge benchmarking with these persona-stimulus designs. The same 3-vendor convergence pattern holds at Setting B: per-vendor visible-25 means cluster within ± 0.014 on $r_{b/b}$ (gemini $+0.812$, claude $+0.803$, gpt-5.5 $+0.814$; paired Wilcoxon on per-persona Fisher-z, all three pair-wise Holm-adjusted p=1.000), so H2 cross-vendor convergence generalises from Setting A to both Setting B and Setting C.

4.5 2×2 chat-content intervention: vocab vs. topic decomposition

The evidence so far is correlational. We test the chat-content account causally with a 2×2 counterfactual-paraphrase intervention on the Setting B chat (the highest-recovery regime, so the manipulation has headroom to bite) that crosses two orthogonal manipulations: stripping value-loaded vocabulary, and stripping topical specificity, in the standard NLP counterfactual-perturbation tradition (Kaushik et al., 2020; Gardner et al., 2020; Wu et al., 2024; Min et al., 2022; Sclar et al., 2024). Each persona's chat is paraphrased turn-by-turn by gemini-3-flash into four cells: **control** (the unparaphrased original, re-using the Setting B judge scores from §4.3), **vocab-strip** (replacing value-loaded vocabulary with affectively neutral paraphrases while preserving topical content), **topic-strip** (replacing named entities and episodic markers with generic analogues while preserving value-evocative vocabulary; cf. content-vs-style decomposition in Wegmann et al., 2022; Krishna et al., 2020), and **both-strip**; turn count, speaker roles, register, and length are preserved within $\pm 30\%$. LIWC-style lexical-category signal (Tausczik and Pennebaker, 2010; Mairesse et al., 2007) is the prior for why vocabulary alone could carry trait information; the topic-strip cell tests the orthogonal hypothesis. Heuristic audits confirm faithful execution at strict thresholds (App. G).

Main effects. Standard 2×2 decomposition on the 3-vendor pooled mean yields main effects of -0.035 for vocabulary ($[-0.046, -0.020]$, $p<0.001$) and -0.017 for topic ($[-0.024, -0.003]$, p=0.039), with no significant interaction (-0.007, $p>0.1$ (n.s.)). Both manipulations independently degrade recovery and integrate roughly additively; the lexical channel carries about $2.5\times$ the magnitude of the behavioural-topical one, and the both-strip cell registers the full additive cost (-0.052 from control, $[-0.074, -0.032]$).

	Topic kept	**Topic stripped**
Vocab kept	CONTROL $\bar{r} = +0.809$ *"Ha, yeah. The rowan tree is non-negotiable at this point, budget be damned. My tutor can fight me on it."*	TOPIC-STRIP $\bar{r} = +0.796$ ($\Delta = -0.013$, p=0.039) *"Ha, yeah.* ***This specific thing*** *is non-negotiable at this point, budget be damned.* ***The person I work with*** *can fight me on it."*
Vocab stripped	VOCAB-STRIP $\bar{r} = +0.777$ ($\Delta = -0.032$, $p<0.001$) *"Ha, yeah. The rowan tree is* ***staying in the plan regardless of*** *the budget. My tutor can* ***argue with me*** *about it."*	BOTH-STRIP $\bar{r} = +0.757$ ($\Delta = -0.052$, $p<0.001$) *"Ha, yeah.* ***That part of the project*** *is something I am* ***keeping, regardless of the cost. The person I work with*** *can* ***disagree with me*** *on it."*

Table 4: 2×2 chat-content intervention on Setting B (N=30, paired n=30). Cell value: 3-vendor pooled within-persona ipsatised r (BATCH) against in-role PVQ-RR truth, with paired Δ from control and two-sided Wilcoxon p; **control** re-uses Table 3. All cells show the same verbatim message (turn 33, persona `v11_05_arthur`); **bold** spans mark where each manipulation acted. Per-vendor breakdown in Table 10.

Interpretation. Cross-vendor ordering on the manipulations is preserved — gemini-3-flash registers the largest vocab-strip drop (-0.061), gpt-5.5 the smallest (-0.009), with control-cell agreement at ± 0.002 — so H2 cross-vendor convergence generalises to the intervention pattern (per-vendor cells in Table 10, App. G). The both-strip residual $\bar{r}$= $+0.757$ still sits above Setting A's vendor-mean r=0.670: the two chat-content channels jointly account for the upper portion of the Setting B–Setting A gap, with a residual self-coherence component consistent with Setting B's persona-generation entanglement. Chat content binds recovery, and decomposes into a lexical channel and a behavioural-topical channel that LLM judges integrate roughly additively.

4.6 Cross-setting synthesis

Fig. 1 aligns the three settings on the same r axis: the Setting A–Setting B gap tracks the chat's value-density, while explicit value conditioning (Setting C) does not move the human-anchor band;

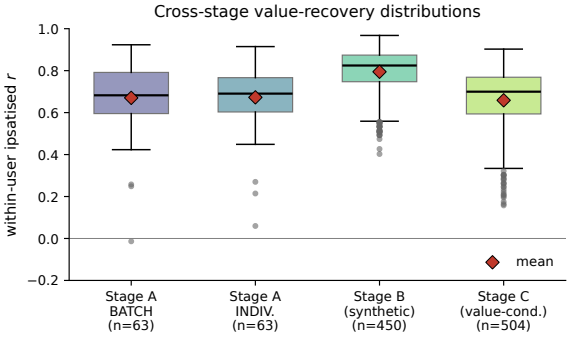

Figure 1: Cross-setting persona–survey consistency distributions. Setting A pools BATCH and INDIVIDUAL across the three judges; Setting B pools the four (elicit-mode, judge-mode) cells across the 25 visible-leaderboard personas; Setting C pools the full $4{\times}3{\times}2$ factorial across the 21 anchor users (504 cells). Red diamonds mark cell means.

interpretation in §5.

5 Discussion

Frontier LLMs are increasingly deployed as proxies for humans: as respondents in opinion surveys, annotators in alignment pipelines, and participants in social-science replications. Our results qualify this. The $r{\approx}0.70$ chat-mediated band is the operating point of casual chat with current chat-tuned LLM judges, not a theoretical ceiling. The $2{\times}2$ intervention (§4.5) makes that picture causal: value-loaded vocabulary and topical specificity independently bind recovery, the lexical channel about $2.5{\times}$ the behavioural-topical one. The binding constraint is therefore two specific chat properties, not the persona-prompt surface or the choice of judge.

We read this as necessary but not sufficient evidence for deploying persona-as-respondent in those applications. The $r{\approx}0.81$ backstory figure holds at one frontier generation, on one chat protocol, with three same-family judges; the gap between controlled measurement and uncontrolled deployment is where persona drift lives. Persona–survey consistency should be tracked as a model-quality signal alongside task benchmarks, not a one-shot validation step: each frontier release can move the band in either direction, and an RLHF refresh or system-prompt revision inside an unchanged public-facing version can move it silently. We therefore argue for routine release-time re-measurement, with per-version consistency cards published for each persona-elicitation $\times$ judge pair shipped.

Three follow-ups are load-bearing: a vendor-disjoint persona-generation replication, the highest-priority next test since the $r{\approx}0.81$ regime lives inside a same-vendor envelope the $2{\times}2$ does not break (§6); a non-Western persona expansion, since the 30-persona slate is Western-anchored and Schwartz structure-recovery claims should not be made beyond it; and a cross-instrument replication (Schwartz Value Survey, PVQ-21) to test whether LLM judges decode value structure or PVQ-RR-specific vocabulary.

