Role-Playing Is Not Personality: Acquiescence in LLM Personality Measurement

Anonymous ACL submission

Abstract

Large language models are increasingly used as agents with different personalities in surveys, social simulation, and public-opinion studies. A common practice is to give an LLM a personality label, ask it to answer human personality questionnaires, and treat the resulting scores as the agent's personality. But this assumes that these questionnaires measure personality in LLMs the same way they do in humans. We test this assumption with a 221-item personality battery across 20 LLMs and 17 persona settings, collecting 375,700 responses. The results show that this assumption does not hold. LLMs tend to agree with questionnaire statements regardless of whether the statements point in opposite directions. This acquiescence bias makes reverse-coded items inconsistent, distorts reliability scores, and collapses the expected personality factor structure. Persona prompts provide an explicit behavioural script for the model to follow, whereas personality traits imply latent dispositions that produce stable but imperfect patterns of behaviour, including inconsistency, conflict, and socially undesirable tendencies. If questionnaire scores reflected a stable trait, the model should answer semantically opposite items in opposite ways. Instead, LLMs often agree with both a statement and its reverse, suggesting that the score is driven by a general tendency to agree rather than by the intended trait. Therefore, persona prompting shows that LLMs can imitate personality roles, but questionnaire scores alone are not reliable evidence of stable personality traits for LLM agents or social simulations. LLMs enact personas rather than simulate personalities.

1 Introduction

Large language models (LLMs) are increasingly treated as agents with personalities. Prior work administers human personality questionnaires to LLMs, computes trait scores, and reports personality profiles (Miotto et al., 2023; Bhandari et al., 2025; Serapio-García et al., 2025; Sorokovikova et al., 2023; Lee et al., 2025; Jiang et al., 2024; Huang et al., 2023). These scores are also used to support "silicon samples" for surveys and behavioural experiments (Argyle et al., 2023; Horton, 2023; Aher et al., 2023; Park et al., 2023; Ferreira et al., 2025), as well as persona-conditioned simulations of opinion dynamics, debate, and polarisation (Chuang et al., 2024; Cau et al., 2025; Chuang et al., 2025). However, these applications rely on a basic assumption: that human personality questionnaires actually measure meaningful traits in LLMs. This paper tests that assumption.

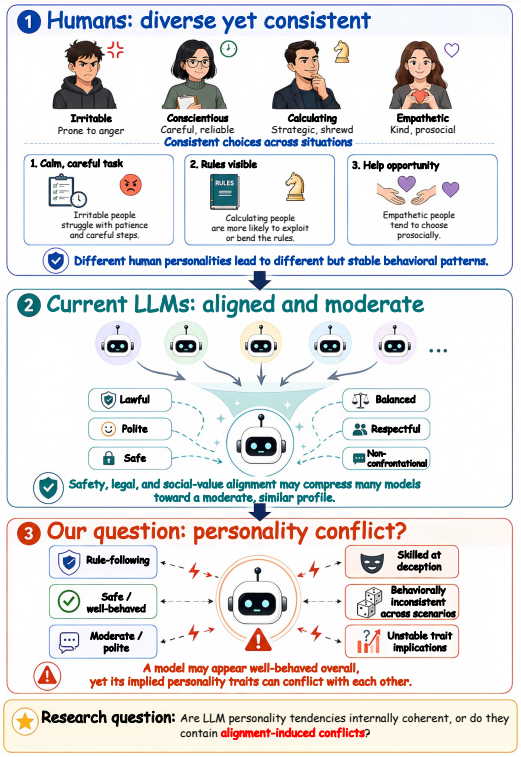

Figure 1: The question this paper asks. Rather than reporting LLM personality scores, we test whether the instruments produce valid measurement at all.

The distinction is important. An LLM may follow a prompt such as "act as an extraverted per-

son" and give answers that look extraverted. But this does not mean that a personality questionnaire has measured a stable trait. A valid trait measure should give consistent answers to items with the same meaning and opposite answers to items with opposite meanings. It should also recover the expected factor structure, agree with related scales, and remain comparable across models and prompt conditions (Cronbach, 1951; Lord and Novick, 1968). These checks are standard in human psychometrics, but are rarely applied to LLM personality scores.

We evaluate this measurement foundation directly. We administer a 221-item battery, including IPIP-NEO-120, SD3, ZKPQ-50-CC, and EPQR-A, to 20 LLMs under 17 persona settings, collecting 375,700 responses. Across standard psychometric checks, the instruments fail to produce valid measurement. The failures trace to a common response style: acquiescence. LLMs often agree with statements regardless of whether the statements point in opposite directions. This makes forward- and reverse-coded items inconsistent, distorts reliability, collapses the expected five-factor structure into three factors, and leaves model identity explaining less than 1% of the variance.

