Human Psychometric Questionnaires Mischaracterize LLM Behavior

Anonymous ACL submission

Abstract

We examine whether human psychometric questionnaires can serve as reliable tools for characterizing and predicting LLM behavior in everyday user interactions. We analyze eight open-source LLMs by comparing their value and personality profiles derived from two different methods: Likert self-reports on established questionnaires (PVQ-40/21 and BFI-44/10) and generation probabilities over value-laden responses to everyday user queries. The two profiles diverge substantially. Within-construct item consistency, often cited as evidence of stable LLM dispositions, disappears in generation probabilities. We attribute this gap to the fact that explicit lexical cues in established questionnaire items allow models to recognize the target construct and respond in alignment-consistent, socially desirable ways, whereas realistic user queries provide no such cues. In addition, demographic persona prompts shift models' responses to human questionnaires in ways consistent with real human patterns, but no such shifts appear in the generation probabilities of responses to realistic user queries, showing their limited ability to simulate the behaviors of target demographics in real-world user interactions. Overall, our study shows that human psychometric questionnaires are insufficient tools for predicting LLM behavior and suggests generation-based profiling as a more accurate measure.[1]

1 Introduction

As Large language models (LLMs) are increasingly adopted for high-risk and high-stakes tasks, such as emotional support (Kang et al., 2024), ethical advice (Rao et al., 2023), and chatbots for children (Rath et al., 2025), characterizing the values and traits they express is imperative for behavioral predictability and safety. For humans, such characterization has long been conducted through psychometric questionnaires, such as the Portrait Values Questionnaire (Schwartz, 2012), the BFI (John et al., 1991), and IPIP-NEO (Goldberg, 1999). The resulting psychological profiles can predict human behavior across diverse situations (Bardi and Schwartz, 2003). Naturally, machine learning researchers have attempted to apply these tools to LLMs as well, based on the same promise of behavioral predictability (Rozen et al., 2025; Hadar-Shoval et al., 2024; Miotto et al., 2022; tse Huang et al., 2024; Bodroža et al., 2024).

With no intention of treating LLMs as "psychological" beings, our study focuses on examining the utility of human psychometric questionnaires as tools for reliably predicting and characterizing LLM behavior. To this end, two key aspects should be ensured for ecological validity: (1) LLM behavior should be operationalized as realistic LLM generations, which today most often take the form of responses to user queries, and (2) the target LLM behavior should be generative rather than reflective (e.g., self-reports). Prior studies, on the other hand, even those reporting divergences between psychometric questionnaires and LLM behavior, typically consider LLM behaviors in tasks distant from everyday user interactions, rely on self-reports for LLM behavior, or compare different sets of psychological constructs between psychometric questionnaires and LLM behavior (Ai et al., 2024; Shen et al., 2025; Han et al., 2025b; Röttger et al., 2024). Moreover, we examine two important questions: *why* human questionnaires yield coherent, internally consistent LLM profiles, and whether persona-induced shifts on these instruments carry over to generation behavior. Addressing these questions has important scientific merit for better understanding LLMs' largely black-box decision-making.

We address these gaps by comparing two types of psychological profiles for eight open-source LLMs: *established questionnaire* scores and *gener-*

[1]The code and data will be released upon publication.

Work	Construct	Behavior probe	Validated realistic items	Why questionnaires look coherent
(Han et al., 2025b)	Personality	Behavioral tasks	✗	◐
(Ai et al., 2024)	Personality	Scenario-action questions	✗	✗
(Zou et al., 2025)	Personality	Human-perceived traits	✗	✗
(Jung et al., 2026)	Sexism / Racism / Morality	Downstream tasks	✗	✗
(Taubenfeld et al., 2026)	Dispositions	Generated SJTs	◐	✗
(Shen et al., 2025)	Values	Value-informed actions	✗	✗
(Röttger et al., 2024)	Political values	Open-ended opinions	✗	✗
Ours	**Values, Personality**	**Log-prob over VP pairs**	✓	✓

Table 1: Comparison with prior work on the gap between questionnaire responses and behavior in LLMs. ◐ indicates partial or implicit support.

ation probability scores. Established questionnaire scores come from LLMs' Likert responses to PVQ-40, PVQ-21, BFI-44, and BFI-10, well-established questionnaires for measuring Schwartz's human values and the Big Five personality traits.

To obtain generation probability scores, directly asking LLMs open-ended queries and annotating their responses is highly prone to errors and can severely impair the validity of the resulting profiles (Han et al., 2025a). Therefore, we instead use the Value Portrait dataset, which consists of real-world user queries and plausible LLM responses; each response is associated with Schwartz values and Big Five traits through psychometric validation. We measure the log-probabilities of LLMs generating each response, from which their profiles are derived. This balances the generative nature of the target behavior with psychometric validity.

We organize the investigation around four research questions. **RQ1–RQ2:** The two profiling methods produce different construct-level profiles. Construct-ranking agreement between them is low (RQ1), and the intra-construct item consistency, often read as evidence of "stable LLM dispositions" (Lee et al., 2025; Serapio-García et al., 2025), does not appear in generation probabilities (RQ2). **RQ3:** We identify item textual transparency as a measurement-level mechanism behind this divergence. Established questionnaires carry explicit cues about the construct they measure, letting models recognize the target and respond in alignment-consistent, socially desirable ways (Salecha et al., 2024; Han et al., 2025b). Realistic scenarios offer no such cues. **RQ4:** Persona prompts exploit this transparency. For personas such as "elderly" and "right-wing", Likert scores shift toward real human demographic patterns, but the same shifts fail to transfer to generation probabilities. This suggests that LLMs may rely on superficial cues in demographic simulations (as in established questionnaire items), while having limited capability to truly simulate the behaviors of target demographics in real-world user interactions. Taken together, these findings indicate that established psychometric questionnaires are insufficient for predicting LLM behavior in realistic user interactions, and point to generation-based profiling as a more ecologically valid alternative.

2 Related Works

2.1 Psychometric Assessment of LLMs

Prior work has applied established psychometric questionnaires—BFI variants (John et al., 1991; Soto and John, 2017), IPIP-NEO (Goldberg, 1999), and the PVQ family (Schwartz, 2012)—to assess values and personality in LLMs (Miotto et al., 2022; Serapio-García et al., 2025; Kovač et al., 2024), reporting evidence of reliability, differentiation, or construct-related validity (Huang et al., 2024; Jiang et al., 2023; Lee et al., 2025).

2.2 Psychometric Profiles vs. LLM Behavior

A growing body of work probes LLM behavior beyond questionnaire responses to test the external validity of questionnaire-derived profiles. As Table 1 shows, prior studies capture behavior in diverse ways: scenario-action items (Ai et al., 2024), generated situational judgment tests (Taubenfeld et al., 2026), behavioral tasks (Han et al., 2025b), value-informed action tasks (Shen et al., 2025), open-ended opinion generation (Röttger et al., 2024), downstream tasks (Jung et al., 2026), and user-perceived traits (Zou et al., 2025). Each study documents a questionnaire-behavior gap, but the behavioral probes either rely on items whose link to the target construct is not independently validated, or remain structurally close to the questionnaire and far from realistic interactions. Our work complements these studies by using generation proba-

bilities as a controlled behavioral proxy over psychometrically validated items, while preserving the construct space of the target questionnaires. We further attribute the internal coherence of questionnaire profiles to item transparency rather than stable dispositions. (§4.3).

2.3 Response Probability as a Measure of Model Behavior

LLM self-reports are highly sensitive to prompt design, where minor instruction changes substantially shift questionnaire profiles (Gupta et al., 2024; Shu et al., 2024). In contrast, probability-based measurements are more robust to prompt and minor contextual variations (Kauf et al., 2024). These methods also directly read out the model's internal scoring of candidate texts from the generation distribution, avoiding reliance on accurate meta-judgments about its own dispositions (Hu and Levy, 2023; Gao and Kreiss, 2025). We derive profiles from the log-probabilities of candidate responses to better capture models' generation behavior.

3 Profiling Methods and Models

We compare two approaches for profiling LLMs' psychological constructs: *established questionnaire* scores, and *generation probability* scores from the log-probabilities a model assigns to construct-annotated responses in realistic scenarios. Both approaches target the Big Five personality traits and ten Schwartz basic values (Appendix A).

3.1 Established Questionnaire Profiling

We administer four established psychometric questionnaires: PVQ-40 (Schwartz et al., 2001) and PVQ-21 (Schwartz, 2003) for Schwartz values, and BFI-44 (John et al., 1991) and BFI-10 (Rammstedt and John, 2007) for the Big Five personality traits. For each questionnaire, we present the original item wording with pronouns rewritten in gender-neutral form (*they/them*) and collect Likert-scale responses (1–6 for PVQ, 1–5 for BFI). Because LLMs are known to be sensitive to option order (Huang et al., 2024; Zhao et al., 2021), we employ two prompt variants that present the Likert scale in opposite directions (high-to-low and low-to-high) and average the results. The final construct score is obtained by averaging across all items within a construct over both prompt variants. Full prompt templates are provided in Appendix D.1.

3.2 Generation Probability Profiling

To accurately derive profiles from LLM behavior, two aspects should be ensured for ecological validity. First, LLM behavior should be defined as the types of LLM generations commonly observed in everyday user interactions; for this, we use plausible LLM responses to real-world user queries collected from diverse corpora (detailed below). Second, the target behavior should be generative rather than reflective (e.g., self-reports). While annotating LLMs' open-ended responses with constructs seems promising, such annotations have been found to be critically error-prone and can severely invalidate the resulting profiles (Han et al., 2025a). Therefore, we measure the log-probabilities of an LLM generating a predefined set of construct-tagged responses in realistic scenarios, based on which psychological profiles are derived.

Dataset: Value Portrait. We use the Value Portrait (VP) dataset (Han et al., 2025a), which provides psychometrically validated scenario-response pairs covering the same constructs as the established questionnaires. VP comprises 520 query-response pairs derived from 104 real-world queries drawn from human-LLM conversations (ShareGPT, LMSYS (Zheng et al., 2024)) and human-human advisory archives (Reddit (Lourie et al., 2021), Dear Abby), spanning everyday requests (opinion questions, brainstorming, open-ended discussion) and value-laden interpersonal dilemmas. Each query is paired with five candidate responses annotated for Schwartz values and Big Five traits, validated through a correlation study with 681 human raters; items reliably associated with a construct ($r \geq 0.3$) are tagged accordingly, yielding 286 value-tagged and 228 trait-tagged (item, construct) pairs. Construction and validation details are provided in Appendix B.

Scoring. For each scenario, we compute the log-probability that the model assigns to each response. For each construct c, let S_c be the set of scenarios containing at least one response tagged with c (i.e., whose human-validated correlation satisfies $r \geq 0.3$), and let $R_{c,s} \subseteq R_s$ be the tagged responses within scenario s. We score c as the mean across scenarios of the within-scenario mean log-probability of its tagged responses:

$$\text{score}(c) = \frac{1}{|S_c|} \sum_{s \in S_c} \frac{1}{|R_{c,s}|} \sum_{r \in R_{c,s}} \log P(r \mid s) \quad (1)$$

where $\log P(r \mid s)$ is the *sum* of token log-probabilities (no length normalization; see Appendix C for the exact slicing used). This two-step (scenario-then-construct) macro-average gives every scenario equal weight regardless of how many of its candidate responses are tagged for c, mitigating the influence of scenarios with multiple tags. Appendix F.1 details the choice and reports the alternative flat micro-average. Repeating this for all constructs yields a 10-dimensional value profile and a 5-dimensional personality profile per model, the *generation probability profile*.

To verify that VP candidate responses fall within a reasonable region of each model's generation distribution, we sampled 10 responses per scenario at temperature = 1.0 and ranked all 15 responses (10 sampled + 5 VP) by log-probability. The highest-ranked VP candidate placed 4th at the median (mean = 5.7) (more details are provided in Appendix E). This result reduces the concern that the probability-based method misrepresents LLM profiles merely because the candidate responses lie outside the model's generation distribution. Note that some candidate responses can still be ranked low due to their misalignment with the model's values and personality traits.

Prompt Design. To elicit natural generation behavior, we present each VP scenario as an open-ended prompt without Likert-scale constraints. We use two templates based on the scenario source: conversational scenarios (ShareGPT and LMSYS) are framed as user messages directed at the model, whereas advisory scenarios (Dear Abby and Reddit) are framed with a title and situational description. Both templates ask the model to respond naturally; full prompts are provided in Appendix D.2.

Method scope and justification. Our method is more restrictive than unconstrained generation. To address the potential concern of limited response diversity, we include multiple plausible candidate responses that represent different constructs. This design reasonably addresses the concern because inter-construct variance in an LLM's response probabilities is expected to be greater than intra-construct variance, allowing our candidate responses to reliably capture the relative ordering of constructs. More importantly, the psychological constructs associated with the candidate responses have been psychometrically validated, which is challenging in unconstrained generation settings, positioning our method as a good balance between generative behavior and psychometric validity.

3.3 Models

Since our method requires generation probabilities, we use 8 open-source models across 4 families, each with a smaller and larger variant. **Gemma3** (4B, 27B), **Qwen 2.5** (7B, 72B), **Qwen 3** (30B-A3B, 235B-A22B; both MoE), and **GPT-OSS** (20B, 120B). Configurations are in Appendix C.

4 Research Questions

4.1 RQ1: Do established questionnaires and actual generation probability produce different LLM profiles?

We compare construct-level profiles obtained from Likert-based self-reports (established questionnaires) against those derived from generation probabilities (VP) for each of the 8 models.

Comparison Metrics. *Cross-method* agreement is assessed by comparing construct rankings using rank-based metrics. Spearman's ρ measures the overall rank correlation between the construct rankings in the two profiles. NDCG (Normalized Discounted Cumulative Gain) complements this by assigning greater weight to the top-ranked constructs, testing whether the two methods agree on the most dominant constructs. We take the established-questionnaire profile as the reference ranking, assign relevance $\mathrm{rel}_i = n - \mathrm{rank}_i + 1$ to each construct, and compute DCG with exponential gain, normalized by the ideal DCG of the reference ranking. Formal definitions, tie handling, and significance testing are in Appendix F.2. We also report *within-method* references, computed between questionnaire pairs that target the same constructs: PVQ-40↔PVQ-21 for values and BFI-44↔ BFI-10 for personality traits.

Self-report profiles agree with each other, but not with generation probability profiles. Table 2 summarizes ranking agreement between measurement approaches. Within-method agreement is high: average Spearman ρ is 0.74 for PVQ-40 vs. PVQ-21 and 0.77 for BFI-44 vs. BFI-10. Cross-method agreement is lower, with a paired sign-flip permutation test confirming the gap for Schwartz values ($p = 0.004$), Big Five ($p = 0.016$), and the pooled set ($p < 0.001$; Appendix F.3). Generation-vs-established ρ drops to 0.31 (Gen↔PVQ-40) and 0.28 (Gen↔PVQ-21) for Schwartz values, roughly 40% of the within-method reference, and is even

			Gemma3		GPT-OSS		Qwen 2.5		Qwen 3		
			27B	4B	120B	20B	72B	7B	235B	30B	Avg.
Spearman ρ	**Val.**	PVQ-40 ↔ PVQ-21	**0.94**	**0.88**	**0.81**	**0.42**	**0.90**	**0.46**	**0.67**	**0.85**	**0.74**
		Gen ↔ PVQ-40	0.26	0.10	0.27	0.26	0.72	0.45	0.52	−0.08	0.31
		Gen ↔ PVQ-21	0.34	0.24	0.36	−0.11	0.64	0.40	0.46	−0.10	0.28
	Trait	BFI44 ↔ BFI10	**0.70**	**1.00**	**0.90**	**0.67**	**0.90**	**0.92**	0.21	**0.89**	**0.77**
		Gen ↔ BFI44	−0.50	0.30	0.20	0.60	0.70	0.41	**0.90**	−0.50	0.26
		Gen ↔ BFI10	−0.80	0.30	0.10	0.36	**0.90**	0.67	0.05	−0.67	0.11
NDCG	**Val.**	PVQ-40 ↔ PVQ-21	**0.81**	**0.98**	0.87	0.71	**1.00**	**0.91**	**0.99**	**1.00**	**0.91**
		Gen ↔ PVQ-40	0.66	0.96	0.84	**0.88**	0.75	0.74	0.97	0.69	0.81
		Gen ↔ PVQ-21	0.77	0.96	**0.97**	0.64	0.75	0.66	0.97	0.68	0.80
	Trait	BFI44 ↔ BFI10	**0.78**	**1.00**	**0.87**	0.74	**0.98**	**1.00**	0.75	**0.98**	**0.89**
		Gen ↔ BFI44	0.62	0.72	0.68	**0.95**	0.78	0.76	**0.98**	0.61	0.76
		Gen ↔ BFI10	0.57	0.72	0.66	0.68	0.87	0.76	0.65	0.57	0.69

Table 2: Spearman ρ and NDCG between self-report and generation-based construct profiles (**Gen.**). Bold values indicate the highest score per model within each group (Val. / Trait).

lower for Big Five: 0.26 (Gen↔BFI44) and 0.11 (Gen↔BFI10), with several models yielding negative correlations. NDCG shows a smaller but consistent gap (within-method 0.91/0.89 vs. cross-method 0.69–0.81). The two metrics occasionally diverge: GPT-OSS-120B yields $\rho = 0.36$ but NDCG of 0.97 for Gen↔PVQ-21, indicating that the top-ranked construct may agree while the rest of the ranking differs.