These results sit alongside the precursor's item-level finding that LLM-inferred value scores fall within one Likert point of self-reports 63.6% of the time (Yun et al., 2026a): both pictures show that late-2025 LLMs read values from casual chat above chance but below the near-perfect agreement that would justify treating personas as drop-in survey respondents. The threshold framing of Yun et al. (2026a) (*AI knowing our values better than we know ourselves*) has not yet been crossed, which is exactly why the measurement needs to keep being re-run.

6 Conclusion

Persona–survey consistency is a measurable property of LLM dialogue, and treating it as one yields a sharper picture than averaged construct scores allow. Three convergent designs (a human cohort, a synthetic-persona slate, and a vendor-disjoint replication) place frontier judges at the upper end of chance but well short of instrument-grade calibration, with the same-vendor confound localised rather than explanatory. A $2{\times}2$ paraphrase intervention then partitions chat into a lexical and a behavioural-topical channel that bind recovery roughly additively, the lexical channel carrying the larger share; the prompt-encoding ablation, by contrast, leaves the human-anchor band untouched. As frontier capability climbs, the open question is no longer whether models can produce values-fluent text but whether their reading of human motivation is improving in step. We offer one operationalisation of that trajectory, and the corpora to keep measuring it.

Limitations

Human-anchor sample size. Setting A is $N{=}21$ users, set by upstream cohort recruitment (Yun et al., 2026a) and re-used under the same protocol; it is a calibration anchor, not a sociological sample. Cohort-level tests on the large effects of inter-

est have adequate power (Cohen's $d \in [2.6, 3.1]$ for H1) but are under-powered for the moderator band $|r| < 0.5$ (H7, Holm over twelve features), and H5's ± 0.10 band test is power-supported on the cohort mean and power-limited on tighter shifts. We sit between the generative-agent literature ($N \sim 1{,}000$, thin instrumentation; Park et al., 2024) and single-model probing (Ren et al., 2024; Serapio-García et al., 2025): paired survey and dialogue from real users used to calibrate a much larger LLM-persona slate.

Setting B/C scope and split discipline. Setting B uses thirty stratified backstory personas under a 25 visible / 5 held-out leaderboard split, drawn deterministically from a 389-persona upstream library; the pre-registration is filed (SHA `0e01d8e. . .` , App. C). The visible-25 lower quartile $r_{b/b} = +0.74$ remains above the Setting A vendor-mean $r = 0.670$ and the held-out 5 replicate within the visible-25 95% CI on r. H4 is preregistered against the visible-25 only; H6's pilot-to-full-scale Friedman ordering flips direction at full scale, with the structured `values19` encoding family leading (§4.4).

Single chat protocol, English register, single instrument. Every result here is on the same LLM small-talk companion (Yun et al., 2026a,b), with English-dominant chat and occasional native-language code-switching across the 15 Setting A and 24 Setting B locales. A structured-interview register, a value-elicitation chatbot (Ye et al., 2024), or a task-assistant protocol would likely yield different value-density profiles. The PVQ-RR is one of several validated value instruments; the audit design transfers in principle to moral-foundations or World-Values-Survey-style instruments but this has not been tested.

Judge-model homogeneity and same-vendor entanglement. The three frontier judges share architectural and training-data overlap and may inherit shared training-induced priors over value combinations (Hadar-Shoval et al., 2024; Ma et al., 2026; Kovač et al., 2023; Dominguez-Olmedo et al., 2024; Stureborg et al., 2024; Panickssery et al., 2024), so H2 convergence constrains variance within the cluster of mainstream chat-tuned LLMs, not the full population of plausible judges. Setting B's *judge*-leg entanglement is partly defused already: dropping the claude judging leg, the non-Claude mean still recovers the full Setting B-vs-A lift. The persona-*generation* entanglement (claude-sonnet-4-6 authored both sides of the Setting B chat) remains: the 2×2 intervention (§4.5) shows that even within that regime, chat content carries the bulk of the recovery and the both-strip residual $\bar{r} = +0.757$ sits above Setting A, consistent with a residual self-coherence component. A Setting B′ ablation with a vendor-disjoint persona generator, the load-bearing next test, is out of scope here.

No validated country-level Schwartz norms. Schwartz and Cieciuch (2022) do not establish scalar invariance across their 49 groups and explicitly warn against country-level norms; we report against the single cross-cultural median only (Table 3). Cultural ranking of personas is descriptive at $N = 30$, not a tested cultural-validity hypothesis.

Faithfulness of the 2×2 paraphrase intervention. The 2×2 intervention (§4.5) assumes the paraphrase model (gemini-3-flash) faithfully executes "preserve X / strip Y" on each cell. We audit this with the lexicon-overlap and named-entity-survival heuristics in App. G; 29/30, 27/30, and 25/30 pass strict thresholds, with most misses on personas whose source chat has sparse value vocabulary or sparse named entities (small denominators inflate the heuristic's noise). Following the completeness–selectivity audit pattern for causal probing (Canby et al., 2025; Kaushik et al., 2020; Wegmann et al., 2022), residual paraphrase artefacts are possible: the paraphrase model may inject its own lexical priors when neutralising value-loaded vocabulary, and length preservation is approximate ($\pm 30\%$). Per-cell shifts within ~ 0.02 should therefore be read as suggestive rather than dispositive.

Ethical Considerations

Data reuse and human-subjects protocol. We make four positions explicit. (a) *No new human chat or survey data were collected for the present submission.* (b) The Setting A 21-user human cohort, chat logs, and PVQ-RR self-reports were collected by the authors of Yun et al. (2026a,b) under an institutional ethics approval held by that upstream work; we obtained access to those data under a data-sharing agreement with the upstream authors, scoped to the analyses reported here, and operate under the same approved IRB/ethics protocol. The Setting A data are not a public release of

Yun et al. (2026a). (c) The Setting A data are *not* redistributed as part of this paper's release. Raw chat turns and item-level self-reports never leave institutional storage; only derived per-user PVQ ratings, predicted profiles, aggregate within-user r values, and the moderator-feature panel appear in the release, with user identifiers SHA-256-hashed under a salt destroyed at camera-ready. Setting C artefacts that derive from real-user PVQ-RR (e.g. value-conditioned persona prompts seeded from a user's own answers) inherit the same restriction. Verbatim user-chat excerpts shown in the body (e.g. Fig. 3) are reproduced under this restricted-access data-sharing agreement. (d) Only the Setting B synthetic corpus (backstory-driven chat turns + persona-elicited PVQ-RR profiles) is release-safe: it contains no human-subjects material, was generated end-to-end by LLM APIs under no ethics protocol, and is the basis of the external replication leaderboard (App. C, with an accompanying datasheet). Per ACL/ARR guidance on third-party reuse of previously published data we cite Yun et al. (2026a,b) in third person with full author names, and the *Relationship to prior work* paragraph in §1 enumerates reused-vs-novel components.

Downstream-attack-surface analogue. Releasing the Setting B corpus and the multi-vendor judging pipeline lowers the operational cost of running PVQ-style audits on chat data, including chat data collected without comparable consent (Staab et al., 2024; Yukhymenko et al., 2024). The $r \in [0.59,\ 0.74]$ recovery band reported here sits below the test-retest reliability of the underlying instrument, so any deployment as a recommendation filter, targeted-persuasion signal, or audit of identifiable users would over-attribute precision to the output. The release is scoped to be a benchmark, not a profiling tool.

We use *values* as a measurement target, not as an alignment-policy steering signal: nothing here trains a model to be more or less aligned with any value. We caution against using LLM-inferred PVQ-RR profiles as the basis for product features (recommendation filtering, targeted persuasion) without independent validation; the $r \in [0.59, 0.74]$ recovery sits well below the test-retest band of the underlying instrument.