Persona prompts do change LLM responses, but they show role-following rather than valid trait measurement: a persona is an explicit behavioural script to be followed, whereas a personality is a latent psychological structure that produces stable but imperfect patterns of behaviour, including conflict, inconsistency, and socially undesirable tendencies. Our results therefore caution against treating questionnaire-derived scores as evidence that LLM agents have stable, human-like personalities, especially when such scores are used to ground social simulation. LLMs enact personas rather than simulate personalities: their apparent trait differences arise from acquiescence and prompt compliance, not stable trait-based personality structures.

2 Related Work

2.1 Personality Measurement in LLMs

Many studies administer human personality questionnaires to LLMs and report personality profiles (Miotto et al., 2023; Serapio-García et al., 2025; Bhandari et al., 2025; Sorokovikova et al., 2023; Lee et al., 2025; Hagendorff, 2024; Jiang et al., 2024; Huang et al., 2023). However, these instruments are often used without testing whether they remain valid for LLMs. Existing work already suggests problems: LLM responses differ from humans (Xie et al., 2025), show limited temporal stability (Bodroža et al., 2024), assume rather than establish measurement invariance (Sühr et al., 2025), and often skip psychometric validation (Chen et al., 2025). Partial reliability and behavioural analyses further question whether self-reported profiles reflect actual behaviour (Schaeckermann et al., 2024; Han et al., 2025). These scores also support "silicon sample" studies (Argyle et al., 2023; Horton, 2023; Aher et al., 2023; Ferreira et al., 2025) and persona-based social simulations (Chuang et al., 2024; Cau et al., 2025; Chuang et al., 2025; Park et al., 2023), where weak measurement can undermine downstream conclusions (Larooij and Törnberg, 2025; Mannekote et al., 2025). Unlike work that asks what personality an LLM appears to have, we ask whether the questionnaires measure meaningful traits at all.

2.2 Response Style and Acquiescence Bias

Acquiescence bias, or agreeing regardless of item direction, is a classic survey response style (Couch and Keniston, 1960; Crowne and Marlowe, 1960; Paulhus, 1991). Psychometric studies model it as a separate factor and show that it can distort scale structure across populations (Billiet and McClendon, 2000; Baumgartner and Steenkamp, 2007). In LLMs, prior work reports social desirability and acquiescent responding (Acerbi and Stubbersfield, 2024; Salecha and Ireland, 2024; Responsible NLP Group, 2025), while forced-choice and ipsative designs have been proposed to reduce response-style effects (Brown and Maydeu-Olivares, 2011; Treaux et al., 2025). We extend this line by measuring how acquiescence damages validity: reverse-item consistency, reliability, factor structure, and cross-instrument convergence.

2.3 Alignment Effects on LLM Behavior

Alignment methods such as RLHF, constitutional AI, instruction tuning, and preference optimisation shape LLM behaviour (Ouyang et al., 2022; Bai et al., 2022; Wei et al., 2022; Christiano et al., 2017; Rafailov et al., 2023). Prior work links alignment to sycophancy, agreeableness, homogenisation, and value-laden behaviour (Sharma et al., 2024; Zheng et al., 2025; Wang et al., 2025; Kirk et al., 2023; Perez et al., 2023). We study alignment from a measurement perspective: whether aligned models share a common response style, and how far per-

Table 1: Psychometric instruments in the 221-item battery.

Instrument	Items	Format	Domains	Rev.
IPIP-NEO-120	120	Likert-5	5	41
SD3	27	Likert-5	3	5
ZKPQ-50-CC	50	T/F	5	12
EPQR-A	24	Y/N	4	5
Total	**221**	3 formats	**17**	**63**

Table 2: All 20 tested LLMs grouped by model family.

Family	Models
Anthropic	Claude Opus 4.6, Sonnet 4.6
OpenAI	GPT-5.5, GPT-5.2
DeepSeek	V3.2, V4-Flash, V4-Pro
Google	Gemini-3-Flash, Gemini-3.1-Flash-Lite, Gemini-3.1-Pro, Gemini-3-Pro
Alibaba	Qwen3-235B-A22B, Qwen3.5-122B-A10B, Qwen3.5-397B-A17B
Moonshot	Kimi-K2.5, Kimi-K2.6
MiniMax	M2.7
Zhipu	GLM-4.6V, GLM-4.7, GLM-5.1

sona prompting can move them from that default.