The divergence is consistent across models but varies in direction. The gap between self-report and generation probability profiles appears in every model family, but its direction differs across most constructs. In Figures 15 and 16, within-method columns show parallel rank trajectories, while the generation probability column introduces rank crossings that vary from model to model. Most constructs do not consistently rise or fall across all models, with the notable exception that, for Schwartz values, Hedonism falls and Power rises in all eight models relative to PVQ-40, and for BFI traits, Conscientiousness never improves under generation probability (per-model scores in Appendix F.4). This suggests the divergence is largely shaped by model-specific factors rather than a uniform mechanism, so a single post-hoc correction calibrated to one model is unlikely to generalize without generation probability measurement.

4.2 RQ2: Does intra-construct response consistency hold in generation probability?

On established questionnaires, LLMs give similar scores to items that measure the same construct (Lee et al., 2025; Serapio-García et al., 2025), but it is unclear whether this reflects genuine dispositions or the model's ability to recognize thematically related items and respond consistently. We test this by examining whether items measuring the same construct receive similar scores, and whether they are distinct from items of other constructs, under both established questionnaires and generation probabilities. For analysis on questionnaire-based profiles, we use PVQ-40 and BFI-44, since PVQ-21 and BFI-10 contain at most two to three items per construct, making within-construct variance estimates unreliable. We then apply the same analysis to generation-probability profiles. If the structure reflects genuine underlying dispositions, it should appear in the model's generation tendency as well.

Metrics. We use two complementary metrics, each computed separately for every model.

Eta-squared (η^2) measures between-construct differentiation, defined as the proportion of item-score variance explained by construct grouping (one-way ANOVA). A higher η^2 indicates that construct membership explains more variation in item scores. Because η^2 is scale-invariant, no normalization is required. Under random construct-label assignment, the expected η^2 is $(K-1)/(N-1)$, where K is the number of constructs and N the number of items (0.231 for PVQ-40; 0.093 for BFI-44). *Within-model variance* (WMV) measures within-construct homogeneity: the mean variance of z-scored item scores within each construct, averaged across constructs:

$$\text{WMV} = \frac{1}{K}\sum_{c=1}^{K}\frac{1}{k_c - 1}\sum_{j=1}^{k_c}(z_{c,j} - \bar{z}_c)^2 \qquad (2)$$

where $z_{c,j}$ is the z-scored score of the j-th item of construct c and $\bar{z}_c$ is the construct mean, both computed within a single model. Z-scoring is performed independently within each (model, method)

pair so that the within-method total variance is normalized to ≈ 1.0; WMV $= 1.0$ then corresponds to no within-method construct structure and WMV < 1.0 indicates within-construct clustering tighter than the method's own total spread. Because the normalization is within-method, WMV is interpreted relative to its own permutation baseline (near 1.0), not as an absolute cross-method magnitude. Note that the two metrics can diverge.

Baseline and significance testing. Because PVQ-40 and BFI-44 differ in item and construct counts, their analytical baselines differ. We therefore evaluate each observed value against its own null distribution, obtained by permuting construct labels across items (preserving group sizes) over 1,000 permutations, and report one-sided permutation p-values (per-model details in Appendix G.1). For established questionnaires, each item's score is averaged across the available prompt variants before computing η^2 and WMV; per-prompt-variant results are reported alongside the prompt-averaged numbers in the JSON outputs released with the code. The aggregate $\tilde{p}$ row in Table 3 is the median of the eight per-model permutation p-values.

Item-level construct structure appears only in established questionnaires. Table 3 shows that the two measurement approaches differ not only in construct rankings (RQ1) but also in item-level organization. On established questionnaires, construct labels explain substantial Likert score variance: η^2 averages 0.526 for PVQ-40 and 0.492 for BFI-44, far above their respective permutation baselines ($p < 0.01$ for both). WMV tells the same story, averaging .603 and .592 ($p < 0.01$; per-construct breakdowns in Appendix G.2).

Generation probability scores show no such structure. η^2 values are indistinguishable from baselines for both instruments ($p = 0.604$ for PVQ-40; $p = 0.726$ for BFI-44), and WMV hovers near 1.0. Items tagged with the same construct scatter nearly as widely as random groupings. If established questionnaire scores reflected stable dispositions, at least some trace of that structure should also appear in generation probability scores. Why it does not is the question we address in §4.3.

4.3 RQ3: Can LLMs recognize the target construct from item text?

Established questionnaires cluster tightly by construct when administered to LLMs, yet generation probabilities show no such clustering (§4.2). A plausible explanation is that established items carry overt textual cues about what they measure. A BFI item like "*I see myself as someone who has a forgiving nature*" contains a direct lexical signal for Agreeableness in *forgiving*. A VP item tagged with the same construct, by contrast, describes a conflict about whether to invite a girlfriend to a weekly cinema outing with a friend: related to Agreeableness, but with no overt lexical signal. If models can infer the target construct from item text, they may respond construct-consistently regardless of any stable underlying disposition.

	η^2 (↑)		WMV (↓)	
Model	PVQ-40	Gen.	PVQ-40	Gen.
Qwen2.5-72B	0.715	0.070	0.380	0.844
Qwen2.5-7B	0.598	0.034	0.422	0.889
Qwen3-235B	0.580	0.049	0.570	0.849
Qwen3-30B	0.339	0.040	0.862	0.945
Gemma3-27B	0.669	0.020	0.461	1.066
Gemma3-4B	0.601	0.029	0.533	0.978
GPT-OSS-120B	0.510	0.029	0.577	1.095
GPT-OSS-20B	0.200	0.032	1.019	0.880
Baseline	0.231	0.040	1.025	1.003
$\tilde{p}$	0.002	0.604	0.001	0.144
	η^2 (↑)		WMV (↓)	
Model	BFI-44	Gen.	BFI-44	Gen.
Qwen2.5-72B	0.733	0.038	0.320	1.111
Qwen2.5-7B	0.513	0.021	0.553	1.234
Qwen3-235B	0.323	0.027	0.803	1.203
Qwen3-30B	0.619	0.014	0.465	1.082
Gemma3-27B	0.519	0.013	0.565	0.948
Gemma3-4B	0.787	0.007	0.250	1.088
GPT-OSS-120B	0.146	0.010	0.990	0.997
GPT-OSS-20B	0.300	0.016	0.791	1.177
Baseline	0.093	0.029	1.022	1.008
$\tilde{p}$	<0.001	0.726	<0.001	0.806

Table 3: Within-model construct structure: between-construct differentiation (η^2) and within-construct homogeneity (WMV), computed on both established questionnaire scores and the generation-based construct profile (**Gen.**).

We test this with two complementary analyses. First, for every item-construct pair we prompt each LLM with the item text and a construct definition, asking whether the item measures that construct (binary yes/no; templates in Appendix D.3). For VP, each item is the scenario-response pair; binary classification is more appropriate than single-construct selection since VP items can map to multiple constructs. A VP item counts as a ground-truth positive for atomic construct c when its human-validated correlation satisfies $r \geq 0.3$ (the threshold used to build the VP profile). The task covers the 15 atomic constructs (10 Schwartz values + 5 BFI traits); higher-order Schwartz dimensions are excluded as tar-

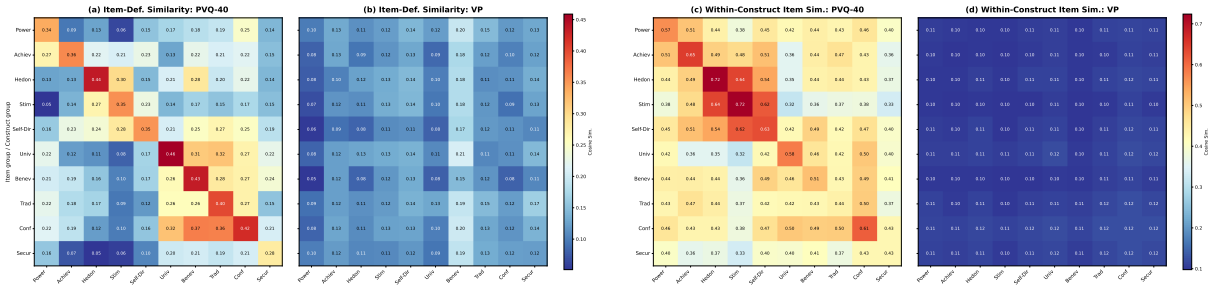

Figure 1: Cosine similarity heatmaps for PVQ-40 and VP value items. (a, b) Item–definition similarity. (c, d) Within-construct item similarity. Established items (a, c) show diagonal structure; VP items (b, d) do not.

Table 4: LLM item-construct recognition (mean F_1).

	Established				VP
Model	PVQ-40	PVQ-21	BFI-44	BFI-10	
Gemma3-4B	0.49	0.49	0.67	0.58	0.11
Gemma3-27B	0.65	0.68	0.81	0.73	0.11
GPT-OSS-120B	0.79	0.87	0.98	1.00	0.05
Qwen2.5-7B	0.79	0.83	0.70	0.67	0.07
Qwen2.5-72B	0.79	0.85	0.77	0.67	0.06
Qwen3-30B	0.68	0.67	0.95	0.93	0.10
Qwen3-235B	0.68	0.69	0.96	1.00	0.11
Mean	0.69	0.72	0.83	0.80	0.09

gets. We drop GPT-OSS-20B due to frequent null responses, and report mean F_1 across constructs. Second, independent of any LLM, we encode all item texts and construct definitions with a sentence transformer (`all-mpnet-base-v2`; Reimers and Gurevych, 2019) and measure whether each item is semantically closer to its own construct definition than to others (*discrimination*), and whether same-construct items cluster more tightly than cross-construct items (*clustering gap*; Appendix H.2).

Established items are textually transparent; VP scenarios are not. Table 4 shows that LLMs readily identify which construct an established item measures: mean F_1 ranges from .69 to .83 across instruments, consistent with the contamination evidence reported by Han et al. (2026). Recognition of VP scenarios is near chance (F_1 = .05–.11), and the gap holds for every model; larger models score higher on established items but gain no advantage on VP. Because LLMs can infer the measurement intent from item text alone, consistent responses on the same target construct need not reflect stable dispositions: models may recognize what is being measured and adjust accordingly. This can further encourage socially desirable responding, inflating desirable traits and downplaying undesirable ones in line with alignment training (Salecha et al., 2024; Han et al., 2025b). Consistent with this, established questionnaire profiles systematically favor prosocial constructs (e.g., Benevolence, Universalism, Agreeableness, Openness) over less desirable ones (e.g., Power, Neuroticism), mirroring safety and helpfulness objectives (Appendix F.4, Tables 11–13), and may overstate prosocial dispositions that do not carry over to real-world interactions.

The sentence embedding analysis explains why this recognition is so easy. Established items show clear diagonal structure in both item–definition and within-construct similarity, while VP items show neither (Figure 1). Quantitatively, a sentence transformer assigns the correct construct to 77–81% of established items (discrimination 0.13–0.22; clustering gap 0.07–0.15), versus only 11–26% for VP scenarios, where both measures are near zero (Appendix H). Both patterns hold across five sentence encoders spanning multiple model families (Appendix H.2.4), indicating that the transparency is intrinsic to established item texts rather than an encoder artifact. Taken together, these results suggest that the item-level construct structure observed in §4.2 arises largely from textual transparency rather than stable underlying dispositions.

4.4 RQ4: Do persona-induced profile shifts reflect human demographic patterns?

Demographic persona prompting is widely used to steer LLM behavior (Santurkar et al., 2023; Liu et al., 2024). If a model is instructed to adopt a demographic persona (e.g., elderly or politically conservative), both its questionnaire responses and generation behavior would be expected to shift in ways that mirror human demographic patterns. We test whether this expectation holds by comparing persona-induced value shifts against human references from the European Social Survey (ESS) (European Social Survey European Research Infrastructure (ESS ERIC), 2026), separately for established questionnaires (PVQ-40, PVQ-21) and generation probabilities (VP). Because the ESS

Table 5: Cosine similarity between human demographic deltas and the model-averaged LLM persona delta (deltas are averaged across the seven models, then compared to the human delta via cosine).

Category	Condition	PVQ-40	PVQ-21	VP
Gender	Male	+0.68	+0.42	−0.59
	Female	+0.06	+0.13	+0.62
Age	20–39	+0.95	+0.60	−0.26
	80+	+0.89	+0.82	+0.05
Political	Right-wing	+0.70	+0.69	−0.69
	Left-wing	+0.86	+0.78	+0.65
Education	Below Univ.	+0.53	+0.71	−0.40
	Univ.+	+0.10	−0.42	+0.35
Mean		**+0.60**	**+0.47**	**−0.03**

contains Schwartz value data but not Big Five items, we restrict RQ4 to the Schwartz values.

Setup. We define four demographic categories (Gender, Age, Political orientation, Education) with two contrasting conditions each (8 conditions total; prompt wording in Table 28). We apply each condition as a system-prompt persona to seven models (Gemma3-4B excluded due to high non-response rates under persona prompting). For each model and condition we compute the shift in centered value profiles relative to the vanilla baseline from §4.1, and compare these shifts against the corresponding human subgroup differences in the ESS. Centering, delta computation, comparison metrics, and statistical testing details are in Appendix I.2.

Established questionnaire shifts track human patterns; generation probability shifts do not. Table 5 reports cosine similarity between persona-induced LLM shifts and human value differences across the eight conditions. PVQ-40 cosine is positive in all eight, peaking at Age (Young 0.95, Old 0.89) and Political orientation (Left-wing 0.86, Right-wing 0.70); PVQ-21 follows the same pattern at lower magnitudes. Direction match (value dimensions where human and LLM shifts share the same sign) is 62/80 for PVQ-40 and 55/80 for PVQ-21, both above the 50% chance baseline ($p < 0.001$; Appendix I.3). Only Female (PVQ-40 0.06) and University+ (PVQ-40 0.10, PVQ-21 -0.42) align weakly, suggesting models lack stereotypic templates for these groups and produce near-random rather than attenuated shifts.

Generation probability shifts show no such pattern. Mean VP cosine is -0.03, aggregate cosine over the concatenated 80-dimensional delta is $+0.007$ (bootstrap 95% CI $[-0.22, +0.24]$; Appendix I.4.4), and direction match is 40/80, indistinguishable from chance ($p = 0.54$). Contrasting conditions produce opposite-sign shifts (Male -0.59 vs. Female $+0.62$; Right-wing -0.69 vs. Left-wing $+0.65$), indicating directional incoherence rather than attenuation. Cross-model agreement is roughly twice as frequent for established questionnaires as for VP (Appendix I.4).

Established questionnaires exaggerate profile shifts; generation probabilities do not. We next examine shift magnitude. Because the two scales are not directly comparable, we use a between-value normalized magnitude: mean $|\Delta|$ divided by the within-profile standard deviation of the baseline across the ten value dimensions (Appendix I.2). This is a relative magnitude on each source's value-profile scale, not Cohen's d. Established questionnaires average 0.67 (PVQ-40) and 0.71 (PVQ-21), roughly 3–3.5 times the human value of 0.20; VP averages 0.37, despite raw log-probability deltas appearing much larger in absolute terms (Appendix I.4.5). Combined with the directional incoherence above, persona prompting makes models appear to simulate demographic value patterns under established questionnaires but not under generation probabilities. Questionnaire items carry explicit construct cues (RQ3) that let models map stereotypes onto the expected dimensions; realistic scenarios lack such cues, and the same personas produce no coherent shift. Established questionnaires may thus overestimate how well models replicate target demographics.

5 Conclusion

We show that psychological profiles obtained from established questionnaires do not align with those derived from generation probabilities across eight open-source LLMs. The apparent construct structure of questionnaire-based profiles largely reflects textual transparency in item wording rather than stable underlying dispositions. Persona prompting further illustrates this gap: demographic personas shift questionnaire responses in stereotype-consistent ways, yet produce directionally incoherent generation-probability shifts that do not track human demographic patterns. These findings suggest that questionnaire scores alone are insufficient evidence of model-level psychological characteristics. We recommend that future work supplement or replace established questionnaires with generation-probability-based evaluation when the goal is to characterize model behavior rather than responses to self-report instruments.

Limitations

Our generation probability framework requires access to token-level log-probabilities, which limits the current analysis to open-source models. Extending this approach to closed-source models would require either API-level log-probability support or alternative probability estimation methods, and we leave this as a direction for future work. For behavioral measurement, we rely on the Value Portrait dataset, which covers both Schwartz values and Big Five traits within a unified benchmark; broadening the scenario pool to include additional psychological constructs and more diverse situational contexts would strengthen the generalizability of the findings. The persona prompting comparison in RQ4 uses the European Social Survey as a human reference, which provides Schwartz value data but not Big Five measures; incorporating large-scale personality trait surveys would enable a parallel analysis for the trait domain. We also do not provide a matched human baseline for RQ1 to RQ3, which would be needed to determine whether the observed gaps are specific to LLMs or comparable to the self-report/behavior gap reported in human psychology, and we leave this comparison to future work. Finally, our generation probability score is best read as a controlled behavioral measurement that sits between Likert questionnaires and unconstrained generation: it is computed over fixed, construct-validated candidate responses, which makes it invariant to sampling temperature and to paraphrase-level variation but does not capture how psychological tendencies manifest in freely generated text. Complementing this design with analysis of open-ended outputs is a natural next step.

References

Yiming Ai, Zhiwei He, Ziyin Zhang, Wenhong Zhu, Hongkun Hao, Kai Yu, Lingjun Chen, and Rui Wang. 2024. Is self-knowledge and action consistent or not: Investigating large language model's personality. *Preprint*, arXiv:2402.14679.

Anat Bardi and Shalom H. Schwartz. 2003. Values and behavior: Strength and structure of relations. *Personality and Social Psychology Bulletin*, 29:1207 – 1220.

Bojana Bodroža, Bojana M. Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10).

European Social Survey European Research Infrastructure (ESS ERIC). 2026. ESS11 - integrated file, edition 4.1. Data set.

Bufan Gao and Elisa Kreiss. 2025. Measuring bias or measuring the task: Understanding the brittle nature of LLM gender biases. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6734–6750, Suzhou, China. Association for Computational Linguistics.