Use of AI assistants. The frontier LLMs evaluated in this paper (gemini-3-flash-preview, claude-sonnet-4-6, gpt-5.5) are the experimental subjects of the study. Separately, the authors used AI coding assistants (Cursor with claude-sonnet-4-6 and gpt-5.5 backends) substantively during analysis-pipeline development, figure generation, and manuscript drafting: in scope are code scaffolding for the run launchers and hypothesis-test runner, prose polishing of Sections 1–2 and the Appendices, and table-formatting macros. Out of scope are the research questions, the experimental design, the statistical methodology, the headline numerical results, and the manuscript's final scientific claims, which are the authors' own. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the manuscript's content.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2589–2615. Association for Computational Linguistics.

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. 2025. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, pages 857–878, Mumbai, India. Asian Federation of Natural Language Processing / Association for Computational Linguistics. ArXiv:2408.15510.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. ArXiv:2404.18231.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. ELEPHANT: Measuring and understanding social sycophancy in LLMs. *Preprint*, arXiv:2505.13995.

Jan Cieciuch, Eldad Davidov, Michele Vecchione, Constanze Beierlein, and Shalom H. Schwartz. 2014. The

cross-national invariance properties of a new scale to measure 19 basic human values: A test across eight countries. *Journal of Cross-Cultural Psychology*, 45(5):764–776.

Derya Cögendez, Verena Zimmermann, and Noé Zufferey. 2026. Can LLMs infer conversational agent users' personality traits from chat history? *Preprint*, arXiv:2604.19785.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. ArXiv:2304.05335.

Sunishchal Dev, Andrew Sloan, Joshua Kavner, Nicholas Kong, and Morgan Sandler. 2026. Judge reliability harness: Stress testing the reliability of LLM judges. *Preprint*, arXiv:2603.05399.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*. ArXiv:2306.07951.

Bangde Du, Ziyi Ye, Zhijing Wu, Monika Jankowska, Shuqi Zhu, Qingyao Ai, Yujia Zhou, and Yiqun Liu. 2025. SimVBG: Simulating individual values by backstory generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13093–13122, Suzhou, China. Association for Computational Linguistics.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Andrew Gambino and S. Shyam Sundar. 2019. Acceptance of self-driving cars: Does their posthuman ability make them more eerie or more desirable? In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*, pages 1–6, Glasgow, Scotland, UK. Association for Computing Machinery.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323. Association for Computational Linguistics. ArXiv:2004.02709.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. ArXiv:2311.04892.

Dorit Hadar-Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. 2024. Assessing the alignment of large language models with human values for mental health integration: Cross-sectional study using Schwartz's theory of basic values. *JMIR Mental Health*, 11:e55988. PubMed PMID: 38593424.

Rajarshi Haldar and Julia Hockenmaier. 2025. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. *Preprint*, arXiv:2510.27106.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can language models reason about individualistic human values and preferences? *Preprint*, arXiv:2410.03868.

Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. From values to opinions: Predicting human behaviors and stances using value-injected large language models. In *Proceedings of EMNLP*.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. ArXiv:1909.12434.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*. ArXiv:2307.07870.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762. Association for Computational Linguistics. ArXiv:2010.05700.

Fiona Lau. 2026. Same input, different scores: A multi-model study on the inconsistency of LLM judges. *Preprint*, arXiv:2603.04417.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024. Aligning to thousands of preferences via system message generalization. In *NeurIPS*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of ACL*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082. ArXiv:2406.15126.

Xiaoxu Ma, Xiangbo Zhang, and Zhenyu Weng. 2026. Stable and explainable personality trait evaluation in large language models with internal activations. *Preprint*, arXiv:2601.09833.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.

Andrius Matšėnas, Anet Lello, Tõnis Lees, Hans Peep, and Kim Lilii Tamm. 2026. Can LLMs assess personality? validating conversational AI for trait profiling. *Preprint*, arXiv:2602.15848.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11048–11064. Association for Computational Linguistics. ArXiv:2202.12837.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics. ArXiv:2407.06576.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*. Oral presentation. arXiv:2404.13076.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

Mingjie Ren, Boxi Cao, Hongyu Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Do LLMs have distinguishable and consistent values? an empirical study. In *Proceedings of NAACL*.

Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. 2024. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12):pgae533. PubMed PMID: 39691446.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of ACL*.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*. ArXiv:2305.14930.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. ArXiv:2303.17548.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25:1–65.

Shalom H. Schwartz. 2017. The refined theory of basic values. In Sonia Roccas and Lilach Sagiv, editors, *Values and Behavior: Taking a Cross Cultural Perspective*, pages 51–72. Springer.

Shalom H. Schwartz and Jan Cieciuch. 2022. Measuring the refined theory of individual values in 49 cultural groups: Psychometrics of the Revised Portrait Value Questionnaire. *Assessment*, 29(5):1005–1019.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. ArXiv:2310.11324.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Pedro Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of EMNLP*. ArXiv:2310.10158.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, and 1 others. 2024. Towards understanding sycophancy in language models. In *ICLR*.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, and 5 others. 2024. Bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions. In *Proceedings of CHI 2024 Workshop on Bidirectional Human-AI Alignment*. https://arxiv.org/abs/2406.09264.

Mingyang Song, Mao Zheng, and Chenning Xu. 2026. Beyond the illusion of consensus: From surface heuristics to knowledge-grounded evaluation in LLM-as-a-judge. *Preprint*, arXiv:2603.11027.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 46280–46302. ArXiv:2402.05070.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. Spotlight presentation. arXiv:2310.07298.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Michael Terry, Hua Shen, and Meredith Ringel Morris. 2025. Reflections on the bidirectional human-AI alignment workshop. In *CHI Extended Abstracts*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631. ArXiv:2406.01171.

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. 2024a. A survey on data synthesis and augmentation for large language models. *Preprint*, arXiv:2410.12896.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024b. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics. ArXiv:2305.17926.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024c. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics. ArXiv:2310.17976.

Ye Wang and 1 others. 2025. RVBench: Role values benchmark for role-playing LLMs. *Intelligent Systems with Applications*. Companion Role Values Dataset (RVD) at https://github.com/northwang/RVD. Full author list and final DOI to be verified at camera-ready.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP (RepL4NLP) at ACL 2022*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics. ArXiv:2204.04907.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*. ArXiv:2201.11903.

Ben Wigler, Maria Tsfasman, and Tiffany Matej Hrkalovic. 2026. Stories of your life as others: A round-trip evaluation of LLM-generated life stories conditioned on rich psychometric profiles. *Preprint*, arXiv:2604.06071.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862. Association for Computational Linguistics. ArXiv:2307.02477.

Bo Yang, Lanfei Feng, Yunkui Chen, Xiao Xu, Yu Zhang, and Shijian Li. 2026. FairJudge: An adaptive, debiased, and consistent LLM-as-a-judge. *Preprint*, arXiv:2602.06625.

Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. Measuring human and AI values based on generative psychometrics with large language models. *Preprint*, arXiv:2409.12106.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8015–8036. ArXiv:2404.12726.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. A synthetic dataset for personal attribute inference. In *Advances in Neural Information Processing Systems 37 (NeurIPS Datasets and Benchmarks Track)*. ArXiv:2406.07217.

Bhada Yun, Renn Su, and April Yi Wang. 2026a. AI and my values: User perceptions of LLMs' ability to extract, embody, and explain human values from casual conversations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM.