3 Methodology

3.1 Psychometric Instruments

Figure 8 provides an overview of the full diagnostic pipeline. We administered a 221-item battery comprising four validated instruments (Table 1): IPIP-NEO-120 (Johnson, 2014), SD3 (Jones and Paulhus, 2014), ZKPQ-50-CC (Aluja et al., 2006), and EPQR-A (Francis et al., 1992). Together they cover 17 domains across three response formats (Likert-5, True/False, Yes/No) and include 63 reverse-coded items.

3.2 Models and Administration Protocol

We tested 20 LLMs spanning 8 model families (Table 2), covering both proprietary and open-weight systems. Technical details for each family are available in the respective model reports (OpenAI, 2023; Gemini Team, Google, 2024; DeepSeek-AI, 2024; Qwen Team, 2024; GLM Team, Zhipu AI, 2024; Moonshot AI, 2025; MiniMax, 2025).

Each model was tested under 17 conditions: a **Default** prompt ("Please answer the following personality questionnaire honestly.") and 16 **MBTI persona** prompts, each assigning a specific MBTI type. All administrations used temperature = 0.7. Each (model, persona, item) cell was sampled 5 times (k=5). Responses are averaged across samples within each cell. This yielded $20 \times 17 \times 221 = 75{,}140$ item-level scores (averaged from 375,700 raw API calls).

3.3 Analysis Pipeline

We applied six checks that psychometricians use before trusting any instrument on a new population.

Internal consistency. We compute Cronbach's α per IPIP domain for each model, using the 17 persona conditions as observations. For a domain with k items, let X_{pi} be the scored response to item i under persona p. Then:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \mathrm{Var}(X_i)}{\mathrm{Var}\left(\sum_{i=1}^{k} X_i\right)}\right) \quad (1)$$

We average α across the 20 models. Human benchmarks come from Johnson (2014).

Factor structure. Let $\mathbf{x}_j \in \mathbb{R}^{17}$ be the vector of domain scores for observation j ($j = 1, \ldots, 340$). We standardize each domain to zero mean and unit variance, then compute the correlation matrix $\mathbf{R}$. We retain factors via the Kaiser criterion ($\lambda_i > 1$) (Kaiser, 1960) and validate with parallel analysis (Horn, 1965): we permute each column of the data matrix independently, recompute eigenvalues, and repeat 500 times. A factor is retained only if its observed eigenvalue exceeds the 95th percentile of the permuted distribution. We extract factor loadings via eigendecomposition of $\mathbf{R}$ followed by varimax rotation (Kaiser, 1958). Factor replication across models is assessed with Tucker's coefficient of congruence (Tucker, 1951).

Reverse-item consistency. For each domain that contains both forward-coded (+) and reverse-coded (−) items, we compute the Pairwise Inconsistency Rate (PIR). For Likert items, a forward-reverse pair (f, r) is inconsistent when both responses are at or above the midpoint, or both are at or below it:

$$\mathrm{PIR} = \frac{1}{|F|\,|R|} \sum_{f \in F} \sum_{r \in R} \mathbf{1}\big[(f \geq 3 \wedge r \geq 3) \vee (f \leq 3 \wedge r \leq 3)\big] \quad (2)$$

where F and R are the sets of forward and reverse item responses (raw Likert values), and the midpoint is 3 on the 1–5 scale. For binary items, inconsistency is $f = r$.

Variance decomposition. We normalize all scored responses to a common 0–1 scale: Likert scores via $x' = (x-1)/4$, binary scores kept as-is.