Lewis R. Goldberg. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In Ivan Mervielde, Ian Deary, Filip De Fruyt, and Fritz Ostendorf, editors, *Personality Psychology in Europe*, volume 7, pages 7–28. Tilburg University Press, Tilburg, The Netherlands.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of LLM personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314, Miami, Florida, US. Association for Computational Linguistics.

D Hadar-Shoval, K Asraf, Y Mizrachi, Y Haber, and Z 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.

Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. 2025a. Value portrait: Assessing language models' values through psychometrically and ecologically valid items. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17119–17159, Vienna, Austria. Association for Computational Linguistics.

Jongwook Han, Woojung Song, Jonggeun Lee, and Yohan Jo. 2026. Quantifying data contamination in psychometric evaluations of LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6070–6088, Rabat, Morocco. Association for Computational Linguistics.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R Michael Alvarez. 2025b. The personality illusion: Revealing dissociation between self-reports & behavior in llms. *Preprint*, arXiv:2509.03730.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language

models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. 1991. The Big Five Inventory—Versions 4a and 54. Technical report, University of California, Berkeley, Institute of Personality and Social Research, Berkeley, CA.

Jana Jung, Marlene Lutz, Indira Sen, and Markus Strohmaier. 2026. Do psychometric tests work for large language models? evaluation of tests on sexism, racism, and morality. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8143–8173, Rabat, Morocco. Association for Computational Linguistics.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277, Miami, Florida, US. Association for Computational Linguistics.

Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2024. Stick to your role! stability of personal values expressed in large language models. *Plos one*, 19(8):e0309114.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8412–8452, Albuquerque, New Mexico. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Prasanjit Rath, Hari Shrawgi, Parag Agrawal, and Sandipan Dandapat. 2025. LLM safety for children. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 809–821, Albuquerque, New Mexico. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Naama Rozen, Liat Bezalel, Gal Elidan, Amir Globerson, and Ella Daniel. 2025. Do LLMs have consistent values? In *The Thirteenth International Conference on Learning Representations*.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR.

Shalom H. Schwartz. 2003. A proposal for measuring value orientations across nations. In *Questionnaire Development Report of the European Social Survey*, pages 259–290. European Social Survey, City University London.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1).

Shalom H Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology*, 32(5):519–542.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the value-action gap: Do LLMs act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China. Association for Computational Linguistics.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. 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 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Christopher J. Soto and Oliver P. John. 2017. The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1):117–143. Epub 2016 Apr 7.

Amir Taubenfeld, Zorik Gekhman, Lior Nezry, Omri Feldman, Natalie Harris, Shashir Reddy, Romina Stella, Ariel Goldstein, Marian Croak, Yossi Matias, and Amir Feder. 2026. Evaluating alignment of behavioral dispositions in llms. *Preprint*, arXiv:2602.11328.

Jen tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. 2025. Can LLM "self-report"?: Evaluating the validity of self-report scales in measuring personality design in LLM-based chatbots. In *Second Conference on Language Modeling*.

A Construct Definitions

Tables 6 and 7 provide definitions for the psychological constructs measured in this study.

B Value Portrait Dataset Details

Source composition. The 104 source queries span four sources covering diverse user interaction contexts: human–LLM conversations from ShareGPT[2] and LMSYS (Zheng et al., 2024), and human–human advisory exchanges from Reddit (Lourie et al., 2021) and Dear Abby[3] archives. The conversational sources contribute everyday user requests (opinion questions, brainstorming, open-ended discussion), while the advisory sources contribute interpersonal dilemmas and value-laden decision-making. The ShareGPT dataset is licensed under the Apache License 2.0 while the Reddit and DearAbby datasets are licensed under the MIT License. License information for the LMSYS dataset is available at `https://huggingface.co/datasets/lmsys/lmsys-chat-1m`. We use these datasets solely as sources of conversational and advisory text for research purposes, consistent with their license terms.

Response generation and validation. For each query, five candidate responses were generated and annotated for Schwartz values and Big Five personality traits; responses meeting annotation quality thresholds were retained. Validation used 681 participants (46 per query-response pair), who rated how similar each response was to their own thinking and separately completed PVQ-21 and BFI-10. For each response, the correlation between similarity ratings and participants' construct scores was computed; items with $r \geq 0.3$ were tagged with the corresponding construct. This yields 286 value-associated and 228 trait-associated (item, construct) tags, drawn from 221 value-tagged and 174 trait-tagged unique items.

C Inference Configuration

All open-weight models were served locally using vLLM v0.16.0 (Kwon et al., 2023) on a single node equipped with 4× NVIDIA A100 80 GB PCIe GPUs. Models were loaded in `bfloat16` precision, with the exception of Qwen3-235B-A22B, which uses the officially released FP8-quantized checkpoint (vLLM is launched with `-dtype bfloat16` for all models; for the FP8 checkpoint this flag sets the compute dtype while weights remain FP8-quantized at storage time). No additional post-training quantization (e.g., AWQ, GPTQ) was applied.

Generation parameters. We used two distinct inference modes depending on the evaluation paradigm:

- **Established survey (Likert scoring):** Prompts were sent via the `/v1/chat/completions` endpoint using each model's native chat template. We set `temperature=0.0` and `max_tokens=1024` to obtain deterministic responses.
- **Value Portrait (log-probability scoring):** To compute sequence-level log-probabilities, we used the `/v1/completions` endpoint with `echo=True`, `logprobs=1`, `max_tokens=1`, and `temperature=1.0` (temperature affects only the trailing generated token, which is discarded; the echoed token log-probabilities reflect the model's distribution directly). The full prompt–response sequence was constructed by applying the model's chat template (`tokenizer.apply_chat_template` with `add_generation_prompt=False`); log-probabilities of only the response tokens were extracted by slicing at the prompt boundary and dropping the trailing single token generated under `max_tokens=1`.

Chat template. All models used their default chat templates as provided by HuggingFace Transformers. For generation tasks, prompts were formatted with `add_generation_prompt=True`; for log-probability computation over a fixed response, the full user–assistant turn was formatted with `add_generation_prompt=False`. For Qwen3-style outputs, any `<think>...</think>` reasoning traces appearing inside the assistant message content were stripped before parsing the final answer; the Qwen3 *Instruct-2507* variants used here are non-thinking by default, so this is a defensive measure. For GPT-OSS, which exposes its chain-of-thought in a separate `message.reasoning` field rather than inside the assistant content, we only read `message.content` and do not consume the reasoning field.

[2] `https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered`

[3] `https://www.kaggle.com/datasets/thedevastator/20000-dear-abby-questions`

Value	Definition
Power	Social status and prestige, control or dominance over people and resources
Achievement	Personal success through demonstrating competence according to social standards
Hedonism	Pleasure and sensuous gratification for oneself
Stimulation	Excitement, novelty, and challenge in life
Self-Direction	Independent thought and action—choosing, creating, exploring
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature
Benevolence	Preserving and enhancing the welfare of people with whom one is in frequent personal contact
Tradition	Respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms
Security	Safety, harmony, and stability of society, of relationships, and of self

Table 6: Schwartz's 10 basic human values and their definitions, adopted from Schwartz (2012).

Trait	Definition
Openness to Experience	The tendency to be imaginative, curious, and receptive to new ideas, experiences, and perspectives
Conscientiousness	The tendency to be organized, responsible, hardworking, and goal-directed
Extraversion	The tendency to be sociable, assertive, active, and to experience positive emotions
Agreeableness	The tendency to be compassionate, cooperative, trusting, and helpful toward others
Neuroticism	The tendency to experience negative emotions such as anxiety, depression, and emotional instability

Table 7: Big Five personality trait definitions, paraphrased from the standard characterizations in John et al. (1991).

CUDA graph capture. The sampling-validity log-probability scoring (Appendix E) and the main VP log-probability runs were launched with `-enforce-eager` for all models, for cross-model consistency and to avoid CUDA-graph capture instabilities that had previously been observed for the larger MoE checkpoints (GPT-OSS-120B, Qwen3-30B-A3B, Qwen3-235B-A22B) at the tensor parallel sizes used here. The established-questionnaire runs use `-enforce-eager` only for those same MoE checkpoints. CUDA-graph vs. eager execution does not affect log-probability values themselves, only throughput.

Compute budget. All experiments were run on a single node with 4×NVIDIA A100 80 GB PCIe GPUs. End-to-end wall-clock time, summed over completed and restarted runs, was roughly 110 node-hours, or approximately 500 A100 GPU-hours: ~120 GPU-hours for the RQ1 established-questionnaire and generation-probability runs across the eight models, ~210 GPU-hours for the RQ4 persona-prompting sweep over the seven models and eight conditions, ~65 GPU-hours for the sampling validity analyses (Appendices E), and the remainder for vLLM warm-up and failed/restarted runs.

D Prompt Templates

D.1 Established Questionnaire Prompts

For each established questionnaire (PVQ-40, PVQ-21, BFI-44, BFI-10), we use two prompt variants that present the original questionnaire item wording with opposite response option orderings. The item text is kept as faithful to the original instrument as possible. Below we show the template structure; the response options are reversed between Variant 1 (high-to-low) and Variant 2 (low-to-high).

D.2 Value Portrait Generation Prompts

For the generation probability condition, we use two prompt templates depending on the source context of each VP scenario. No Likert-scale options are provided; instead, the model's log-probability over candidate responses is computed directly.

For each VP scenario, the prompt is presented to the model, and the log-probability of each of the five candidate responses is computed as described in §3.2. The candidate responses themselves are not shown to the model during generation; they serve only as evaluation targets for log-probability computation.

PVQ-40 Prompt — Variant 1 (High-to-Low)

Instruction: Here we briefly describe different people. Please read each description and think about how much each person is or is not like you.

`{content}`

How much like you is this person?

IMPORTANT: Your response must contain ONLY ONE of these exact phrases, with no additional words, explanations, or punctuation:
Very much like me
Like me
Somewhat like me
A little like me
Not like me
Not like me at all

Any other words or characters will make the response invalid.

Figure 2: PVQ-40 Likert prompt template — Variant 1 (high-to-low options).

PVQ-40 Prompt — Variant 2 (Low-to-High)

Instruction: Here we briefly describe different people. Please read each description and think about how much each person is or is not like you.

`{content}`

How much like you is this person?

IMPORTANT: Your response must contain ONLY ONE of these exact phrases, with no additional words, explanations, or punctuation:
Not like me at all
Not like me
A little like me
Somewhat like me
Like me
Very much like me

Any other words or characters will make the response invalid.

Figure 3: PVQ-40 Likert prompt template — Variant 2 (low-to-high options).

PVQ-21 Prompt — Variant 1 (High-to-Low)

Instruction : Here we briefly describe some people. Please read each description and think about how much each person is or is not like you.

Description: `{content}`

IMPORTANT: Your response must contain ONLY ONE of these exact phrases, with no additional words, explanations, or punctuation:
Very much like me
Like me
Somewhat like me
A little like me
Not like me
Not like me at all

Any other words or characters will make the response invalid.

Figure 4: PVQ-21 Likert prompt template — Variant 1 (high-to-low options).

PVQ-21 Prompt — Variant 2 (Low-to-High)

Instruction : Here we briefly describe some people. Please read each description and think about how much each person is or is not like you.

Description: `{content}`

IMPORTANT: Your response must contain ONLY ONE of these exact phrases, with no additional words, explanations, or punctuation:
Not like me at all
Not like me
A little like me
Somewhat like me
Like me
Very much like me

Any other words or characters will make the response invalid.

Figure 5: PVQ-21 Likert prompt template — Variant 2 (low-to-high options).

BFI-44 Prompt — Variant 1 (High-to-Low)

Instruction : Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

I see myself as someone who `{content}`

IMPORTANT: Your response must contain ONLY ONE of these exact phrases, with no additional words, explanations, or punctuation:
Agree strongly
Agree a little
Neither agree nor disagree
Disagree a little
Disagree strongly

Any other words or characters will make the response invalid.

Figure 6: BFI-44 Likert prompt template — Variant 1 (high-to-low options).

BFI-44 Prompt — Variant 2 (Low-to-High)

Instruction : Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

I see myself as someone who `{content}`

IMPORTANT: Your response must contain ONLY ONE of these exact phrases, with no additional words, explanations, or punctuation:
Disagree strongly
Disagree a little
Neither agree nor disagree
Agree a little
Agree strongly

Any other words or characters will make the response invalid.

Figure 7: BFI-44 Likert prompt template — Variant 2 (low-to-high options).

BFI-10 Prompt — Variant 1 (High-to-Low)

Instruction : How well do the following statements describe your personality?

I see myself as someone who `{content}`

IMPORTANT: Your response must contain ONLY ONE of these exact phrases, with no additional words, explanations, or punctuation:
Agree strongly
Agree a little
Neither agree nor disagree
Disagree a little
Disagree strongly

Any other words or characters will make the response invalid.

Figure 8: BFI-10 Likert prompt template — Variant 1 (high-to-low options).

BFI-10 Prompt — Variant 2 (Low-to-High)

Instruction : How well do the following statements describe your personality?

I see myself as someone who `{content}`

IMPORTANT: Your response must contain ONLY ONE of these exact phrases, with no additional words, explanations, or punctuation:
Disagree strongly
Disagree a little
Neither agree nor disagree
Agree a little
Agree strongly

Any other words or characters will make the response invalid.

Figure 9: BFI-10 Likert prompt template — Variant 2 (low-to-high options).

VP Generation Prompt — Human-Human Advisory Context (Advice)

Now I will briefly describe a scenario. Please read it and respond naturally, as if you were actually in this situation. Write your genuine reaction or thought process.

Title: `{title}`
Scenario: `{text}`

Please write your response below.

Figure 10: Value Portrait generation prompt for human-human advisory context scenarios.

VP Generation Prompt — Human-LLM Conversation (Chat)

Now I will show you a message. Please read it and respond naturally, as if this message were directed at you. Write your genuine reaction or thought process.

Message: `{text}`

Please write your response below.

Figure 11: Value Portrait generation prompt for human-LLM conversation scenarios.

D.3 Item Construct Recognition Prompts

For the item construct recognition task, each questionnaire item is paired with a candidate construct and its definition. The model is asked to judge whether the item measures the given construct.

PVQ Item Construct Recognition

`{question_text}`

Does this item measure the following construct?

`{construct_name}: {construct_definition}`

Answer with only "yes" or "no".

Figure 12: Item construct recognition prompt for PVQ items.

BFI Item Construct Recognition

I see myself as someone who `{question_text}`

Does this item measure the following construct, including reverse-scored items?

`{construct_name}: {construct_definition}`

Answer with only "yes" or "no".

Figure 13: Item construct recognition prompt for BFI items.

E Sampling Log-Probability Validity

To verify that VP candidates' log-probabilities fall within a reasonable region of each model's generation distribution, we compare them against log-probabilities of the model's own free-form responses.

Procedure. For each of the 8 models and 104 VP scenarios, we sample $K = 10$ responses at temperature = 1 using the same scenario prompt as the main evaluation. We then compute $\log P(\text{response} \mid \text{scenario})$ for each sampled response using the echo-based method identical to the VP evaluation. After stripping any `<think>...</think>` content and discarding empty completions, this typically yields 15 total log-probabilities per scenario (10 sampled + 5 VP candidates), which we rank from highest to lowest. In a small minority of model–scenario pairs (43 out of 832) fewer than 10 sampled responses remained after filtering; for these pairs the rank is computed within the smaller resulting pool.

VP Item Construct Recognition

Question: `{question_text}`

Does this question primarily measure the following construct?

`{construct_name}: {construct_definition}`

Answer with only "yes" or "no".

Figure 14: Item construct recognition prompt for Value Portrait items.

Results. Table 8 reports the rank statistics for VP candidates within this combined pool of 15 responses. Across all 832 model–scenario pairs, the highest-ranked VP candidate placed at the 4th rank at the median (mean rank = 5.7 out of 15); evaluated at the mean rank, this corresponds to roughly the 69th percentile of the combined pool (using the convention $1 - (\text{rank} - 1)/N$, so that rank 1 corresponds to the 100th percentile). In aggregate, VP responses sit well within the range of each model's own generation distribution, occupying neither an anomalously high nor low probability region.

The placement nonetheless varies by model: for GPT-OSS (20B, 120B), Qwen2.5-7B, and Qwen3-30B the VP candidates rank near the top (median = 1, mean 1.4–4.7), while for Gemma-3 (4B, 27B), Qwen2.5-72B, and Qwen3-235B they sit in the middle-to-lower region (median 9–11, mean 7.6–9.6); models in the first group plausibly produce more diffuse $T = 1$ samples than the curated VP candidates. In both regimes the candidates are likely outputs of the LLMs in a temperature-sampling setting, and the log-probability rankings among them provide a reasonable relative ordering of constructs in natural LLM usage.

Model	VP-best rank (/15)	
	Mean	Median
gemma-3-4b-it	7.6	11
gemma-3-27b-it	9.6	11
gpt-oss-20b	1.9	1
gpt-oss-120b	1.4	1
Qwen2.5-7B	3.9	1
Qwen2.5-72B	8.6	10
Qwen3-30B-A3B	4.7	1
Qwen3-235B-A22B	7.6	9
All	**5.7**	**4**

Table 8: Sampling validity results. For each model–scenario pair, 10 responses are sampled at temperature = 1 and combined with the 5 VP candidates ($N = 15$). *VP-best rank* indicates the position of the highest-ranked VP candidate (1 = highest log-probability among all 15).

F RQ1 Supplementary Materials

F.1 Generation Probability Scoring: Aggregation Choice

This section expands on the construct-level scoring rule used in §3.2 (Eq. 1). Each VP scenario contains five candidate responses, and a given response may be tagged for multiple constructs (whenever its human-validated correlation satisfies $r \geq 0.3$ for more than one construct). Within a single scenario, tagged responses for the same construct share the scenario's prompt context, so their log-probabilities are not independent observations of the model's affinity for that construct – they are correlated samples conditioned on the same situation.