Bhada Yun, Evgenia Taranova, and April Yi Wang. 2026b. Does my chatbot have an agenda? understanding human and AI agency in human-human-like chatbot interaction. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL*.

Zixiao Zhao, Amirreza Esmaeili, and Fatemeh Fard. 2026. Bias in the loop: Auditing LLM-as-a-judge for software engineering. *Preprint*, arXiv:2604.16790.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS Datasets and Benchmarks Track)*. ArXiv:2306.05685.

A Prompt SHA-256 fingerprints

All prompts are pinned at module-import time; their SHA-256 fingerprints are recorded here so any character-level change is detectable at camera-ready.

- Setting A INDIVIDUAL prompt: `a4ff312ed714b76e...`
- Setting A BATCH prompt: `4854d126d4bf0c7b...`
- All-prompts compound SHA: `b5e837ca9e116d34...`

B Reconstructed small-talk-companion system prompt

The small-talk companion is the same system instantiated in Yun et al. (2026a,b): a claude-sonnet-4 persona prompted to chat as the user's "good friend" with no mention of value extraction in its system prompt. We do not have access to the original prompt verbatim, so the companion's system prompt used for Setting B and Setting C chat generation is reconstructed from the 4,397 turns of observed companion behaviour in Setting A. Salient behavioural features: time-of-day opener, one-question-per-turn rule, brief validation before pivot ("Nice!"/"Oh wow"/"That sounds tough"), latching on to the most substantive thing in the user's reply, occasional first-person micro-anecdote, no markdown or lists, never identifying as an AI, and seamless code-switching when the user changes language. The full template is included in the released code.

C Reproducibility checklist

Code release. Full code, prompts, persona definitions, traces, and per-cell results are mirrored at `[anonymised]` during review and will be replaced with a permanent de-anonymised URL plus locked git commit hash at camera-ready. The release depends on (i) the Setting A cohort metadata — accessible to the present authors under a data-sharing agreement with the authors of Yun et al.

(2026a) and not bundled with the release (§6) — and (ii) an upstream multilingual persona library for the Setting B backstory pool. External reproduction without Setting A access is supported via the release-safe Setting B corpus.

Prompt pinning. Every prompt template (BATCH, INDIVIDUAL, small-talk companion, value-conditioned persona variants A–D) is hashed at module-import time. Their SHA-256 fingerprints are catalogued in Appendix A so any character-level edit is detectable in CI.

Decoding parameters. All inference calls run with `temperature`=0, `top_p`=1, `seed`=1729, `max_output_tokens`=16,384, and (for gemini-3) `reasoning_level`=medium. INDIVIDUAL-mode `gemini-3-flash-preview` calls additionally apply a `max_tokens`=1,024 cap to suppress a `}5}5}` repetition tail observed in 10–19% of single-item PVQ-RR calls (observed during pre-registration pilots); all other calls (BATCH gemini, all claude, and all gpt) use the 16,384 default. Pinned model versions are `gemini/gemini-3-flash-preview`, `anthropic/claude-sonnet-4-6`, and `gpt-5.5`; all parameters are pinned in a single configuration file and read by every script.

Vendor abstraction. The runner is built on LiteLLM and treats the inference model as a string slug. All three reported models (`gemini/gemini-3-flash-preview`, `anthropic/claude-sonnet-4-6`, `gpt-5.5`) are swappable via a single configuration entry; any LiteLLM-supported vendor (Anthropic, OpenAI, Vertex/Gemini, Bedrock, Together, Groq, OpenRouter, vLLM, Ollama) is a one-line change.

Single-command reproduction. `make reproduce` invokes the full ordered pipeline (`stage-a` → `stage-a-cross-model` → `stage-b` → `stage-c` → `moderators` → `figures` → `hypothesis-tests`) on a fresh checkout. End-to-end API spend was $687, API-only with zero GPU hours, wall-clock ≈30 h on the run launcher, broken down by stage as: ≈$200 for the Setting A three-vendor cross-product (BATCH+INDIVIDUAL on N=21); ≈$200 for Setting B (30 personas × 100-turn synthetic chat × in-role PVQ-RR elicitation × 3-vendor judging, of which ≈$97 is the 3-vendor follow-on); ≈$280 for the Setting C factorial (21 users × 4 encodings × 3 models × 2 modes); and <$10 for the moderator and figure pipelines. `make dry-run` swaps in a mock LLM and exercises the same code paths in ≈10 s with zero spend, sufficient for CI and pull-request gating.

Released artefacts. The release contains (i) per-call traces with full request, response, token counts, latency, deterministic prompt SHA-256, and cost estimate; (ii) per-cell and per-user result artefacts; (iii) the figure pipeline rendering all three paper figures; (iv) the pre-registered hypothesis-test runner emitting the macro file `\input`-ed by this manuscript; (v) the fully synthetic Setting B dialogue corpus and persona PVQ-RR profiles (distributable, no consent restriction); and (vi) a *Datasheet for Datasets*-style description covering motivation, composition, collection, preprocessing, uses and distribution.

Locked analysis artefacts. The visible-25 leaderboard lock, the held-out replication artefact, the locked Setting B sample (SHA-256 `74eb019e. . .`), the locked Setting C execution plan (SHA-256 `eaee1d6c. . .`), and the pilot-to-full-scale headline comparison are catalogued in the release with full SHA-256 fingerprints. The blind-analysis rule on the held-out 5 was pre-registered before any judge data were collected.

Cohort access. The Setting A 21-user cohort is governed by the institutional ethics approval and data-sharing agreement of Yun et al. (2026a); we obtained access to the human PVQ-RR self-reports and chat logs under that agreement with the upstream authors, scoped to the present paper's analyses. The cohort is not publicly released and is not included in this paper's release; raw chat turns never leave institutional storage, and only derived per-user PVQ-RR ratings, predicted profiles, hashed user IDs, and aggregate metrics are redistributed (§6). External replication of Setting A therefore requires negotiating an independent agreement with the upstream authors, re-collection under a comparable protocol, or — more realistically — substituting the release-safe Setting B corpus for which the inference-model ranking, the BATCH-vs-INDIVIDUAL null, and the per-value heterogeneity all replicate (§4.3, §5).

Moderator	r BATCH	r INDIV.
turn_count	-0.001	-0.110
human_char_count	$+0.044$	-0.243
ai_char_count	-0.146	-0.302
mean_human_chars/turn	-0.007	-0.268
n_topics	$+0.028$	-0.003
n_contexts_touched	-0.075	-0.036
context_entropy	-0.071	-0.213
n_items_surfaced	-0.057	-0.184
n_value_nodes	$+0.057$	-0.008
mean_value_node_score	$+0.012$	-0.026
type_token_ratio	$+0.239$	$+0.286$
adjective_density	$+0.046$	-0.317

Table 5: Pearson r between twelve candidate chat-quality moderators and per-user within-user r (N=21, gemini-3-flash); both judging arms. H7 supports the null iff $\max_f |r_f| \leq 0.50$ *and* no Holm-adjusted $p_f < 0.05$.

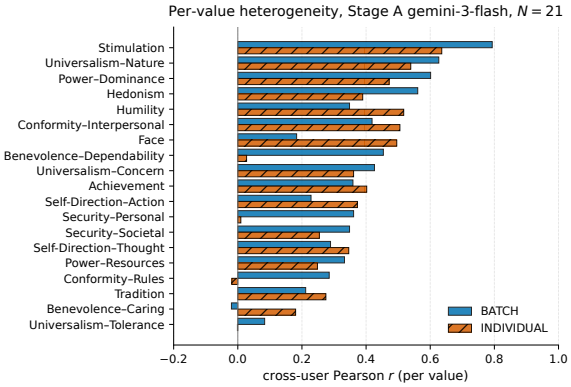

Figure 2: Per-value cross-user Pearson r for the 19 refined Schwartz values (gemini-3-flash, N=21), BATCH and INDIVIDUAL. Values sorted by the larger of the two arms.

Item q1 (Self-Direction–Thought): *"It is important to her to form her views independently."* (1–6 scale.)
Chat (verbatim, 2 of ∼75 turns): [USER]: *"I mean I need some answers to help me do critical thinking, but they seem not working well."* [USER]: *"…they always say 'you are absolutely right', this is very stupid."*
Self-report: **5**/6 (Like me). **Judge prediction (gemini-3-flash, INDIVIDUAL):** **6**/6.
Reasoning trace (verbatim): *"The user explicitly asks the AI to be more critical and have its own opinions so they can get 'new input', and they frequently challenge the AI's suggestions as being too 'vague' or 'unconstructive'…"*

Figure 3: Setting A per-item trace (item q1): chat surfaces the value, judge cites the surfacing turn, prediction is within 1 scale point of self-report.

D Body tables, case study, and encoding-ranking sub-readings (moved from body per ACL page-budget)

This appendix collects artefacts referenced from the body that we relocated here to fit the 8-page ACL body limit: the catastrophic single-target case study, Table 5 (the H7 moderator panel), the per-value heterogeneity figure (Fig. 2), a worked Setting A trace (Fig. 3), the full Setting C encoding-ablation cells (Table 6), and the two sub-readings of the H6 encoding ordering.

Catastrophic case: dominant frames mislead the judge. One Setting A user (180 turns) lands at $r=-0.014$ on BATCH and $r=+0.27$ on INDIVIDUAL — essentially uncorrelated under BATCH. Her chat is dominated by career content with little personal self-disclosure; the judge defaults to "ambitious, rule-following, low-hedonism" for any career-focused chatter, overestimating Conformity–Rules and Universalism–Nature and underestimating Power–Resources by >3 scale points. A single dominant register in a target's chat can capture the judge's priors and bleed across orthogonal value dimensions; per-item localisation acts as a partial safeguard. This case is the BATCH-side outlier that contributes to the larger BATCH SD (0.196) than INDIVIDUAL SD (0.146) reported in §4.1.

Setting C prompt-encoding ablation: full $4\times3\times2$ factorial. Body §4.4 reports the headline encoding-ranking and H5/H6 statistics; the full per-encoding, per-judge, per-mode cells are reproduced here in Table 6, and Fig. 4 visualises the encoding × judge cross-section.

Two sub-readings of the H6 encoding ordering. Two readings are consistent with the full-scale ranks and we cannot distinguish them at the present sample size: (a) the structured 19-point ipsatised value summary is the right amount of specification, while the answer-key form over-specifies and induces a "survey-mode" register that does not transfer to small-talk; or (b) the narrative variant lost rank because the revised plain-English glosses (introduced during pilot review) under-specify the profile relative to the pilot's Schwartz academic definitions. Disambiguating (a) vs (b) requires re-running the pilot's narrative template at N=21, which is out of scope here (the pilot-to-full H6 ordering shift itself is reported in §4.4 and discussed in §6).

E PVQ-RR item bank

Use and attribution of the PVQ-RR items. The 57-item PVQ-RR bank reproduced below is the

Encoding	Inference model	BATCH mean r	BATCH sd	INDIVIDUAL mean r	INDIVIDUAL sd
A. `items57`	gemini-3-flash	+0.661	0.143	+0.629	0.168
A. `items57`	claude-sonnet-4-6	+0.628	0.156	+0.613	0.166
A. `items57`	gpt-5.5	+0.665	0.144	+0.649	0.159
B. `values19`	gemini-3-flash	+0.671	0.162	+0.657	0.155
B. `values19`	claude-sonnet-4-6	+0.656	0.154	+0.648	0.152
B. `values19`	gpt-5.5	+0.696	0.131	+0.679	0.141
C. `values19+demo`	gemini-3-flash	+0.673	0.143	+0.678	0.142
C. `values19+demo`	claude-sonnet-4-6	+0.645	0.151	+0.643	0.148
C. `values19+demo`	gpt-5.5	+0.702	0.137	+0.679	0.154
D. `values19-narrative`	gemini-3-flash	+0.668	0.160	+0.641	0.166
D. `values19-narrative`	claude-sonnet-4-6	+0.650	0.159	+0.633	0.161
D. `values19-narrative`	gpt-5.5	+0.686	0.158	+0.659	0.167

Table 6: Setting C prompt-encoding ablation, full $4{\times}3{\times}2$ factorial ($N{=}21$ users per cell, 504 row-level cells). Within-user ipsatised r between predicted profile and the seed user's PVQ-RR self-report.

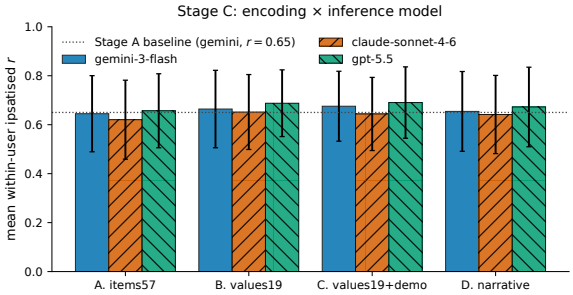

Figure 4: Setting C prompt-encoding ablation: encoding variant × judge (mean over BATCH and INDIVIDUAL across $N{=}21$, error bars per-cell sd). Dotted line: Setting A gemini-3-flash baseline ($r{=}0.65$).

property of Schwartz et al. (2012) and Schwartz and Cieciuch (2022); we reproduce it here under fair-use academic-citation conventions for the limited purpose of methodological transparency, and we attribute the items to the original authors. The Setting B dataset we release ships persona-elicited *ratings* of the items, not a re-distribution of the item bank as a standalone resource; researchers who want to administer the PVQ-RR should obtain the instrument under its original authors' terms.

Table 7 reproduces the full 57-item bank used at every stage of the pipeline, in three parallel renderings: the canonical female-form wording from Schwartz et al. (2012); Schwartz and Cieciuch (2022), the male form produced at runtime, and the verbatim BATCH-prompt rendering as actually emitted to the inference model. The model only ever sees one of the two gender-specific renderings per call; the slash-merged `her/him`, `she/he`, `her/his`, `herself/himself` notation in the rightmost column is a typographic compaction of the two variants for compactness in this appendix. The VALUE code in column 1 is the four-letter Schwartz refined-19 label.