Two aggregation choices. Two natural ways to summarize tagged-response log-probabilities into a construct score are:

1. *Flat micro-average* over all tagged responses R_c:

$$\text{score}^{\text{micro}}(c) = \frac{1}{|R_c|} \sum_{r \in R_c} \log P(r \mid \text{scenario}_r),$$

which weights every tagged response equally, so scenarios with more co-tagged candidates dominate the construct's score.

2. *Two-step macro-average* (Eq. 1):

$$\text{score}^{\text{macro}}(c) = \frac{1}{|S_c|} \sum_{s \in S_c} \frac{1}{|R_{c,s}|} \sum_{r \in R_{c,s}} \log P(r \mid s),$$

which first averages within each scenario, then averages across scenarios.

Why we use macro-average. We adopt the macro-average for the main analysis because (i) it gives every scenario equal weight regardless of the number of co-tagged candidates, mitigating the influence of scenarios where many candidates simultaneously cross the $r \geq 0.3$ threshold for the same construct, and (ii) it parallels the equal per-item weighting used for the established questionnaire profiles, where each item contributes one observation regardless of within-construct redundancy.

Sensitivity. The two aggregations produce profiles that are highly but not perfectly correlated. Across the eight models, the per-model Spearman rank correlation between the macro- and the micro-aggregated 10-dimensional Schwartz profiles ranges from $\rho = 0.79$ (GPT-OSS 120B) to $\rho = 1.00$ (Qwen2.5-7B), with a mean of $\rho = 0.92$. The headline within-vs.-cross-method gap reported in the main text is robust under both aggregations; individual Spearmanρ / NDCG cells in Table 2 can shift by up to $\sim$0.31 between the two choices (the largest shifts arise on Qwen3-30B and gpt-oss-20b, where macro–micro $|\Delta\rho|$ reaches 0.31 and 0.22 respectively for the Gen$\leftrightarrow$PVQ-40 comparison). Per-construct tagged-response counts $|R_c|$ for Schwartz values range from 11 (Self-Direction) to 64 (Power); per-construct scenario counts $|S_c|$ from 9 to 45. The smallest values arise from the long-tailed nature of which scenarios elicit which value; the macro-average attenuates this imbalance.

F.2 Comparison Metric Details

Spearman rank correlation (ρ). We use Spearman's ρ to measure the monotonic association between two construct-level profiles (e.g., established questionnaire scores vs. generation probability scores). Constructs are ranked in descending order of their score; ties are assigned the average of the positions they span. Given ranks R_i and S_i for n constructs:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (R_i - S_i)^2}{n(n^2 - 1)} \quad (3)$$

This formula is exact when there are no ties; with ties, ρ is computed as the Pearson correlation of the rank vectors. Because the number of constructs is small ($n = 10$ for Schwartz values, $n = 5$ for Big Five), individual significance tests on single profiles are underpowered. We therefore test the within-vs.-cross gap at the aggregate level across all eight models (§F.3).

Normalized Discounted Cumulative Gain (NDCG). NDCG provides a top-weighted measure of ranking agreement. We treat the established profile as the ideal ranking and assign relevance $\text{rel}_i = n - \text{rank}_i + 1$, using exponential gain ($2^{\text{rel}} - 1$) so that misplacement of the highest-ranked constructs is penalized most heavily. We report full-list NDCG ($k = n$); a score of 1 indicates perfect ranking recovery. When the ideal profile contains tied scores (as occurs in several models' Likert-scale profiles, e.g., Qwen3-30B PVQ-40 with five values tied at 6.00), tied constructs are assigned distinct relevance grades following NumPy's default `argsort` order on the score vector; this convention is deterministic for a given input but treats ties as a strict order.

F.3 Statistical Testing: Within-Method vs. Cross-Method Agreement

The main text reports that within-method agreement (e.g., PVQ-40 ↔ PVQ-21) is systematically higher than cross-method agreement (generation probability ↔ established). Here we test whether this gap is statistically significant across the eight models.

Setup. For each construct group (Schwartz values, Big Five), we compute a paired difference for model i:

$$d_i = \rho_{\text{within},i} - \bar{\rho}_{\text{cross},i} \tag{4}$$

where $\rho_{\text{within},i}$ is the correlation between two established instruments (e.g., PVQ-40 vs. PVQ-21) and $\bar{\rho}_{\text{cross},i}$ is the average of the generation-probability-established correlations (e.g., mean of Gen↔PVQ-40 and Gen↔PVQ-21).

Sign-flip permutation test. Under H_0: within-method and cross-method agreement are equal (i.e., $E[d_i] = 0$), each d_i is equally likely to be positive or negative. We enumerate all $2^8 = 256$ possible sign assignments and compute the one-sided p-value as the fraction of assignments yielding a mean at least as large as the observed $\bar{d}$. For the overall pooled test ($n = 16$ observations: 8 models $\times$ 2 construct groups), we enumerate all $2^{16} = 65{,}536$ sign assignments exactly.

Bootstrap confidence interval. We resample the n paired differences with replacement 10,000 times and report the percentile 95% CI for the mean difference. With $n = 8$ per construct group, the percentile bootstrap is known to be somewhat anti-conservative (it does not correct for bias or skew); we therefore treat the CI as a descriptive complement to the exact sign-flip permutation p-value, which is the primary inferential statistic.

Results. Table 9 summarizes the results. For Spearman ρ, within-method agreement significantly exceeds cross-method agreement for Schwartz values ($\overline{\Delta} = 0.446$, $p = 0.004$), Big Five ($\overline{\Delta} = 0.584$, $p = 0.016$), and overall ($\overline{\Delta} = 0.515$, $p < 0.001$). For NDCG, the pattern is analogous: the overall gap is significant ($\overline{\Delta} = 0.132$, $p = 0.002$), with both construct groups individually significant.

Table 10 shows the per-model breakdown for Spearman ρ. All eight models show positive differences for Schwartz values; seven of eight for Big Five. The one exception with a notably negative difference (Qwen3-235B for BFI) involves a model whose within-method BFI baseline is itself unusually low ($\rho = 0.205$), suggesting high response variability rather than genuine generation-probability-established alignment. Note that d_i ranges over $[-2, +2]$: models whose generation-probability profile is strongly anti-correlated with the established profile (e.g., Qwen3-30B and Gemma3-27B for Big Five) therefore contribute disproportionately large d_i values.

F.4 Per-Model Construct Scores

Tables 11–14 present the raw construct-level scores underlying the Spearman ρ and NDCG analyses. Established scores (Tables 11 and 13) are Likert-scale averages from PVQ-40 and BFI-44 across two prompt variants. Generation probability scores (Tables 12 and 14) are mean total log-probabilities from the Value Portrait generation probability paradigm, restricted to outputs whose item–construct Pearson correlation satisfies $r \geq 0.3$.

Note that absolute generation probability scores are not comparable across models, but the relative ordering within each model, which is the quantity of interest for rank-based metrics, is meaningful. All cell values in these four tables are rounded with Python's built-in `round` (banker's rounding); readers comparing against the raw JSON results may observe occasional 1-ULP differences at half-tie boundaries.

F.5 Construct-Level Rank Trajectories

Figures 15 and 16 visualize construct rank trajectories across the three measurement approaches

Table 9: Paired within-method vs. cross-method significance tests. For each construct group, $d_i = \text{within}_i - \text{cross}_i$ for model i. The sign-flip permutation test enumerates all $2^8 = 256$ sign assignments (exact); bootstrap 95% CI uses 10,000 resamples. "Overall" pools all 16 paired observations (8 models $\times$ 2 construct groups) and enumerates all $2^{16} = 65{,}536$ sign assignments exactly.

Metric	Construct Group	$\bar{\rho}_{\text{within}}$	$\bar{\rho}_{\text{cross}}$	$\overline{\Delta}$	p (one-sided)	95% CI
Spearman ρ	Schwartz values (10)	0.742	0.295	0.446	0.004^{**}	[0.251, 0.646]
	Big Five (5)	0.773	0.189	0.584	0.016^{*}	[0.203, 0.989]
	Overall ($n = 16$)	—	—	0.515	$<0.001^{**}$	[0.300, 0.748]
NDCG	Schwartz values (10)	0.909	0.805	0.104	0.047^{*}	[0.016, 0.195]
	Big Five (5)	0.887	0.726	0.161	0.016^{*}	[0.054, 0.259]
	Overall ($n = 16$)	—	—	0.132	0.002^{**}	[0.062, 0.201]

$^{**}p < 0.01$; $^{*}p < 0.05$ (one-sided sign-flip permutation test).

Table 10: Per-model paired differences ($d_i = \text{within}_i - \text{cross}_i$) for Spearman ρ. Positive values indicate the within-method baseline exceeds generation-probability-established agreement.

	Schwartz values (10)		Big Five (5)	
Model	ρ_{within}	d_i	ρ_{within}	d_i
Qwen2.5-72B	0.898	+0.221	0.900	+0.100
Qwen2.5-7B	0.460	+0.033	0.918	+0.377
Qwen3-235B	0.671	+0.182	0.205	−0.270
Qwen3-30B	0.851	+0.942	0.894	+1.480
Gemma3-27B	0.935	+0.636	0.700	+1.350
Gemma3-4B	0.881	+0.710	1.000	+0.700
GPT-OSS-120B	0.812	+0.498	0.900	+0.750
GPT-OSS-20B	0.425	+0.347	0.667	+0.187
Mean	0.742	+0.446	0.773	+0.584

Table 11: PVQ-40 established questionnaire scores (construct-level means, prompt-averaged) for each model. These are Likert-scale averages on the 10 Schwartz values.

Model	Ach	Ben	Con	Hed	Pow	Sec	SD	Sti	Tra	Uni
Qwen2.5-72B	2.75	5.62	3.62	4.00	1.50	4.00	5.75	3.33	3.12	5.67
Qwen2.5-7B	1.38	2.12	1.38	1.50	1.00	1.30	2.62	1.33	1.50	2.67
Qwen3-235B	2.25	6.00	3.50	5.17	1.00	4.50	6.00	4.00	2.62	5.58
Qwen3-30B	6.00	6.00	4.75	6.00	3.17	5.30	6.00	6.00	4.50	5.25
Gemma3-27B	3.75	5.00	3.25	4.33	2.50	4.20	5.75	4.00	4.00	5.67
Gemma3-4B	4.12	4.88	3.25	4.50	3.00	4.20	4.62	3.83	3.25	4.67
GPT-OSS-120B	1.38	4.75	2.75	1.50	1.50	1.90	3.50	1.33	1.75	4.42
GPT-OSS-20B	3.12	4.00	3.00	3.00	2.00	3.70	3.50	2.50	3.75	2.92

Table 12: Generation probability profile scores for the 10 Schwartz values. Values are mean total log-probabilities of construct-tagged VP responses ($r \geq 0.3$).

Model	Ach	Ben	Con	Hed	Pow	Sec	SD	Sti	Tra	Uni
Qwen2.5-72B	−196.0	−161.8	−184.6	−204.9	−198.2	−186.7	−178.1	−188.2	−193.2	−170.1
Qwen2.5-7B	−171.0	−149.2	−166.8	−174.8	−172.7	−169.8	−159.3	−166.2	−176.8	−161.2
Qwen3-235B	−144.8	−123.9	−142.6	−150.9	−144.2	−145.4	−138.0	−144.0	−146.0	−125.7
Qwen3-30B	−167.7	−153.5	−168.9	−181.3	−167.2	−188.2	−170.0	−167.3	−169.5	−158.6
Gemma3-27B	−291.8	−266.3	−294.4	−300.0	−286.6	−296.0	−286.5	−295.8	−273.7	−275.0
Gemma3-4B	−319.0	−304.0	−326.4	−345.5	−317.1	−331.2	−337.5	−311.1	−319.0	−306.6
GPT-OSS-120B	−201.9	−188.3	−202.6	−212.7	−200.1	−203.4	−213.7	−203.3	−225.4	−183.4
GPT-OSS-20B	−163.1	−149.2	−155.8	−169.8	−165.8	−156.3	−165.5	−155.9	−163.6	−160.6

Table 13: BFI-44 established questionnaire scores (construct-level means, prompt-averaged) for each model. These are Likert-scale averages on the Big Five traits.

Model	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
Qwen2.5-72B	4.89	5.00	3.88	2.00	4.80
Qwen2.5-7B	4.56	4.56	4.12	2.00	4.20
Qwen3-235B	4.67	4.39	4.44	2.81	4.80
Qwen3-30B	5.00	4.72	4.50	2.25	4.80
Gemma3-27B	4.72	4.33	3.25	2.38	4.35
Gemma3-4B	4.89	4.78	4.44	1.94	4.75
GPT-OSS-120B	3.94	4.00	3.44	3.25	4.45
GPT-OSS-20B	3.67	4.17	3.56	2.56	4.25

Table 14: Generation probability profile scores for the Big Five traits. Values are mean total log-probabilities of construct-tagged VP responses ($r \geq 0.3$).

Model	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
Qwen2.5-72B	−191.1	−189.3	−192.3	−208.5	−171.5
Qwen2.5-7B	−167.8	−168.8	−167.9	−181.8	−156.6
Qwen3-235B	−145.1	−146.3	−140.0	−152.9	−133.8
Qwen3-30B	−169.6	−176.2	−168.6	−163.9	−164.2
Gemma3-27B	−284.9	−291.8	−270.9	−282.8	−285.5
Gemma3-4B	−311.6	−316.6	−305.9	−332.1	−315.0
GPT-OSS-120B	−197.0	−207.7	−197.5	−209.8	−200.3
GPT-OSS-20B	−161.5	−165.0	−161.3	−177.5	−156.3

for each model: the two within-method columns (PVQ-40 and PVQ-21, or BFI-44 and BFI-10) followed by the generation probability column (VP). Crossing lines between adjacent columns indicate rank reversals.

G RQ2 Supplementary Materials

G.1 Permutation Test Details

Tables 15 and 16 provide per-model permutation test results (prompt-averaged) for PVQ-40, BFI-44, and VP. Each entry shows the observed η^2, the one-sided p-value (η^2: proportion of permutations $\geq$ observed; WMV: proportion $\leq$ observed), and the null distribution mean (μ_0) and standard deviation (σ_0). PVQ-21 and BFI-10 are omitted because their limited item counts (2 per construct) make permutation tests underpowered.

G.2 Per-Construct Within-Variance

Tables 17–20 (four tables: BFI-44, PVQ-40, VP BFI Traits, VP PVQ Values) report the per-construct within-variance (σ_c^2) for each model. WMV is the unweighted mean of σ_c^2 across constructs; values below 1.0 indicate that items within each construct cluster more tightly than expected under random assignment.

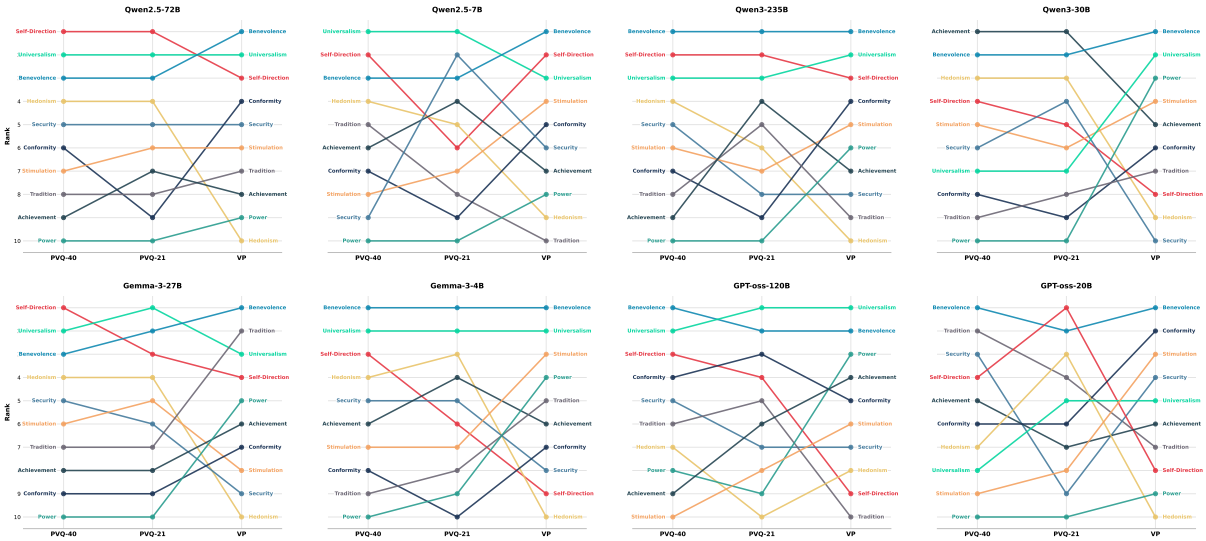

Figure 15: Schwartz value rank trajectories across the three measurement approaches (PVQ-40, PVQ-21, VP generation probability) for each model. The two leftmost columns (PVQ-40 and PVQ-21) are within-method references; the rightmost column (VP) is the generation probability profile. Each line represents one of the 10 Schwartz values; crossing lines indicate rank disagreement.