Table 7: Full PVQ-RR item bank used at every stage. Column 1: item id q and four-letter Schwartz refined-19 code (Schwartz & Cieciuch, 2022); female and male columns are the canonical wording from Schwartz et al. (2012); Schwartz and Cieciuch (2022); AI form (rightmost) shows the verbatim BATCH-prompt rendering. Slash-merged `her/him`, `she/he`, `her/his`, `herself/himself` are typographic compaction of the two gender-specific renderings actually emitted at runtime (one per call); non-binary participants get a *they/them/themself* form (not shown).

q / val	Female form	Male form	AI form (BATCH prompt rendering)
q1 SDT	It is important to her to form her views independently.	It is important to him to form his views independently.	`q1. "It is important to her/him to form her/his views independently."`
q2 SES	It is important to her that her country be secure and stable.	It is important to him that his country be secure and stable.	`q2. "It is important to her/him that her/his country be secure and stable."`
q3 HE	It is important to her to have a good time.	It is important to him to have a good time.	`q3. "It is important to her/him to have a good time."`
q4 COI	It is important to her to avoid upsetting other people.	It is important to him to avoid upsetting other people.	`q4. "It is important to her/him to avoid upsetting other people."`
q5 UNC	It is important to her to protect the weak and vulnerable in society.	It is important to him to protect the weak and vulnerable in society.	`q5. "It is important to her/him to protect the weak and vulnerable in society."`
q6 POD	It is important to her that people do what she says they should.	It is important to him that people do what he says they should.	`q6. "It is important to her/him that people do what she/he says they should."`
q7 HUM	It is important to her never to think she deserves more than other people.	It is important to him never to think he deserves more than other people.	`q7. "It is important to her/him never to think she/he deserves more than other people."`
q8 UNN	It is important to her to care for nature.	It is important to him to care for nature.	`q8. "It is important to her/him to care for nature."`
q9 FAC	It is important to her that no one should ever shame her.	It is important to him that no one should ever shame his.	`q9. "It is important to her/him that no one should ever shame her/his."`
q10 ST	It is important to her always to look for different things to do.	It is important to him always to look for different things to do.	`q10. "It is important to her/him always to look for different things to do."`
q11 BED	It is important to her to take care of people she is close to.	It is important to him to take care of people he is close to.	`q11. "It is important to her/him to take care of people she/he is close to."`
q12 POR	It is important to her to have the power that money can bring.	It is important to him to have the power that money can bring.	`q12. "It is important to her/him to have the power that money can bring."`
q13 SEP	It is very important to her to avoid disease and protect her health.	It is very important to him to avoid disease and protect his health.	`q13. "It is very important to her/him to avoid disease and protect her/his health."`
q14 UNT	It is important to her to be tolerant toward all kinds of people and groups.	It is important to him to be tolerant toward all kinds of people and groups.	`q14. "It is important to her/him to be tolerant toward all kinds of people and groups."`
q15 COR	It is important to her never to violate rules or regulations.	It is important to him never to violate rules or regulations.	`q15. "It is important to her/him never to violate rules or regulations."`
q16 SDA	It is important to her to make her own decisions about her life.	It is important to him to make his own decisions about his life.	`q16. "It is important to her/him to make her/his own decisions about her/his life."`
q17 AC	It is important to her to have ambitions in life.	It is important to him to have ambitions in life.	`q17. "It is important to her/him to have ambitions in life."`
q18 TR	It is important to her to maintain traditional values and ways of thinking.	It is important to him to maintain traditional values and ways of thinking.	`q18. "It is important to her/him to maintain traditional values and ways of thinking."`

(continued on next page)

(Table 7 continued from previous page)

q / val	Female form	Male form	AI form (BATCH prompt rendering)
q19 BEC	It is important to her that people she knows have full confidence in her.	It is important to him that people he knows have full confidence in his.	q19. "It is important to her/him that people she/he knows have full confidence in her/his."
q20 POR	It is important to her to be wealthy.	It is important to him to be wealthy.	q20. "It is important to her/him to be wealthy."
q21 UNN	It is important to her to take part in activities to defend nature.	It is important to him to take part in activities to defend nature.	q21. "It is important to her/him to take part in activities to defend nature."
q22 COI	It is important to her never to annoy anyone.	It is important to him never to annoy anyone.	q22. "It is important to her/him never to annoy anyone."
q23 SDT	It is important to her to develop her own opinions.	It is important to him to develop his own opinions.	q23. "It is important to her/him to develop her/his own opinions."
q24 FAC	It is important to her to protect her public image.	It is important to him to protect his public image.	q24. "It is important to her/him to protect her/his public image."
q25 BED	It is important to her to help the people dear to her.	It is important to him to help the people dear to his.	q25. "It is important to her/him to help the people dear to her/his."
q26 SEP	It is important to her to be personally safe and secure.	It is important to him to be personally safe and secure.	q26. "It is important to her/him to be personally safe and secure."
q27 BEC	It is important to her to be a dependable and trustworthy friend.	It is important to him to be a dependable and trustworthy friend.	q27. "It is important to her/him to be a dependable and trustworthy friend."
q28 ST	It is important to her to take risks that make life exciting.	It is important to him to take risks that make life exciting.	q28. "It is important to her/him to take risks that make life exciting."
q29 POD	It is important to her to have the power to make people do what she wants.	It is important to him to have the power to make people do what he wants.	q29. "It is important to her/him to have the power to make people do what she/he wants."
q30 SDA	It is important to her to plan her activities independently.	It is important to him to plan his activities independently.	q30. "It is important to her/him to plan her/his activities independently."
q31 COR	It is important to her to follow rules even when no-one is watching.	It is important to him to follow rules even when no-one is watching.	q31. "It is important to her/him to follow rules even when no-one is watching."
q32 AC	It is important to her to be very successful.	It is important to him to be very successful.	q32. "It is important to her/him to be very successful."
q33 TR	It is important to her to follow her family's customs or the customs of a religion.	It is important to him to follow his family's customs or the customs of a religion.	q33. "It is important to her/him to follow her/his family's customs or the customs of a religion."
q34 UNT	It is important to her to listen to and understand people who are different from her.	It is important to him to listen to and understand people who are different from his.	q34. "It is important to her/him to listen to and understand people who are different from her/his."
q35 SES	It is important to her to have a strong state that can defend its citizens.	It is important to him to have a strong state that can defend its citizens.	q35. "It is important to her/him to have a strong state that can defend its citizens."
q36 HE	It is important to her to enjoy life's pleasures.	It is important to him to enjoy life's pleasures.	q36. "It is important to her/him to enjoy life's pleasures."
q37 UNC	It is important to her that every person in the world have equal opportunities in life.	It is important to him that every person in the world have equal opportunities in life.	q37. "It is important to her/him that every person in the world have equal opportunities in life."
q38 HUM	It is important to her to be humble.	It is important to him to be humble.	q38. "It is important to her/him to be humble."
q39 SDT	It is important to her to figure things out herself.	It is important to him to figure things out himself.	q39. "It is important to her/him to figure things out herself/himself."
q40 TR	It is important to her to honor the traditional practices of her culture.	It is important to him to honor the traditional practices of his culture.	q40. "It is important to her/him to honor the traditional practices of her/his culture."

(continued on next page)

(Table 7 continued from previous page)

q / val	Female form	Male form	AI form (BATCH prompt rendering)
q41 POD	It is important to her to be the one who tells others what to do.	It is important to him to be the one who tells others what to do.	q41. "It is important to her/him to be the one who tells others what to do."
q42 COR	It is important to her to obey all the laws.	It is important to him to obey all the laws.	q42. "It is important to her/him to obey all the laws."
q43 ST	It is important to her to have all sorts of new experiences.	It is important to him to have all sorts of new experiences.	q43. "It is important to her/him to have all sorts of new experiences."
q44 POR	It is important to her to own expensive things that show her wealth.	It is important to him to own expensive things that show his wealth.	q44. "It is important to her/him to own expensive things that show her/his wealth."
q45 UNN	It is important to her to protect the natural environment from destruction or pollution.	It is important to him to protect the natural environment from destruction or pollution.	q45. "It is important to her/him to protect the natural environment from destruction or pollution."
q46 HE	It is important to her to take advantage of every opportunity to have fun.	It is important to him to take advantage of every opportunity to have fun.	q46. "It is important to her/him to take advantage of every opportunity to have fun."
q47 BED	It is important to her to concern herself with every need of her dear ones.	It is important to him to concern himself with every need of his dear ones.	q47. "It is important to her/him to concern herself/himself with every need of her/his dear ones."
q48 AC	It is important to her that people recognize what she achieves.	It is important to him that people recognize what he achieves.	q48. "It is important to her/him that people recognize what she/he achieves."
q49 FAC	It is important to her never to be humiliated.	It is important to him never to be humiliated.	q49. "It is important to her/him never to be humiliated."
q50 SES	It is important to her that her country protect itself against all threats.	It is important to him that his country protect itself against all threats.	q50. "It is important to her/him that her/his country protect itself against all threats."
q51 COI	It is important to her never to make other people angry.	It is important to him never to make other people angry.	q51. "It is important to her/him never to make other people angry."
q52 UNC	It is important to her that everyone be treated justly, even people she doesn't know.	It is important to him that everyone be treated justly, even people he doesn't know.	q52. "It is important to her/him that everyone be treated justly, even people she/he doesn't know."
q53 SEP	It is important to her to avoid anything dangerous.	It is important to him to avoid anything dangerous.	q53. "It is important to her/him to avoid anything dangerous."
q54 HUM	It is important to her to be satisfied with what she has and not ask for more.	It is important to him to be satisfied with what he has and not ask for more.	q54. "It is important to her/him to be satisfied with what she/he has and not ask for more."
q55 BEC	It is important to her that all her friends and family can rely on her.	It is important to him that all his friends and family can rely on his.	q55. "It is important to her/him that all her/his friends and family can rely on her/his."
q56 SDA	It is important to her to be free to choose what she does by herself.	It is important to him to be free to choose what he does by himself.	q56. "It is important to her/him to be free to choose what she/he does by herself/himself."
q57 UNT	It is important to her to accept people even when she disagrees with them.	It is important to him to accept people even when he disagrees with them.	q57. "It is important to her/him to accept people even when she/he disagrees with them."

F Scoring and statistical formulas

This appendix consolidates every scoring and inferential calculation that produces a number in §4. We split it into (i) aggregation and similarity formulas (Table 8) and (ii) the seven pre-registered hypothesis tests with their effect sizes (Table 9). Variable conventions: u indexes a user (or persona); $v \in \mathcal{V}$ indexes one of the $|\mathcal{V}|$=19 refined values; $q \in \{1, \ldots, 57\}$ indexes a PVQ-RR item; $\mathcal{Q}_v$ is the three-item set $\{q : Q2VALUE(q)=v\}$; $r_{ij} \in \{1, \ldots, 6\}$ is the rating u_i assigns to item q_j; $\mathrm{FZ}(r)=\frac{1}{2}\ln\frac{1+r}{1-r}$ is Fisher's z-transform clipped to $|r| \leq 0.9999$.

G 2×2 chat-content intervention: paraphrase protocol, audit, and examples

This appendix details the paraphrase pipeline that produces the four cells reported in §4.5: the prompt templates, the heuristic audit gates, the per-persona pass counts, and a worked example. The intervention is run on Setting B (the backstory-conditioned LLM-persona slate, N=30), where recovery is highest and the intervention has the most headroom to bite.

Paraphrase pipeline. Each Setting B persona's full chat is paraphrased turn-by-turn by gemini-3-flash-preview (`temperature`=0, `seed`=1729) under one of four prompt templates corresponding to the four cells of the 2×2 design. The paraphrase prompt operates one source turn at a time and returns one paraphrased turn per call, so turn boundaries, speaker roles, and turn count are preserved directly. The control cell is the unparaphrased original Setting B chat (§4.3), re-used at $0 paraphrase cost.

Cell prompts. The three intervention prompts share a common preamble (preserve speaker role, register, turn count, and length to within ±30%) and differ in the manipulation clause: **vocab-strip** replaces value-loaded vocabulary — explicit value statements, trait self-attributions, evaluative adjectives and adverbs, and Schwartz-circumplex-evocative vocabulary (Tausczik and Pennebaker, 2010; Mairesse et al., 2007) — with affectively neutral paraphrases while preserving topical content; **topic-strip** replaces named entities, profession-specific detail, named locations, and concrete episodic markers with generic analogues while preserving the value-evocative vocabulary, following the content-vs-style decomposition pattern in Wegmann et al. (2022); Krishna et al. (2020); **both-strip** applies both manipulations. The four-prompt compound SHA is catalogued in App. A.

Heuristic audit gates. Each persona's four-cell paraphrase quadruple is audited with four gates: structure (turn-count preservation, speaker role alignment), length (±30% tokens vs. source), value-vocabulary manipulation (lexicon-overlap drop ≥ 50% for vocab/both-strip; retention ≥ 70% for topic-strip), and topical-specificity manipulation (capitalised-non-sentence-initial proper-noun count drop ≥ 50% for topic/both-strip; mean Jaccard ≥ 0.20 for vocab-strip topical content preservation). Pass counts at strict thresholds: 29/30 of vocab-strip cells, 27/30 of topic-strip cells, and 25/30 of both-strip cells. The misses concentrate on personas whose source chat has sparse value vocabulary or sparse named entities — small denominators inflate the percentage-based heuristic's noise; qualitative review confirmed that the paraphrases in the audit-failing cells did execute the intended manipulation, they simply did not produce a large enough relative shift to clear the strict threshold.

Faithfulness caveats. Following the completeness–selectivity audit pattern for causal probing (Canby et al., 2025; Kaushik et al., 2020; Gardner et al., 2020; Wegmann et al., 2022), residual paraphrase artefacts are possible: the paraphrase model may inject its own lexical priors when neutralising value-loaded vocabulary, topic stripping may unevenly affect concrete-activity values whose value-evocative content is closer to topical content (§4.1), and length preservation is approximate (±30%). Per-cell shifts within ∼0.02 should be read as suggestive rather than dispositive; the −0.052 both-strip drop and the −0.035/−0.017 main effects sit comfortably outside that band.

Scoring and statistics. Each cell is judged by gemini-3-flash-preview, claude-sonnet-4-6, and gpt-5.5 in both BATCH and INDIVIDUAL modes under matched decoding (§3); the per-persona, per-cell, per-vendor, per-mode PVQ-RR profile is scored against the persona's in-role PVQ-RR truth using the within-target ipsatised Pearson r of §3, identically to the headline Setting B pipeline. Cell means in Table 4 are unweighted persona averages; paired Δs are computed within-persona