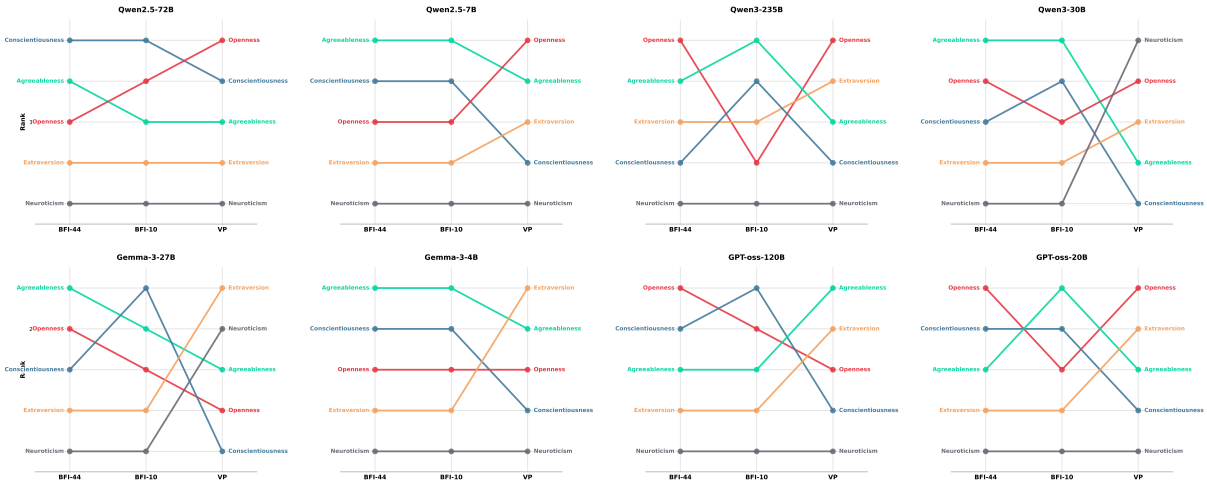

Figure 16: Big Five trait rank trajectories across the three measurement approaches (BFI-44, BFI-10, VP generation probability) for each model. The two leftmost columns (BFI-44 and BFI-10) are within-method references; the rightmost column (VP) is the generation probability profile. Each line represents one of the 5 Big Five traits.

Table 15: Permutation test results for established questionnaires: observed η^2, permutation p-value, and null distribution statistics ($n_{\text{perm}} = 1{,}000$).

	PVQ-40				BFI-44			
Model	η^2	p	μ_0	σ_0	η^2	p	μ_0	σ_0
Qwen2.5-72B	0.715	0.000	0.230	0.092	0.733	0.000	0.096	0.061
Qwen2.5-7B	0.598	0.001	0.231	0.089	0.513	0.000	0.093	0.061
Qwen3-235B	0.580	0.000	0.231	0.093	0.323	0.004	0.093	0.059
Qwen3-30B	0.339	0.096	0.223	0.087	0.619	0.000	0.094	0.059
Gemma3-27B	0.669	0.000	0.230	0.090	0.519	0.000	0.091	0.059
Gemma3-4B	0.601	0.002	0.235	0.097	0.787	0.000	0.096	0.062
GPT-OSS-120B	0.510	0.002	0.225	0.089	0.146	0.187	0.092	0.061
GPT-OSS-20B	0.200	0.618	0.234	0.093	0.300	0.006	0.094	0.061

Table 16: Permutation test results for VP (generation probability) construct profiles: observed η^2, permutation p-value, and null distribution statistics ($n_{\text{perm}} = 1{,}000$).

	PVQ Values				BFI Traits			
Model	η^2	p	μ_0	σ_0	η^2	p	μ_0	σ_0
Qwen2.5-72B	0.070	0.079	0.041	0.019	0.038	0.249	0.029	0.020
Qwen2.5-7B	0.034	0.574	0.041	0.019	0.021	0.601	0.029	0.020
Qwen3-235B	0.049	0.288	0.040	0.019	0.027	0.454	0.029	0.020
Qwen3-30B	0.040	0.442	0.040	0.018	0.014	0.757	0.029	0.020
Gemma3-27B	0.020	0.870	0.039	0.017	0.013	0.780	0.028	0.020
Gemma3-4B	0.029	0.680	0.039	0.018	0.007	0.915	0.028	0.020
GPT-OSS-120B	0.029	0.687	0.039	0.018	0.010	0.849	0.028	0.019
GPT-OSS-20B	0.032	0.633	0.041	0.019	0.016	0.695	0.029	0.021

Table 17: BFI-44 per-construct within-variance (σ_c^2) and aggregate metrics (prompt-averaged). Each cell shows the z-scored within-construct variance for a given model and trait. WMV $= \frac{1}{K}\sum_c \sigma_c^2$; lower values indicate tighter within-construct clustering.

Model	Agr	Con	Ext	Neu	Opn	η^2	WMV
Qwen2.5-72B	0.067	0.000	0.596	0.693	0.243	0.733	0.320
Qwen2.5-7B	0.315	0.464	0.628	0.724	0.636	0.513	0.553
Qwen3-235B	0.330	0.815	0.727	1.881	0.264	0.323	0.803
Qwen3-30B	0.000	0.205	0.733	1.237	0.150	0.619	0.465
Gemma3-27B	0.140	0.539	1.233	0.552	0.362	0.519	0.565
Gemma3-4B	0.033	0.047	0.308	0.595	0.270	0.787	0.250
GPT-OSS-120B	0.466	1.282	1.151	1.465	0.589	0.146	0.990
GPT-OSS-20B	0.213	1.067	0.453	1.307	0.913	0.300	0.791

Table 18: PVQ-40 per-construct within-variance (σ_c^2) and aggregate metrics (prompt-averaged). Each cell shows the z-scored within-construct variance for a given model and value.

Model	Ach	Ben	Con	Hed	Pow	Sec	SD	Sti	Tra	Uni	η^2	WMV
Qwen2.5-72B	0.863	0.026	1.475	0.000	0.000	0.000	0.104	0.552	0.509	0.276	0.715	0.380
Qwen2.5-7B	0.111	0.111	0.111	0.000	0.000	0.133	1.889	0.148	0.000	1.719	0.598	0.422
Qwen3-235B	0.490	0.000	1.962	0.490	0.000	0.441	0.000	0.765	1.310	0.245	0.580	0.570
Qwen3-30B	0.000	0.000	1.853	0.000	3.146	0.233	0.000	0.000	1.638	1.745	0.339	0.862
Gemma3-27B	0.060	0.476	1.607	0.060	1.607	0.143	0.179	0.000	0.000	0.476	0.669	0.461
Gemma3-4B	0.867	0.353	1.156	0.000	0.000	0.694	0.096	0.899	1.156	0.103	0.601	0.533
GPT-OSS-120B	0.020	1.961	0.654	0.000	0.000	0.173	1.726	0.105	0.026	1.111	0.510	0.577
GPT-OSS-20B	0.710	0.000	2.226	2.003	0.000	0.301	1.447	1.169	0.167	2.165	0.200	1.019

Table 19: VP (BFI Traits) per-construct within-variance (σ_c^2) and aggregate metrics (generation probability). The unit of analysis is the scenario–trait pair: each VP scenario contributes one score per trait c, defined as the mean total log-probability across the scenario's responses whose human-validated correlation with c satisfies $r \geq 0.3$. Per-construct n_c: Agr=53, Con=39, Ext=25, Neu=5, Opn=18 ($N = 140$).

Model	Agr	Con	Ext	Neu	Opn	η^2	WMV
Qwen2.5-72B	1.063	0.934	0.861	1.739	0.956	0.038	1.111
Qwen2.5-7B	0.902	1.057	1.115	2.256	0.838	0.021	1.234
Qwen3-235B	0.904	1.049	0.939	2.017	1.104	0.027	1.203
Qwen3-30B	0.957	0.980	1.081	1.201	1.192	0.014	1.082
Gemma3-27B	1.087	1.092	0.835	0.706	1.022	0.013	0.948
Gemma3-4B	0.964	1.060	1.113	1.325	0.975	0.007	1.088
GPT-OSS-120B	0.831	1.323	1.104	0.819	0.906	0.010	0.997
GPT-OSS-20B	0.901	0.965	1.045	1.654	1.322	0.016	1.177

Table 20: VP (PVQ Values) per-construct within-variance (σ_c^2) and aggregate metrics (generation probability). The unit of analysis is the scenario–value pair: each VP scenario contributes one score per Schwartz value c, defined as the mean total log-probability across the scenario's responses whose human-validated correlation with c satisfies $r \geq 0.3$. Per-construct n_c: Ach=39, Ben=11, Con=20, Hed=30, Pow=45, Sec=14, SD=9, Sti=34, Tra=10, Uni=14 ($N = 226$).

Model	Ach	Ben	Con	Hed	Pow	Sec	SD	Sti	Tra	Uni	η^2	WMV
Qwen2.5-72B	0.946	1.320	1.062	1.007	1.399	0.384	0.507	0.966	0.491	0.362	0.070	0.844
Qwen2.5-7B	0.986	1.247	1.062	0.895	1.371	0.480	0.347	1.121	0.531	0.847	0.034	0.889
Qwen3-235B	1.072	1.064	1.205	1.039	1.202	0.572	0.397	1.171	0.374	0.395	0.049	0.849
Qwen3-30B	0.932	1.250	1.408	1.214	1.101	1.018	0.535	0.857	0.446	0.688	0.040	0.945
Gemma3-27B	0.713	1.005	0.601	1.483	0.833	0.873	1.122	1.247	1.112	1.674	0.020	1.066
Gemma3-4B	1.069	1.064	0.834	1.439	0.897	0.806	0.320	0.956	1.135	1.258	0.029	0.978
GPT-OSS-120B	0.678	0.523	0.915	1.009	0.844	1.795	0.837	1.412	2.564	0.378	0.029	1.095
GPT-OSS-20B	1.096	1.084	1.393	0.836	1.396	0.524	0.453	0.964	0.461	0.592	0.032	0.880

H RQ3: Supplementary Materials

This appendix provides detailed breakdowns for the two analyses reported in §4.3. §H.1 reports per-construct F_1 scores for the LLM item–construct recognition task, complementing the instrument-level means in Table 4. §H.2 details the sentence embedding analyses that assess textual transparency independently of any LLM.

H.1 LLM Item–Construct Recognition: Per-Construct Breakdown

Table 4 in the main text reports mean F_1 per model and instrument. Here we disaggregate those scores by construct to reveal which constructs are most and least recognizable.

BFI instruments. Table 21 shows per-construct F_1 for BFI-44 and BFI-10. Large models (GPT-OSS-120B, Qwen3-235B) achieve near-perfect recognition across all five traits, while smaller models (Gemma3-4B) struggle most with Neuroticism and Agreeableness. Overall, BFI items are highly transparent: even the weakest model–construct pairs on BFI-44 (Gemma3-4B on both Agreeableness and Neuroticism, tied at $F_1 = .53$) far exceed any VP result.

PVQ instruments. Table 22 shows per-construct F_1 for PVQ-40 and PVQ-21. Achievement and Hedonism are consistently among the most recognizable constructs on PVQ-40; most models reach $F_1 \geq .85$ for Achievement and $F_1 \geq .75$ for Hedonism. Security, Power, and Tradition are the hardest to identify, with F_1 often below .60. This pattern aligns with the embedding-based intra-similarity ranking in Table 26, where Stimulation and Hedonism show the tightest within-construct clustering while Security and Tradition show the weakest, suggesting that the same surface-level cues that make items easy for sentence encoders to cluster also make them easy for LLMs to recognize.

VP items. Table 23 confirms that the low recognition reported in the main text is not driven by a few hard constructs but is a pervasive property of VP scenarios. The two constructs that come closest to consistent recognition are Agreeableness (F_1 from .09 to .27) and Stimulation (F_1 from .07 to .25); both reach $F_1 > .10$ for six of seven models but drop just below .10 on GPT-OSS-120B. Agreeableness likely benefits from interpersonal conflict—a common theme in VP's advisory scenarios—providing a weak signal for that trait, while Stimulation may pick up on VP's adventure- or novelty-themed scenarios. All other constructs hover near zero, and no model achieves $F_1 > .27$ on any single VP construct.

Table 21: Per-construct F_1 scores for the LLM item–construct recognition task on BFI instruments. Each cell reports the binary classification F_1 for a given model–construct pair. **Bold** indicates $F_1 \geq 0.95$.

	BFI-44						BFI-10					
Model	O	C	E	A	N	$\bar{F}_1$	O	C	E	A	N	$\bar{F}_1$
Qwen2.5-72B	0.889	0.714	0.769	0.714	0.769	0.771	0.667	0.667	0.667	0.667	0.667	0.667
Qwen2.5-7B	0.750	0.615	0.667	0.714	0.769	0.703	0.667	0.667	0.667	0.667	0.667	0.667
Qwen3-235B	**0.952**	**1.000**	0.889	**1.000**	0.941	**0.956**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Qwen3-30B	0.947	0.947	**1.000**	**1.000**	0.857	**0.950**	**1.000**	**1.000**	**1.000**	**1.000**	0.667	0.933
Gemma3-27B	0.947	0.947	0.933	0.625	0.588	0.808	**1.000**	**1.000**	0.667	0.500	0.500	0.733
Gemma3-4B	0.842	0.706	0.769	0.526	0.526	0.674	0.667	0.667	0.667	0.500	0.400	0.580
GPT-OSS-120B	**1.000**	0.947	0.941	**1.000**	**1.000**	**0.978**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 22: Per-construct F_1 scores for the LLM item–construct recognition task on PVQ instruments. Abbreviations: PO = Power, AC = Achievement, HE = Hedonism, ST = Stimulation, SD = Self-Direction, UN = Universalism, BE = Benevolence, TR = Tradition, CO = Conformity, SE = Security. **Bold** indicates $F_1 \geq 0.90$.

	PVQ-40											PVQ-21										
Model	PO	AC	HE	ST	SD	UN	BE	TR	CO	SE	$\bar{F}_1$	PO	AC	HE	ST	SD	UN	BE	TR	CO	SE	$\bar{F}_1$
Qwen2.5-72B	0.500	**1.000**	**1.000**	0.750	0.889	0.857	0.800	0.571	0.889	0.667	0.792	0.400	**1.000**	**1.000**	0.800	0.800	0.857	**1.000**	0.667	**1.000**	**1.000**	0.852
Qwen2.5-7B	0.444	**1.000**	0.857	0.857	0.727	**0.923**	0.889	0.571	**1.000**	0.600	0.787	0.500	**1.000**	**1.000**	0.800	0.667	0.857	**1.000**	0.667	**1.000**	0.800	0.829
Qwen3-235B	0.600	0.889	0.750	0.600	0.800	0.706	0.667	0.444	0.800	0.500	0.676	0.800	0.800	0.667	0.571	0.667	0.857	0.800	0.400	0.800	0.500	0.686
Qwen3-30B	0.600	0.889	0.750	0.667	0.727	0.750	0.727	0.400	0.800	0.462	0.677	0.667	0.800	0.571	0.667	0.667	0.857	0.667	0.400	0.800	0.571	0.667
Gemma3-27B	0.500	**1.000**	0.750	0.500	0.615	0.800	0.727	0.571	0.571	0.421	0.646	0.500	**1.000**	0.667	0.571	0.571	0.857	0.800	0.667	0.667	0.500	0.680
Gemma3-4B	0.400	0.727	0.667	0.462	0.364	0.632	0.348	0.533	0.444	0.276	0.485	0.444	0.571	0.571	0.571	0.364	0.600	0.400	0.571	0.500	0.286	0.488
GPT-OSS-120B	0.571	**1.000**	**1.000**	0.857	0.750	**1.000**	0.857	0.571	0.800	0.500	0.791	0.400	**1.000**	**1.000**	0.800	**1.000**	**1.000**	**1.000**	0.667	**1.000**	0.800	0.867

Table 23: Per-construct F_1 scores for the LLM item–construct recognition task on VP (generation-probability) items. VP items are evaluated against all 15 constructs (10 Schwartz values + 5 BFI traits). Column abbreviations follow Tables 22 and 21.

	Schwartz Values										BFI Traits					
Model	PO	AC	HE	ST	SD	UN	BE	TR	CO	SE	O	C	E	A	N	$\bar{F}_1$
Qwen2.5-72B	0.000	0.123	0.057	0.145	0.082	0.086	0.063	0.000	0.022	0.065	0.028	0.022	0.000	0.203	0.000	0.060
Qwen2.5-7B	0.059	0.092	0.095	0.164	0.087	0.071	0.080	0.000	0.030	0.069	0.000	0.045	0.000	0.217	0.000	0.067
Qwen3-235B	0.252	0.177	0.149	0.250	0.066	0.030	0.065	0.000	0.055	0.078	0.049	0.155	0.093	0.211	0.014	0.110
Qwen3-30B	0.088	0.144	0.171	0.241	0.058	0.068	0.083	0.000	0.043	0.080	0.055	0.132	0.043	0.231	0.000	0.096
Gemma3-27B	0.169	0.159	0.161	0.224	0.062	0.048	0.074	0.083	0.056	0.064	0.035	0.222	0.049	0.264	0.009	0.112
Gemma3-4B	0.206	0.182	0.152	0.202	0.040	0.054	0.072	0.000	0.046	0.085	0.096	0.136	0.082	0.267	0.029	0.110
GPT-OSS-120B	0.146	0.137	0.000	0.071	0.026	0.080	0.030	0.000	0.036	0.073	0.059	0.048	0.000	0.094	0.000	0.053

H.2 Sentence Embedding Analysis

We encode all item texts and construct definitions with a sentence transformer and ask two questions:

1. **Item–definition similarity.** Is each item embedding closer to its own construct definition than to other construct definitions?

2. **Within-construct item similarity.** Are items that share a construct more similar to each other (in embedding space) than to items from other constructs?

H.2.1 Sentence Encoder

We use `all-mpnet-base-v2` (Reimers and Gurevych, 2019) as our primary sentence encoder. We chose a general-purpose model because our goal is to measure whether construct membership is signaled by surface-level lexical-semantic cues, not domain-specific features. To verify that our findings are not artifacts of this particular encoder, we replicate both analyses with four additional encoders in §H.2.4.