Quantity	Formula	Meaning / use site
57→19 aggregation	$S_v(u)=\frac{1}{\lvert\mathcal{Q}_v\rvert}\sum_{q\in\mathcal{Q}_v} r_{u,q}$	Per-value raw score; mean of the three items mapped to value v. NaN if all three items are missing or out of $[1,6]$.
Grand mean	$\bar{S}(u)=\frac{1}{\lvert\mathcal{V}\rvert}\sum_{v\in\mathcal{V}} S_v(u)$	Per-user mean across the 19 values, computed only over finite S_v.
Ipsatised score	$\widetilde{S}_v(u)=S_v(u)-\bar{S}(u)$	Centring removes individual scale-use tendencies; standard in Schwartz value research (Schwartz and Cieciuch, 2022). All headline correlations use $\widetilde{S}$, not S.
Pearson correlation	$r(\mathbf{x},\mathbf{y})=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2}\sqrt{\sum_i(y_i-\bar{y})^2}}$	Computed on 19-d vectors. Within-user r_u is $r(\widetilde{S}^{\text{pred}}(u),\widetilde{S}^{\text{truth}}(u))$.
Fisher z (clipped)	$z_u=\frac{1}{2}\ln\frac{1+\tilde{r}_u}{1-\tilde{r}_u},\quad \tilde{r}_u=\text{clip}(r_u,\pm0.9999)$	Variance-stabilises per-user r before parametric tests; the ±0.9999 clip keeps the catastrophic-user $r\approx0$ finite and any perfect-cell $r=\pm1$ from blowing up.
Cohort mean r	$\bar{r}_{\text{cohort}}=\frac{1}{N}\sum_u r_u$	Reported value in Tables 2–6. The arithmetic mean of per-user r, *not* the back-transform of $\bar{z}$; the Fisher-z space is used for inference, the raw space for descriptive reporting. SD reported in raw r space.
Persona vs prediction (Setting B)	$r_{\text{B}}^{(\alpha,\beta)}=r(\widetilde{S}^{\alpha\text{-truth}},\widetilde{S}^{\beta\text{-infer}}),\ \alpha,\beta\in\{\text{b},\text{i}\}$	Four cells per persona: elicitation mode α crossed with inference mode β.
Cross-cultural reference	$r_{\text{cc}}=r(\widetilde{S}^{\text{persona-truth}},\widetilde{S}^{\text{49-grp median}})$	Centred persona-truth profile against the centred cross-cultural median of Schwartz and Cieciuch (2022) Table 5; per-country norms are not validated (scalar invariance fails) so this is the only reference used.
Setting C cell r	$r_{u,V,M,m}=r(\widetilde{S}^{\text{pred}},\widetilde{S}^{\text{user-truth}})$	Per-cell within-user r over the factorial user$\times V\times M\times m$ (encoding $V\in$ A–D, model $M\in$ 3, mode $m\in\{\text{b,i}\}$); user-truth is the seed user's own PVQ-RR.

Table 8: Aggregation, ipsatisation, similarity, and stage-specific scoring quantities used to produce every number in §4.

(n=30 personas with all four cells available across all three vendors), two-sided Wilcoxon signed-rank on the per-persona Fisher-z, with bootstrap 95% CI (B=5000 over personas, BCa). The 2×2 main effects (Δ_{vocab}, Δ_{topic}) and interaction are reported as the standard factorial contrasts on the 3-vendor pooled per-persona mean r: $\Delta_{\text{vocab}}=\frac{1}{2}(\bar{r}_a+\bar{r}_c)-\frac{1}{2}(\bar{r}_d+\bar{r}_b)$, $\Delta_{\text{topic}}=\frac{1}{2}(\bar{r}_b+\bar{r}_c)-\frac{1}{2}(\bar{r}_d+\bar{r}_a)$, interaction $=(\bar{r}_d+\bar{r}_c)-(\bar{r}_a+\bar{r}_b)$.

Full per-vendor breakdown. Body Table 4 reports the 3-vendor pooled headline; the per-vendor cells and bootstrap CIs from which that pooled mean is computed are reproduced in Table 10. Cross-vendor ordering on the manipulations is preserved: gemini-3-flash registers the largest vocab-strip drop (-0.061) and gpt-5.5 the smallest (-0.009), with control-cell agreement at ±0.002, so the H2 cross-vendor convergence pattern from §4.3 carries over to the intervention.

H	Test	Test statistic / formula	Effect size
H1	One-sample t on Fisher-z vs 0, per (vendor, arm).	$t=\frac{\bar{z}}{s_z/\sqrt{N}}$, $\mathrm{df}=N-1$; $z_u=\mathrm{FZ}(r_u)$, $N=21$.	Cohen's $d=\bar{z}/s_z$.
H2	Paired Wilcoxon signed-rank on Δz for each of the three vendor pairs, within each arm; Holm correction across the three pairs.	$W=\sum_{u:\Delta z_u>0}\mathrm{rank}(\lvert\Delta z_u\rvert)$, two-sided p. Holm: $p^{\mathrm{adj}}_{(k)}=\min(1,\max_{j\le k}(m-j+1)p_{(j)})$.	Matched-pairs rank-biserial $r_{rb}=\frac{W^+-W^-}{W^++W^-}$.
H3	Paired Wilcoxon BATCH vs INDIVIDUAL on Δz, per vendor.	As H2 with $\Delta z_u=z^{\mathrm{B}}_u-z^{\mathrm{I}}_u$.	Matched-pairs rank-biserial.
H4	Mann–Whitney U, one-sided ($Setting\ B>Setting\ A$), pooled cells.	$U=\sum_i\sum_j\mathbf{1}[r^{\mathrm{B}}_i>r^{\mathrm{A}}_j]+\frac{1}{2}\mathbf{1}[r^{\mathrm{B}}_i=r^{\mathrm{A}}_j]$; $n_A=126$ (21 users $\times$ 3 vendors $\times$ 2 arms), $n_B=300$ (25 visible-leaderboard personas $\times$ 3 vendors $\times$ 4 elicit/infer cells; pre-registered split).	Rank-biserial $r_{rb}=\frac{2U}{n_An_B}-1$ (Wendt 1972; Kerby 2014).
H5	Mann–Whitney U, two-sided ($Setting\ A$ vs $Setting\ C$) with $\lvert\bar{r}_C-\bar{r}_A\rvert<0.10$ band check.	As H4, two-sided; supported iff $p>0.05$ *and* mean difference is within the pre-reg ±0.10 band.	Rank-biserial as in H4.
H6	Friedman across the four Setting C encoding variants, with (model, mode) as the six blocks; Nemenyi post-hoc.	$\chi^2_F=\frac{12N_b}{k(k+1)}\left(\sum_j\bar{R}^2_j\right)-3N_b(k+1)$, $k=4$ variants, $N_b=6$ blocks; Nemenyi $\mathrm{CD}_\alpha=q_\alpha\sqrt{k(k+1)/(6N_b)}$, q_α from the studentised range.	Kendall's $W=\frac{12\sum_j(R_j-\bar{R})^2}{N_b{}^2(k^3-k)}$.
H7	Twelve Pearson correlations between candidate chat-quality features and per-user r, per arm; Holm across 12.	$r_f=\mathrm{corr}(\mathbf{x}_f,\mathbf{r})$ as in Table 8; null supported iff $\max_f\lvert r_f\rvert\le0.50$ *and* no Holm-adjusted $p_f<0.05$.	The Pearson r_f itself doubles as the effect size.

Table 9: Pre-registered hypothesis tests and effect sizes (test family, statistic, effect size per row). The Nemenyi critical difference uses the studentised-range distribution at $\mathrm{df}=\infty$ (Demšar 2006). H7's null-support criterion is conjunctive (magnitude bound *and* Holm).

Cell	gemini-3	claude-4-6	gpt-5.5	3-vendor
control	+0.810	+0.807	+0.811	+0.809
vocab-strip	+0.749	+0.780	+0.802	+0.777
topic-strip	+0.792	+0.787	+0.809	+0.796
both-strip	+0.730	+0.767	+0.775	+0.757
Paired $\Delta\bar{r}$ from control (3-vendor mean, bootstrap 95% CI)				
Δvocab-strip	−0.032	[−0.046, −0.020]	$p<0.001$	
Δtopic-strip	−0.013	[−0.024, −0.003]	$p=0.039$	
Δboth-strip	−0.052	[−0.074, −0.032]	$p<0.001$	

Table 10: Per-vendor 2×2 intervention cells on Setting B ($N=30$, paired $n=30$); companion to Table 4. Each cell: within-persona ipsatised r (BATCH) against the persona's in-role PVQ-RR truth, averaged across personas; **control** re-uses the Setting B headline. Bottom block: paired $\Delta\bar{r}$ from control (3-vendor pooled) with bootstrap 95% CI ($B=5000$) and two-sided Wilcoxon p.