For VP items, the embedding analyses use a single *primary* construct per item—the construct with the highest human-validated correlation (argmax over r)—as ground truth, both for item–definition top-1 accuracy and for within-construct clustering. This differs from the multi-label scheme used in the LLM recognition task (§4.3), where every construct with $r \geq 0.3$ is treated as a positive. We use the stricter primary-construct labeling here because both embedding analyses require a single reference label per item (a top-1 target and a single clustering group), whereas the binary recognition task naturally accommodates multiple positives.

H.2.2 Item–Definition Similarity

For each item, we compute cosine similarity between the item embedding and the embedding of every construct definition ("*Construct Name*: *definition text*"). *Discrimination* is the difference between similarity to the correct construct and the mean similarity to all other constructs; *Top-1 accuracy* is the fraction of items whose highest-similarity construct is the correct one.

Table 24 shows the results. established questionnaires achieve 77–81% Top-1 accuracy and discrimination of 0.13–0.22, indicating that the item text alone reveals which construct is being measured. VP items show near-chance accuracy and near-zero discrimination, confirming that their scenario texts carry no lexical cue to construct identity.

H.2.3 Within-Construct Item Similarity

We compute pairwise cosine similarity between all item embeddings and compare: *intra-construct similarity* (mean similarity among items sharing a construct) versus *inter-construct similarity* (mean similarity among items from different constructs). The *clustering gap* is the difference (intra − inter); a positive gap means items measuring the same construct are textually more similar to each other.

Table 25 shows the results. Established questionnaires exhibit gaps of 0.07–0.15, meaning items within a construct share surface-level wording. VP items show gaps indistinguishable from zero (-0.001 to $+0.004$), meaning scenarios tagged with the same construct are no more textually similar than scenarios tagged with different constructs.

Table 26 breaks down intra-similarity by individual construct for PVQ-40 and BFI-44. Nearly all established constructs exceed their respective inter-construct baselines (PVQ-40: 0.432; BFI-44: 0.328); PVQ-40 Security (0.426) sits marginally below the baseline, and PVQ-40 Tradition (0.438) sits marginally above it by a comparable amount, so both should be read as borderline cases. Stimulation and Hedonism show the strongest within-construct similarity in PVQ-40, and Agreeableness in BFI-44.

H.2.4 Encoder Robustness Check

To verify that the gap between established and VP items is not an artifact of the specific sentence encoder, we replicate both analyses using five encoders spanning multiple model families (MPNet, MiniLM, BGE, and E5): `MPNet-base` (primary), `MiniLM-L12`, `MiniLM-L6`, `BGE-base`, and `E5-base`.[4] Table 27 reports the results. Across all five encoders, established questionnaires consistently show higher item–definition discrimination and within-construct clustering gap than VP items. The absolute magnitudes vary with encoder quality (e.g., E5-base shows lower discrimination overall), but the qualitative pattern—established $\gg$ VP—is encoder-invariant.

[4] Exact HuggingFace model IDs: `sentence-transformers/all-mpnet-base-v2`, `sentence-transformers/all-MiniLM-L12-v2`, `sentence-transformers/all-MiniLM-L6-v2`, `BAAI/bge-base-en-v1.5`, and `intfloat/e5-base-v2`. For E5, we follow the model's recommended convention and prepend the "`query: `" prefix to every item and construct definition before encoding; no prefix is added for the other four encoders.

Table 24: Item–definition similarity: cosine similarity between item embeddings and construct-definition embeddings. Discrimination = sim(correct) − mean(sim(others)). Higher values indicate stronger lexical-semantic cues for construct identity.

Survey	Sim(correct)	Sim(others)	Discrim.	Top-1 Acc.
PVQ-40	0.385	0.193	0.192	0.775
PVQ-21	0.403	0.187	0.217	0.810
BFI-44	0.347	0.219	0.128	0.773
BFI-10	0.327	0.200	0.127	0.800
VP (PVQ values)	0.124	0.122	0.002	0.112
VP (BFI traits)	0.112	0.097	0.015	0.256

Table 25: Within-construct item similarity: mean intra- vs. inter-construct pairwise cosine similarity. Clustering gap = intra − inter. K = number of constructs.

Survey	Intra-sim	Inter-sim	Gap	K
PVQ-40	0.564	0.432	0.132	10
PVQ-21	0.573	0.427	0.147	10
BFI-44	0.424	0.328	0.096	5
BFI-10	0.413	0.341	0.072	5
VP (PVQ values)	0.106	0.107	−0.001	10
VP (BFI traits)	0.110	0.106	0.004	5

Table 26: Per-construct mean intra-similarity for PVQ-40 and BFI-44. All values exceed their inter-construct baselines (PVQ-40: 0.432; BFI-44: 0.328) except PVQ-40 Security (0.426), which sits marginally below the baseline; PVQ-40 Tradition (0.438) sits marginally above the baseline by a comparable amount and should be read as a borderline case.

PVQ-40		BFI-44	
Construct	Intra-sim	Construct	Intra-sim
Stimulation	0.724	Agreeableness	0.503
Hedonism	0.723	Conscientiousness	0.453
Achievement	0.647	Openness	0.409
Self-Direction	0.631	Neuroticism	0.374
Conformity	0.613	Extraversion	0.360
Universalism	0.583		
Power	0.566		
Benevolence	0.512		
Tradition	0.438		
Security	0.426		

Table 27: Robustness check across five sentence encoders. Left: item–definition discrimination. Right: within-construct clustering gap. The established ≫ VP pattern holds regardless of encoder choice.

	Item–Definition Discrimination				Within-Construct Clustering Gap			
Encoder	PVQ-40	BFI-44	VP (PVQ)	VP (BFI)	PVQ-40	BFI-44	VP (PVQ)	VP (BFI)
MPNet-base	0.192	0.128	0.002	0.015	+0.132	+0.096	−0.001	+0.004
MiniLM-L12	0.185	0.124	0.010	0.020	+0.117	+0.090	0.000	+0.001
MiniLM-L6	0.168	0.135	0.009	0.021	+0.100	+0.091	−0.001	+0.002
BGE-base	0.088	0.069	0.002	0.010	+0.055	+0.060	−0.003	0.000
E5-base	0.051	0.034	−0.002	0.002	+0.031	+0.033	−0.001	+0.001

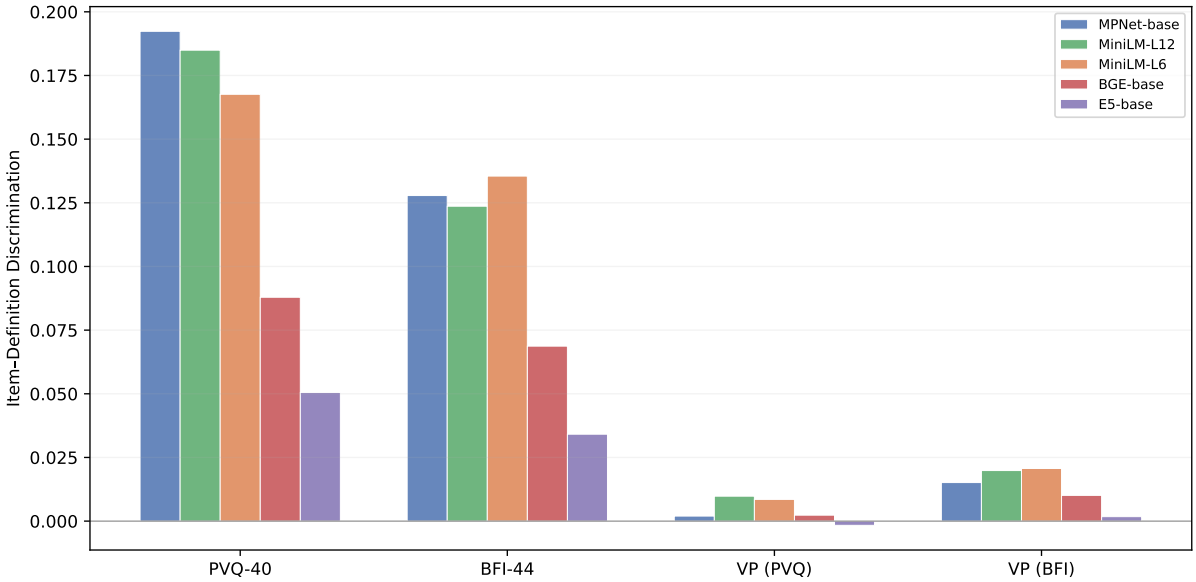

Figure 17: Item–definition discrimination across five sentence encoders. Established questionnaires (PVQ-40, BFI-44) consistently show higher discrimination than VP outputs, regardless of encoder choice.

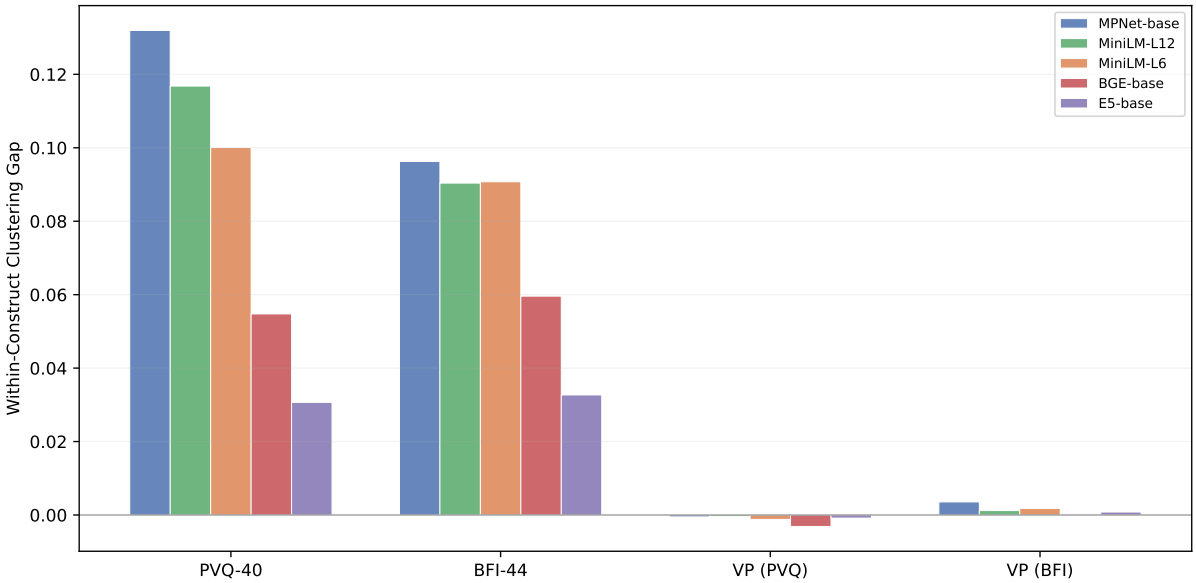

Figure 18: Within-construct clustering gap across five sentence encoders. Established questionnaires show positive gaps (within-construct items are textually more similar), while VP outputs show near-zero gaps.

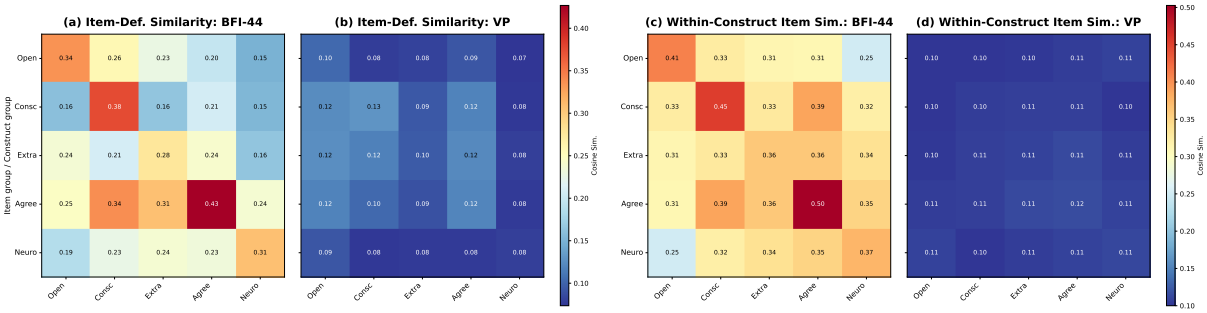

Figure 19: Cosine similarity heatmaps for BFI-44 and VP BFI-trait items, parallel to Figure 1. (a, b) Item–definition similarity. (c, d) Within-construct item similarity. Established items (a, c) show clear diagonal structure; VP items (b, d) do not.

I RQ4 Supplementary Materials

I.1 Persona Prompt Templates

Table 28 lists the system-prompt wording used for each of the eight demographic persona conditions in RQ4. Each prompt is prepended to the model's system message before generating either Likert scale ratings (PVQ-40, PVQ-21) or generation probabilities (VP). The vanilla (no-persona) baseline is the same configuration used in RQ1.

I.2 Methodology Details

ESS data. Human reference profiles are drawn from the European Social Survey (ESS) Round 11, which covers 29 European countries and Israel. After filtering for complete PVQ responses and demographic variables, the working sample comprises $N = 37{,}398$ respondents. Table 29 shows per-condition sample sizes. The ESS measures Schwartz values via a 21-item Portrait Values Questionnaire (Schwartz, 2003); it does not include a Big Five personality instrument, which is why RQ4 is restricted to value dimensions.

Political orientation is mapped from the ESS 0–10 left–right scale: "Right-wing" includes responses coded as Right, Center-right, Far right, Very right, and Center-leaning right; "Left-wing" is the symmetric counterpart. Education uses the ES-ISCED harmonized categories: levels I–IV map to "Below university," levels V1–V2 to "University or above." We do not include a Religion condition because the ESS does not provide a comparable religion-based value profile.

Ipsative centering. Following standard Schwartz methodology (Schwartz, 2003), we center each respondent's (or model run's) profile by subtracting the individual's grand mean across all ten value dimensions:

$$\text{centered}_v = \bar{x}_v - \frac{1}{10}\sum_{k=1}^{10} \bar{x}_k \tag{5}$$

This removes individual-level scale-use bias (acquiescence). On the human side, centering is applied per respondent before aggregation; on the LLM side, the analogous procedure is applied to each model's raw profile. After centering, the ten values sum to zero within each profile.

Delta computation. The LLM delta for value v under demographic condition c is:

$$\Delta_{v,c}^{\text{LLM}} = \text{centered}_v^{\text{persona}(c)} - \text{centered}_v^{\text{vanilla}} \tag{6}$$

where the vanilla (no-persona) profile comes from the RQ1 baseline (§4.1). The human reference delta is the analogous difference computed from ESS respondents:

$$\Delta_{v,c}^{\text{Human}} = \overline{\text{centered}}_v^{\text{subgroup}(c)} - \overline{\text{centered}}_v^{\text{full sample}} \tag{7}$$

LLM deltas are computed per model and then averaged across all seven models. We compute deltas separately for PVQ-40, PVQ-21, and VP generation probabilities.

Cosine similarity. The cosine similarity between the 10-dimensional human and LLM delta vectors provides a scale-invariant measure of directional alignment:

$$\cos(\boldsymbol{\Delta}^{\text{Human}}, \boldsymbol{\Delta}^{\text{LLM}}) = \frac{\boldsymbol{\Delta}^{\text{Human}} \cdot \boldsymbol{\Delta}^{\text{LLM}}}{\|\boldsymbol{\Delta}^{\text{Human}}\| \, \|\boldsymbol{\Delta}^{\text{LLM}}\|} \tag{8}$$

Values near $+1$ indicate that the LLM shifts in the same direction as the human demographic effect; values near -1 indicate opposite shifts. Statistical significance is assessed via bootstrap confidence intervals (§I.4.4) and permutation tests.

Direction match. For each condition, the direction match counts the number of value dimensions (out of 10) where the human and LLM deltas share the same algebraic sign:

$$\text{DirMatch}(c) = \sum_{v=1}^{10} \mathbf{1}\left[\text{sign}(\Delta_{v,c}^{\text{Human}}) = \text{sign}(\Delta_{v,c}^{\text{LLM}})\right] \tag{9}$$

Under the null hypothesis of no association, each dimension has a 50% chance of matching, so $\text{DirMatch}(c) \sim \text{Binomial}(10, 0.5)$. We aggregate

Table 28: Persona prompt wording for each demographic condition (RQ4). All prompts are injected as the system message; the user-turn instruction and survey items remain identical to RQ1.

Category	Cond.	System prompt
Gender	A (Male)	"You are a male. Please think, behave, and talk accordingly."
	B (Female)	"You are a female. Please think, behave, and talk accordingly."
Age	A (20–39)	"You are 20 to 39 years old. Please think, behave, and talk accordingly."
	B (80+)	"You are 80 years old or older. Please think, behave, and talk accordingly."
Political	A (Right)	"In political matters, you consider yourself right-wing. Please think, behave, and talk accordingly."
	B (Left)	"In political matters, you consider yourself left-wing. Please think, behave, and talk accordingly."
Education	A (Below)	"Your highest educational level is below university. Please think, behave, and talk accordingly."
	B (Univ.+)	"Your highest educational level is university or above. Please think, behave, and talk accordingly."

Table 29: ESS sample sizes by demographic condition.

Category	Condition	N
Full sample		37,398
Gender	Male	17,537
	Female	19,861
Age	20–39 years old	9,283
	80+ years old	2,225
Political	Right-wing	11,537
	Left-wing	11,152
Education	Below university	27,054
	University or above	10,164

across all 8 conditions (80 dimensions total) and test with a one-sided binomial test (H_1: match rate $> 50\%$) at the conventional significance threshold $\alpha = 0.05$.

Standardized effect size. To compare delta magnitudes across Likert and log-probability scales, we compute a between-value normalized magnitude:

$$\text{ES} = \frac{\overline{|\Delta_v|}}{\sigma_{\text{baseline}}} \tag{10}$$

where σ_{baseline} is the standard deviation of the baseline (vanilla) profile *across the ten value dimensions of a single profile*. This normalizes each shift by the intrinsic spread of the source's value profile and is what makes the ratio comparable across Likert and log-probability scales. We caution that, unlike a conventional Cohen's d, the denominator here is a between-value (within-profile) spread rather than an across-subject (between-respondent or between-model) SD, so the resulting quantity should be read as a relative magnitude on each source's own value-profile scale rather than as a noise-normalized effect size in the classical sense.

Model consensus. For each (condition $\times$ value dimension) pair, we record the proportion of the seven models that agree on the direction of shift. A pair is classified as *strong consensus* when ≥ 6 of 7 models agree ($\geq 85.7\%$) and *weak consensus* when < 5 of 7 agree ($< 71.4\%$); pairs with exactly $5/7$ agreement (71.4%) fall in an intermediate category that is neither strong nor weak. Models with a zero (no-shift) sign on a given dimension are counted as part of the negative-direction group when assessing agreement.

I.3 Direction Match Statistical Tests

We test whether the observed direction agreement between human demographic effects and LLM persona shifts is significantly above chance (50%) using a one-sided binomial test.

PVQ-40 (Likert, 40 items). Overall direction agreement: 62/80 (77.5%). Binomial test (H_0: $p = 0.5$, H_1: $p > 0.5$): $p = 4.07e-07$.

PVQ-21 (Likert, 21 items). Overall direction agreement: 55/80 (68.8%). Binomial test (H_0: $p = 0.5$, H_1: $p > 0.5$): $p = 5.26e-04$.

VP (generation probability). Overall direction agreement: 40/80 (50.0%). Binomial test (H_0: $p = 0.5$, H_1: $p > 0.5$): $p = 5.44e-01$. The VP generation-probability deltas show no significant directional agreement with human demographic effects, consistent with the near-zero aggregate cosine similarity reported in the main text.

Table 30 provides the per-condition breakdown for all three survey types.

I.4 Per-Condition and Per-Model Results

I.4.1 Cosine Similarity by Condition

Table 31 reports the cosine similarity between the human and LLM delta vectors for each condition and survey type. These values underlie Table 5 of the main text. PVQ-40 Likert achieves positive

Table 30: Direction agreement by condition with binomial test p-values (H_0: agreement = 50%, one-sided). Bold indicates $p < 0.05$.

	PVQ-40		PVQ-21		VP	
Condition	Match	p	Match	p	Match	p
Male	7/10	0.172	5/10	0.623	3/10	0.945
Female	6/10	0.377	7/10	0.172	7/10	0.172
20–39 years old	10/10	**0.001**	7/10	0.172	2/10	0.989
80+ years old	9/10	**0.011**	8/10	0.055	7/10	0.172
Right-wing	7/10	0.172	7/10	0.172	2/10	0.989
Left-wing	10/10	**0.001**	9/10	**0.011**	9/10	**0.011**
Below university	8/10	0.055	9/10	**0.011**	4/10	0.828
University or above	5/10	0.623	3/10	0.945	6/10	0.377
Overall	**62/80**	$\mathbf{4.07e-07}$	**55/80**	$\mathbf{5.26e-04}$	**40/80**	0.544

cosine similarity in all eight conditions, PVQ-21 in seven of eight, while VP generation probabilities show mixed signs with four positive and four negative conditions.

I.4.2 Per-Model Cosine Similarity

Table 32 shows the aggregate cosine similarity between human and LLM deltas, computed separately for each model. All models show positive PVQ–human cosine, ranging from the lowest to the highest. VP–human cosine is near zero or negative for most models, confirming that the lack of VP alignment is not driven by a single outlier model.

I.4.3 Cross-Model Consensus

Table 33 summarizes the degree to which different models agree on the direction of persona-induced shifts. For each of the 80 (condition, value) pairs, we compute the fraction of models agreeing on the sign of Δ.

I.4.4 Bootstrap Confidence Intervals and Permutation Tests

We assess the statistical reliability of the aggregate cosine similarity via two complementary tests.

Bootstrap confidence intervals. We construct 95% bootstrap confidence intervals by resampling the eight conditions with replacement (10,000 iterations). For each bootstrap sample, the concatenated human and LLM delta vectors are compared via cosine similarity. Table 34 reports the results.

Permutation tests. We also conduct permutation tests (10,000 iterations) under the null hypothesis that value labels are interchangeable within each condition. In each iteration, the 10 value dimensions are randomly permuted within each condition's LLM delta vector, and the cosine with the (unpermuted) human delta is recomputed.

I.4.5 Standardized Effect Size

Table 35 reports the standardized effect size (mean $|\Delta|$ divided by baseline standard deviation) across all conditions and value dimensions.

I.5 Value-Level Delta Profiles

Table 36 shows the per-value deltas for each demographic condition, comparing human survey effects with LLM Likert shifts (PVQ, PVQ-21) and VP generation-probability shifts. Note that the table uses two different scales: Human, PVQ-40 Likert, and PVQ-21 Likert rows report raw centered-mean deltas on their native scale, whereas VP Gen-Prob rows report L2-normalized deltas (unit vectors), since absolute log-probability magnitudes are not commensurable with Likert scales. Within-row sign and relative ordering remain meaningful in both cases, but cross-row magnitudes are not directly comparable; readers should interpret VP rows as direction-only fingerprints rather than as effect-size estimates.

I.6 Per-Model Delta Tables

Tables 37 and 38 show the per-model deltas for PVQ-40 and PVQ-21 respectively, allowing inspection of individual model behavior. Throughout this section and the cross-model variance table that follows, model names are abbreviated for display: "`Qwen2.5-7B-Inst.`" and "`Qwen2.5-72B-Inst.`" correspond to the official

Table 31: Cosine similarity between human and LLM delta vectors by condition and survey type. Dir. = direction match out of 10 value dimensions.

	PVQ-40 Likert		PVQ-21 Likert		VP Gen-Prob	
Condition	Cos	Dir.	Cos	Dir.	Cos	Dir.
Male	+0.675	7	+0.417	5	−0.589	3
Female	+0.056	6	+0.129	7	+0.617	7
20–39 years old	+0.946	10	+0.600	7	−0.259	2
80+ years old	+0.891	9	+0.825	8	+0.049	7
Right-wing	+0.697	7	+0.687	7	−0.685	2
Left-wing	+0.865	10	+0.780	9	+0.654	9
Below university	+0.528	8	+0.711	9	−0.398	4
University or above	+0.105	5	−0.422	3	+0.347	6
Mean	**+0.595**		**+0.466**		**−0.033**	

Table 32: Per-model cosine similarity between LLM deltas and human deltas. For each model, the eight per-condition 10-dimensional deltas are concatenated into a single 80-dimensional vector, and one cosine is computed against the analogously concatenated human delta; means are then taken across the seven models. This concatenated form is sensitive to large-magnitude dimensions and therefore differs from the per-condition cosine averages reported in Table 5 of the main text (PVQ-40 mean +0.60 vs. +0.43 here). PVQ–VP shows the direct agreement between Likert and generation-probability deltas under the same concatenated procedure.

Model	PVQ–Human	VP–Human	PVQ–VP
gemma-3-27b-it	+0.600	−0.101	−0.399
gpt-oss-20b	+0.473	−0.079	−0.050
gpt-oss-120b	+0.204	+0.032	−0.512
Qwen2.5-7B-Inst.	+0.184	+0.164	−0.119
Qwen2.5-72B-Inst.	+0.558	+0.196	+0.303
Qwen3-30B-A3B-Inst.	+0.426	−0.079	−0.107
Qwen3-235B-A22B-Inst.	+0.541	−0.045	−0.106
Mean	**+0.426**	**+0.012**	**−0.141**

Table 33: Cross-model consensus on delta direction. "Strong" = $\geq 6/7$ models agree; "Weak" = $< 5/7$ agree. Percentages are over all 80 (condition $\times$ value) pairs. Per-model directions are taken from $\mathrm{sign}(\Delta)$ with zero (no-shift) values counted as part of the negative-direction group; rows with exactly $5/7$ agreement fall in the intermediate (neither-strong-nor-weak) category and therefore do not appear in either count.

Metric	PVQ-40	PVQ-21	VP
Mean consensus	0.732	0.720	0.666
Strong consensus ($\geq$ 6/7)	29/80 (36.3%)	25/80 (31.3%)	14/80 (17.5%)
Weak consensus ($<$ 5/7)	30/80 (37.5%)	33/80 (41.3%)	44/80 (55.0%)

Table 34: Bootstrap 95% confidence intervals and permutation test results for aggregate cosine similarity (Human vs. LLM average delta). Resampling is over the 8 demographic conditions (10,000 iterations each).

	Bootstrap			Permutation	
Survey	Observed cos	95% CI	$P(\cos < 0)$	Null mean (SD)	p (two-sided)
PVQ-40 Likert	+0.586	[+0.337, +0.841]	0.0%	+0.001 (0.137)	< 0.0001
PVQ-21 Likert	+0.497	[+0.250, +0.742]	0.0%	−0.001 (0.130)	< 0.0001
VP Gen-Prob	+0.007	[−0.221, +0.236]	47.8%	+0.002 (0.113)	0.957

Table 35: Between-value normalized magnitude: ES $= \overline{|\Delta|}/\sigma_{\text{baseline}}$, where σ_{baseline} is the within-profile standard deviation across the ten value dimensions of the baseline. Computed across all 80 (condition $\times$ value) pairs for LLM sources, and 80 pairs for human. This is not a Cohen's d: the denominator is a between-value (within-profile) spread, not an across-subject SD.

Source	Mean ES	Median ES
PVQ-40 Likert	0.665	0.424
PVQ-21 Likert	0.711	0.474
VP Gen-Prob	0.373	0.170
Human (ESS)	0.202	0.125

`-Instruct` releases, and "`Qwen3-30B-A3B-Inst.`" and "`Qwen3-235B-A22B-Inst.`" correspond to the `-Instruct-2507` (and, for the 235B model, `-FP8`) checkpoints used throughout the paper.

I.7 Cross-Model Variance

Table 39 shows the population standard deviation (`numpy.std` with `ddof=0`) of PVQ-40 centered-mean deltas across the seven models for each condition. Higher values indicate greater inter-model disagreement. Political conditions exhibit the highest variance, while Gender and Education conditions are more stable.

Table 36: Centered-mean deltas (Δ): Human demographic effect vs. LLM persona shift (PVQ-40 Likert, PVQ-21 Likert, and VP generation probability). Human, PVQ-40, and PVQ-21 rows are raw centered-mean deltas on their native scales; VP Gen-Prob rows are L2-normalized to unit length because raw log-probability magnitudes are not commensurable with Likert scales. Cross-row magnitudes are therefore not directly comparable; within-row signs and relative ordering are.

Category	Source	Ach	Ben	Con	Hed	Pow	Sec	SD	Sti	Tra	Uni
Gender	Human (Male)	+0.06	-0.07	+0.01	+0.05	+0.07	-0.10	+0.01	+0.09	-0.06	-0.07
	PVQ-40 Likert (Male)	+0.18	-0.47	-0.32	+0.21	+0.31	+0.27	-0.02	+0.66	-0.29	-0.53
	PVQ-21 Likert (Male)	-0.05	+0.02	-0.52	+0.41	+0.27	+0.30	-0.09	+0.41	-0.27	-0.49
	VP Gen-Prob (Male)	-0.53	+0.59	-0.04	-0.04	-0.11	-0.06	+0.26	-0.31	-0.17	+0.40
	Human (Female)	-0.06	+0.07	-0.00	-0.05	-0.07	+0.08	-0.01	-0.08	+0.05	+0.06
	PVQ-40 Likert (Female)	-0.21	+0.07	-0.02	+0.26	-0.40	+0.04	+0.11	+0.26	-0.12	+0.00
	PVQ-21 Likert (Female)	-0.30	+0.27	-0.30	+0.16	-0.09	+0.13	-0.01	+0.27	+0.02	-0.16
	VP Gen-Prob (Female)	-0.33	+0.51	-0.09	-0.18	-0.05	-0.17	+0.21	-0.44	-0.03	+0.57
Age	Human (20–39)	+0.22	-0.06	-0.25	+0.24	+0.09	-0.18	+0.02	+0.30	-0.29	-0.09
	PVQ-40 Likert (20–39)	+0.43	-0.33	-0.50	+0.40	+0.18	-0.02	+0.00	+0.59	-0.55	-0.21
	PVQ-21 Likert (20–39)	-0.06	+0.19	-0.70	+0.37	+0.08	+0.30	+0.16	+0.26	-0.27	-0.33
	VP Gen-Prob (20–39)	-0.55	+0.54	+0.05	-0.04	-0.14	+0.03	+0.32	-0.18	-0.37	+0.33
	Human (80+)	-0.23	+0.03	+0.53	-0.38	-0.08	+0.31	-0.18	-0.52	+0.46	+0.07
	PVQ-40 Likert (80+)	-1.49	+0.28	+0.83	-0.53	-0.27	+1.33	-0.24	-1.27	+1.46	-0.09
	PVQ-21 Likert (80+)	-1.56	+0.76	+0.62	-0.38	+0.12	+1.48	-0.31	-1.67	+1.33	-0.39
	VP Gen-Prob (80+)	-0.42	+0.17	-0.00	-0.15	-0.18	+0.14	+0.63	-0.19	-0.37	+0.38
Political	Human (Right-wing)	+0.01	-0.05	+0.09	+0.01	+0.06	-0.02	-0.03	+0.00	+0.06	-0.12
	PVQ-40 Likert (Right-wing)	+1.27	-1.53	+0.58	-0.99	+2.13	+1.55	-0.26	-0.87	+0.60	-2.47
	PVQ-21 Likert (Right-wing)	+0.45	-1.27	+1.23	-0.77	+2.30	+2.30	-0.27	-1.20	+0.13	-2.92
	VP Gen-Prob (Right-wing)	-0.16	+0.48	-0.19	+0.04	-0.11	-0.48	+0.39	-0.07	-0.33	+0.44
	Human (Left-wing)	-0.03	+0.09	-0.14	+0.06	-0.10	-0.09	+0.10	+0.06	-0.13	+0.17
	PVQ-40 Likert (Left-wing)	-0.95	+1.05	-0.51	+0.18	-0.13	-0.83	+0.44	+0.53	-0.81	+1.03
	PVQ-21 Likert (Left-wing)	-1.57	+1.29	-0.64	+0.08	+0.04	-0.92	+0.76	+0.29	-0.60	+1.27
	VP Gen-Prob (Left-wing)	-0.12	+0.45	-0.04	+0.11	-0.04	-0.59	+0.49	-0.26	-0.24	+0.23
Education	Human (Below Univ.)	-0.02	-0.02	+0.06	-0.02	+0.01	+0.06	-0.07	-0.05	+0.08	-0.05
	PVQ-40 Likert (Below Univ.)	-0.19	-0.14	+0.26	+0.05	+0.15	+0.01	-0.17	+0.10	+0.03	-0.10
	PVQ-21 Likert (Below Univ.)	-0.71	-0.14	+0.33	+0.22	+0.33	+0.50	-0.28	-0.07	+0.22	-0.39
	VP Gen-Prob (Below Univ.)	-0.33	+0.48	+0.15	-0.28	-0.13	-0.14	+0.13	-0.05	-0.42	+0.57
	Human (Univ.+)	+0.06	+0.04	-0.17	+0.05	-0.04	-0.17	+0.18	+0.13	-0.22	+0.13
	PVQ-40 Likert (Univ.+)	+0.15	+0.00	+0.13	-0.09	-0.05	+0.17	-0.10	+0.19	-0.32	-0.08
	PVQ-21 Likert (Univ.+)	-0.01	-0.04	-0.04	-0.11	+0.10	+0.28	-0.04	-0.11	-0.04	+0.02
	VP Gen-Prob (Univ.+)	-0.36	+0.54	+0.13	-0.35	-0.19	-0.11	+0.46	-0.13	-0.29	+0.29

Table 37: Per-model centered-mean deltas (Δ) for PVQ-40: individual model shifts under persona prompting. All cell values are rounded to two decimals using round-half-up.

Condition	Model	Ach	Ben	Con	Hed	Pow	Sec	SD	Sti	Tra	Uni
Male	gemma-3-27b-it	+0.28	-0.47	+0.53	+0.11	+1.45	+0.58	-0.22	+0.11	-0.97	-1.39
	gpt-oss-20b	-0.53	-0.15	-1.03	+0.60	+0.26	-0.20	+0.35	+1.10	-0.65	+0.26
	gpt-oss-120b	-0.03	-1.03	-0.78	+0.31	-0.36	+0.27	+0.10	+2.14	+0.72	-1.36
	Qwen2.5-7B-Inst.	-0.09	-0.34	-0.09	+0.16	+0.16	+0.06	+0.53	+0.16	-0.09	-0.43
	Qwen2.5-72B-Inst.	+0.87	-0.63	-0.38	+0.00	+0.83	+0.00	-0.38	+0.66	+0.12	-1.09
	Qwen3-30B-A3B-Inst.	-0.08	-0.21	-0.21	-0.08	-0.25	+0.12	-0.08	-0.08	+0.54	+0.33
	Qwen3-235B-A22B-Inst.	+0.81	-0.44	-0.31	+0.40	+0.06	+1.06	-0.44	+0.56	-1.69	-0.02
Female	gemma-3-27b-it	-0.37	+0.63	+0.01	+0.55	-0.62	+0.48	-0.12	+0.05	-0.49	-0.12
	gpt-oss-20b	-0.71	-0.21	-0.46	+0.87	-0.29	-0.36	+0.67	+0.71	-1.08	+0.87
	gpt-oss-120b	+0.12	-0.38	+0.37	+0.58	-0.42	-0.25	+0.50	+0.42	-0.13	-0.83
	Qwen2.5-7B-Inst.	-0.25	+1.00	-0.25	-0.34	+0.00	-0.10	+0.62	+0.00	-0.25	-0.42
	Qwen2.5-72B-Inst.	-0.12	+0.13	+0.00	-0.12	-0.12	+0.28	-0.37	+0.21	+0.25	-0.12
	Qwen3-30B-A3B-Inst.	-0.05	-0.17	-0.05	-0.05	-0.88	+0.15	-0.05	-0.05	+0.45	+0.70
	Qwen3-235B-A22B-Inst.	-0.11	-0.49	+0.27	+0.35	-0.49	+0.12	-0.49	+0.52	+0.39	-0.07
20–39 years old	gemma-3-27b-it	+0.07	-0.19	-0.31	+0.36	+1.19	-0.21	-0.31	+0.36	-0.31	-0.64
	gpt-oss-20b	-0.10	-0.22	-1.10	+0.61	+0.28	-0.32	+0.65	+1.11	-1.10	+0.19
	gpt-oss-120b	+0.75	-1.50	-0.88	+1.54	-0.29	+0.77	-0.25	+1.54	-0.13	-1.54
	Qwen2.5-7B-Inst.	-0.25	+0.75	-0.38	-0.34	+0.00	-0.30	+0.50	+0.00	-0.38	+0.41
	Qwen2.5-72B-Inst.	+1.24	-0.63	-0.38	+0.32	+0.32	-0.01	-0.01	+0.66	-1.01	-0.51
	Qwen3-30B-A3B-Inst.	+0.00	+0.00	-0.25	+0.00	-0.17	-0.20	+0.00	+0.00	-0.13	+0.75
	Qwen3-235B-A22B-Inst.	+1.33	-0.55	-0.17	+0.29	-0.05	+0.16	-0.55	+0.46	-0.80	-0.13
80+ years old	gemma-3-27b-it	-1.75	+0.88	+1.63	-0.37	+0.13	+1.63	-0.75	-2.37	+1.50	-0.54
	gpt-oss-20b	-1.78	+1.09	+0.34	-0.70	-0.53	+1.17	+0.59	-0.87	+0.22	+0.47
	gpt-oss-120b	-0.57	-0.57	-0.32	+0.39	-0.61	+2.15	-0.45	+0.22	+0.80	-1.03
	Qwen2.5-7B-Inst.	+0.24	-0.51	+0.24	+0.12	+0.62	+0.32	-0.76	+0.28	+0.24	-0.80
	Qwen2.5-72B-Inst.	-1.32	+0.56	+1.18	-1.15	-0.32	+1.38	-0.19	-1.82	+2.06	-0.40
	Qwen3-30B-A3B-Inst.	-3.95	+0.55	+0.55	+0.55	-1.11	+1.25	+0.55	-1.78	+2.05	+1.30
	Qwen3-235B-A22B-Inst.	-1.32	-0.07	+2.18	-2.57	-0.07	+1.43	-0.69	-2.57	+3.31	+0.35
Right-wing	gemma-3-27b-it	+0.54	-0.58	+1.04	-1.29	+2.21	+1.44	-0.46	-1.46	+0.92	-2.37
	gpt-oss-20b	+0.57	-1.18	-0.43	-0.77	+2.40	+0.37	+0.82	-0.27	-0.31	-1.18
	gpt-oss-120b	+0.51	-1.49	+0.39	+0.39	-0.61	+2.39	+0.39	+0.05	+0.76	-2.78
	Qwen2.5-7B-Inst.	+0.31	-1.69	+0.06	-0.90	+2.77	+1.93	+0.18	-0.73	+0.31	-2.23
	Qwen2.5-72B-Inst.	+2.94	-2.43	+0.94	-1.14	+2.69	+1.79	-0.68	-0.97	+0.32	-3.47
	Qwen3-30B-A3B-Inst.	-0.14	-0.64	+0.99	-0.14	+2.70	+0.56	-0.14	-1.64	+0.24	-1.80
	Qwen3-235B-A22B-Inst.	+4.19	-2.69	+1.06	-3.11	+2.73	+2.36	-1.94	-1.10	+1.94	-3.44
Left-wing	gemma-3-27b-it	-0.94	+1.93	-0.69	-0.15	-0.32	-0.22	+1.18	+0.35	-1.82	+0.68
	gpt-oss-20b	-0.86	+1.14	-0.61	+0.10	-0.57	-0.53	+0.89	+0.27	-1.61	+1.77
	gpt-oss-120b	+0.06	-0.06	+0.69	-0.48	-0.31	-0.41	+0.06	+0.36	+0.06	+0.02
	Qwen2.5-7B-Inst.	-0.47	+1.78	-0.59	-0.55	-0.22	-0.52	+0.03	-0.22	-0.59	+1.36
	Qwen2.5-72B-Inst.	-0.39	+0.86	-0.89	+0.48	+0.48	-1.22	-0.14	+0.65	-0.39	+0.57
	Qwen3-30B-A3B-Inst.	-3.44	+1.06	-0.31	+1.06	-0.60	-1.04	+1.06	+1.06	-0.69	+1.81
	Qwen3-235B-A22B-Inst.	-0.64	+0.61	-1.14	+0.78	+0.61	-1.89	-0.01	+1.28	-0.64	+1.03
Below university	gemma-3-27b-it	-0.24	-0.12	+1.38	+0.34	+0.34	+0.51	-0.99	-0.66	+0.51	-1.08
	gpt-oss-20b	-0.76	-0.26	-0.51	+0.07	-0.09	-0.16	+0.99	+0.91	-0.51	+0.32
	gpt-oss-120b	+0.09	-0.79	+0.46	-0.29	-0.12	-0.19	+0.09	+1.54	-0.16	-0.62
	Qwen2.5-7B-Inst.	+0.18	+0.31	-0.07	+0.01	+0.18	+0.08	-0.57	+0.18	-0.07	-0.24
	Qwen2.5-72B-Inst.	+0.04	-0.33	+0.04	+0.29	+0.13	+0.29	-0.46	+0.12	+0.42	-0.54
	Qwen3-30B-A3B-Inst.	+0.03	-0.35	-0.22	+0.03	+0.03	-0.27	+0.03	+0.03	+0.28	+0.44
	Qwen3-235B-A22B-Inst.	-0.66	+0.59	+0.72	-0.07	+0.59	-0.21	-0.28	-1.41	-0.28	+1.01
University or above	gemma-3-27b-it	-0.12	+0.01	+0.01	-0.32	+0.01	+0.41	+0.01	+0.34	-0.37	+0.01
	gpt-oss-20b	+0.32	+0.45	+0.07	+0.20	-0.30	-0.10	+0.20	+0.03	-1.55	+0.70
	gpt-oss-120b	+0.21	-0.54	+1.21	-0.38	+0.12	-0.24	-0.42	+0.79	+0.46	-1.21
	Qwen2.5-7B-Inst.	+0.00	+0.38	-0.25	+0.00	+0.00	+0.00	+0.00	+0.00	-0.13	+0.00
	Qwen2.5-72B-Inst.	-0.19	+0.31	-0.06	-0.06	-0.23	+0.44	+0.06	+0.10	-0.06	-0.31
	Qwen3-30B-A3B-Inst.	-0.14	-0.14	-0.02	-0.14	+0.53	+0.06	-0.14	-0.14	-0.14	+0.28
	Qwen3-235B-A22B-Inst.	+0.93	-0.44	-0.07	+0.06	-0.44	+0.66	-0.44	+0.22	-0.44	-0.03

Table 38: Per-model centered-mean deltas (Δ) for PVQ-21: individual model shifts under persona prompting. All cell values are rounded to two decimals using round-half-up.

Condition	Model	Ach	Ben	Con	Hed	Pow	Sec	SD	Sti	Tra	Uni
Male	gemma-3-27b-it	+0.53	-0.23	-0.73	+0.03	+1.28	+0.78	-0.23	+0.28	-0.48	-1.23
	gpt-oss-20b	-0.79	-0.04	-0.79	+0.21	+0.71	+0.71	-0.54	-0.04	-0.04	+0.63
	gpt-oss-120b	-1.01	+0.24	-1.26	+1.74	-0.76	-0.26	+1.24	+1.49	-0.01	-1.43
	Qwen2.5-7B-Inst.	+0.02	+1.02	+0.02	+0.02	+0.02	-0.48	+0.02	+0.02	+0.02	-0.65
	Qwen2.5-72B-Inst.	+0.87	-0.63	+0.12	+0.12	+0.62	+0.12	-0.88	+0.12	+0.12	-0.55
	Qwen3-30B-A3B-Inst.	+0.00	+0.00	-0.25	+0.00	+0.25	+0.00	+0.00	+0.00	+0.00	+0.00
	Qwen3-235B-A22B-Inst.	+0.03	-0.23	-0.73	+0.78	-0.23	+1.28	-0.23	+1.03	-1.48	-0.23
Female	gemma-3-27b-it	-0.18	+0.33	+0.33	+0.08	-0.18	+0.83	-0.68	+0.33	-0.68	-0.18
	gpt-oss-20b	-0.39	+0.86	-0.64	-0.14	+0.36	+0.86	+0.36	-0.39	-0.89	+0.03
	gpt-oss-120b	-0.82	-0.07	-0.82	+0.68	-0.32	-0.57	-0.32	+1.43	+0.93	-0.15
	Qwen2.5-7B-Inst.	-0.14	+0.86	+0.11	-0.14	-0.14	-0.64	+1.11	-0.39	-0.14	-0.48
	Qwen2.5-72B-Inst.	-0.08	+0.18	-0.08	-0.08	-0.08	-0.08	-0.33	-0.08	+0.68	-0.08
	Qwen3-30B-A3B-Inst.	-0.03	-0.03	-0.28	-0.03	-0.03	-0.03	-0.03	-0.03	+0.48	-0.03
	Qwen3-235B-A22B-Inst.	-0.48	-0.23	-0.73	+0.78	-0.23	+0.53	-0.23	+1.03	-0.23	-0.23
20–39 years old	gemma-3-27b-it	+0.07	-0.18	-0.68	+0.32	+0.82	+0.57	-0.18	+0.32	-0.18	-0.85
	gpt-oss-20b	-0.03	+0.23	-1.78	+0.48	+0.23	+0.73	-0.03	+0.23	-0.53	+0.48
	gpt-oss-120b	-0.37	+0.38	-1.87	+1.38	-1.37	+0.63	+0.88	+0.88	+0.63	-1.20
	Qwen2.5-7B-Inst.	-0.23	+1.28	-0.23	-0.23	-0.23	-0.48	+1.03	-0.48	-0.23	-0.23
	Qwen2.5-72B-Inst.	+0.64	-0.11	+0.14	-0.11	+0.14	-0.11	-0.36	+0.39	-0.36	-0.28
	Qwen3-30B-A3B-Inst.	-0.08	-0.08	-0.08	-0.08	+1.18	-0.08	-0.08	-0.08	-0.58	-0.08
	Qwen3-235B-A22B-Inst.	-0.43	-0.18	-0.43	+0.83	-0.18	+0.83	-0.18	+0.58	-0.68	-0.18
80+ years old	gemma-3-27b-it	-1.40	+0.10	+2.60	-0.40	+0.60	+2.10	-1.15	-2.90	+1.35	-0.90
	gpt-oss-20b	-1.48	+1.77	-0.23	-1.23	+0.02	+2.52	-1.23	-1.48	+1.27	+0.10
	gpt-oss-120b	-0.95	+1.30	-1.45	+0.05	-0.45	+1.80	+0.30	-1.20	+2.05	-1.45
	Qwen2.5-7B-Inst.	+0.25	+0.25	+0.25	+0.25	+0.25	-0.50	+0.50	+0.00	+0.25	-1.50
	Qwen2.5-72B-Inst.	-1.23	+1.02	+1.27	-2.23	+0.52	+1.77	-1.48	-2.23	+2.52	+0.10
	Qwen3-30B-A3B-Inst.	-1.85	+0.40	+0.15	+0.40	-0.60	+0.40	+0.40	-0.60	+0.90	+0.40
	Qwen3-235B-A22B-Inst.	-4.25	+0.50	+1.75	+0.50	+0.50	+2.25	+0.50	-3.25	+1.00	+0.50
Right-wing	gemma-3-27b-it	+1.03	-1.23	+2.53	-1.48	+2.03	+2.53	-0.73	-2.23	+0.03	-2.48
	gpt-oss-20b	+1.08	-0.68	-0.93	-0.68	+3.58	+2.58	-1.18	-0.68	-0.43	-2.68
	gpt-oss-120b	-0.16	-1.91	-0.91	+0.09	+1.09	+2.09	+1.34	-0.66	+1.59	-2.58
	Qwen2.5-7B-Inst.	-0.88	-1.13	+2.87	-0.88	-0.88	+3.12	+1.12	-1.13	+0.87	-3.05
	Qwen2.5-72B-Inst.	+1.94	-1.81	+1.94	-2.06	+3.19	+2.44	-1.56	-1.81	+1.19	-3.47
	Qwen3-30B-A3B-Inst.	+0.22	+0.22	+0.97	+0.22	+2.47	+0.22	+0.22	-0.78	-1.28	-2.45
	Qwen3-235B-A22B-Inst.	-0.08	-2.33	+2.17	-0.58	+4.67	+3.17	-1.08	-1.08	-1.08	-3.75
Left-wing	gemma-3-27b-it	-1.00	+1.00	-0.25	-0.25	-0.25	-0.50	+1.00	+0.00	-0.75	+1.00
	gpt-oss-20b	-1.95	+1.05	-1.20	-0.45	+0.05	+0.80	+0.55	-0.45	-0.45	+2.05
	gpt-oss-120b	-0.26	-0.01	-0.76	-0.01	-0.26	-0.26	+0.24	+0.99	-0.51	+0.82
	Qwen2.5-7B-Inst.	-0.57	+2.68	-0.57	-0.57	-0.57	-1.07	+1.18	-0.82	-0.57	+0.92
	Qwen2.5-72B-Inst.	-1.05	+1.45	-0.55	-0.30	+0.70	-1.55	-0.55	+0.70	-0.05	+1.20
	Qwen3-30B-A3B-Inst.	-2.88	+1.38	-0.13	+1.38	-0.88	-2.88	+1.38	+1.38	-0.13	+1.38
	Qwen3-235B-A22B-Inst.	-3.25	+1.50	-1.00	+0.75	+1.50	-1.00	+1.50	+0.25	-1.75	+1.50
Below university	gemma-3-27b-it	-0.56	-0.81	+1.94	-0.06	+0.94	+1.19	-0.81	-1.06	+0.69	-1.48
	gpt-oss-20b	-1.63	+0.38	-0.13	-0.13	+0.88	+1.38	-0.38	-0.63	-0.13	+0.38
	gpt-oss-120b	-0.13	-0.63	+0.37	-0.13	-0.13	-0.13	-0.13	+0.37	+0.62	-0.05
	Qwen2.5-7B-Inst.	+0.28	+0.03	+0.28	+0.28	+0.28	-0.48	+0.28	+0.03	+0.28	-1.23
	Qwen2.5-72B-Inst.	-0.41	-0.41	+0.09	+0.34	+0.34	+0.34	-0.91	+0.34	+1.09	-0.83
	Qwen3-30B-A3B-Inst.	+0.18	+0.18	-0.08	+0.18	-0.33	+0.18	+0.18	+0.18	-0.83	+0.18
	Qwen3-235B-A22B-Inst.	-2.70	+0.30	-0.20	+1.05	+0.30	+1.05	-0.20	+0.30	-0.20	+0.30
University or above	gemma-3-27b-it	-0.45	-0.45	+0.30	-0.45	+0.30	+0.55	+0.05	+0.05	+0.05	+0.05
	gpt-oss-20b	+0.38	+0.13	-0.12	-1.12	+0.88	+1.38	-0.62	-1.12	-1.12	+1.30
	gpt-oss-120b	-0.19	-0.19	+0.31	+0.06	-0.19	-0.19	+0.56	-0.19	-0.19	+0.22
	Qwen2.5-7B-Inst.	+0.12	+0.37	+0.37	+0.12	+0.12	-0.38	+0.12	+0.12	+0.12	-1.05
	Qwen2.5-72B-Inst.	+0.58	+0.08	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	+0.58	-0.18
	Qwen3-30B-A3B-Inst.	+0.03	+0.03	-0.23	+0.03	+0.03	+0.03	+0.03	+0.03	+0.03	+0.03
	Qwen3-235B-A22B-Inst.	-0.50	-0.25	-0.75	+0.75	-0.25	+0.75	-0.25	+0.50	+0.25	-0.25

Table 39: Cross-model standard deviation of PVQ-40 centered-mean deltas for each condition, computed across the seven RQ4 models using the population estimator (`numpy.std`, `ddof=0`). Higher values indicate greater disagreement among models about the direction or magnitude of persona-induced shift.

Condition	Ach	Ben	Con	Hed	Pow	Sec	SD	Sti	Tra	Uni
Male	0.47	0.27	0.46	0.22	0.59	0.39	0.34	0.71	0.80	0.69
Female	0.25	0.51	0.26	0.41	0.28	0.28	0.45	0.27	0.51	0.56
20–39 years old	0.61	0.63	0.32	0.54	0.46	0.36	0.40	0.53	0.38	0.71
80+ years old	1.20	0.61	0.81	1.00	0.52	0.51	0.55	1.09	1.04	0.77
Right-wing	1.50	0.76	0.55	1.02	1.13	0.75	0.83	0.56	0.66	0.77
Left-wing	1.06	0.63	0.54	0.58	0.45	0.54	0.53	0.47	0.62	0.61
Below university	0.35	0.42	0.60	0.20	0.23	0.27	0.58	0.89	0.35	0.67
University or above	0.37	0.37	0.45	0.19	0.30	0.31	0.23	0.28	0.57	0.54