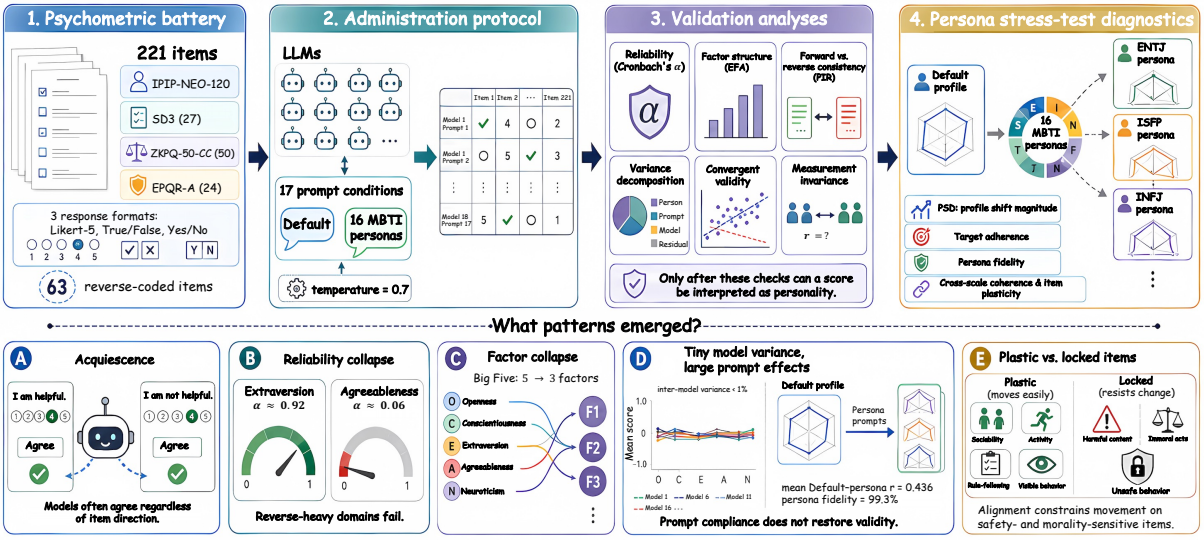

Figure 2: Overview of the methodology. We administer a multi-instrument psychometric battery to LLMs, normalize the response matrix, and run standard validation checks before interpreting any score as evidence of personality.

We then decompose the total sum of squares into marginal components:

$$\begin{aligned} \mathrm{SS}_{\text{total}} &= \sum_i (y_i - \bar{y})^2 \\ &= \mathrm{SS}_{\text{model}} + \mathrm{SS}_{\text{domain}} \\ &\quad + \mathrm{SS}_{\text{persona}} + \mathrm{SS}_{\text{item}} + \mathrm{SS}_{\text{resid}} \end{aligned} \tag{3}$$

where each component groups by the corresponding factor and computes the squared deviation of group means from the grand mean. We run this separately for Likert and binary scales.

Convergent validity. For each of 8 pre-specified construct pairs (e.g., IPIP Neuroticism vs. ZKPQ Neuroticism-Anxiety), we compute Pearson r_p and Spearman r_s across the 20 model-level means. A pair passes if the correlation sign matches the theoretical expectation.

Measurement invariance. For each model, let $\mathbf{d} \in \mathbb{R}^{120}$ be the vector of IPIP-NEO-120 parsed values under the Default persona, and $\mathbf{m}_k$ the same vector under MBTI persona k. We compute:

$$r_k = \mathrm{corr}(\mathbf{d}, \mathbf{m}_k), \quad k = 1, \ldots, 16 \tag{4}$$

Strong invariance requires $r_k > 0.8$ for all persona pairs (Vandenberg and Lance, 2000; Meredith, 1993). We report the mean r across all $20 \times 16 = 320$ pairs.

Persona separation degree (PSD). We quantify how far a persona prompt pushes a model from its default profile. After z-scoring each domain across all $20 \times 17 = 340$ observations, we define:

$$\mathrm{PSD}(m, p) = \frac{1}{\sqrt{17}} \left\| \mathbf{z}_{m,p} - \mathbf{z}_{m,\text{Default}} \right\|_2 \tag{5}$$

where division by $\sqrt{17}$ yields a per-dimension effect size in SD units.

Persona steering diagnostics. PSD measures distance from Default but not direction. We define two targets: an **empirical target** using the leave-one-model-out centroid difference $\mathbf{t}^{\text{emp}}_{m,p} = \bar{\mathbf{z}}_{-m,p} - \bar{\mathbf{z}}_{-m,\text{Default}}$, and a **theoretical target** mapping MBTI letters to expected domain shifts (e.g., E → higher Extraversion, F → higher Agreeableness). For each persona shift $\Delta_{m,p} = \mathbf{z}_{m,p} - \mathbf{z}_{m,\text{Default}}$, we measure target adherence ($\cos(\Delta_{m,p}, \mathbf{t}_{m,p})$) and off-target leakage. We also classify each profile to the nearest leave-one-out persona centroid in z-Euclidean, cosine, and Mahalanobis (Mahalanobis, 1936) distance metrics, yielding a fidelity confusion matrix. Aggregate statistics carry bootstrap 95% confidence intervals (Efron, 1979).

Cross-scale persona coherence and item plasticity. For overlapping constructs measured by multiple instruments (e.g., Extraversion across IPIP/EPQR/ZKPQ), we sign-align each construct and compute a coherence score $|\overline{\Delta}|/\overline{|\Delta|}$, where 1 means all scales move in the same direction and 0 means they cancel. At the item level, we compute each item's mean absolute shift from Default and the rate at which forward/reverse item pairs move in theoretically opposite directions.

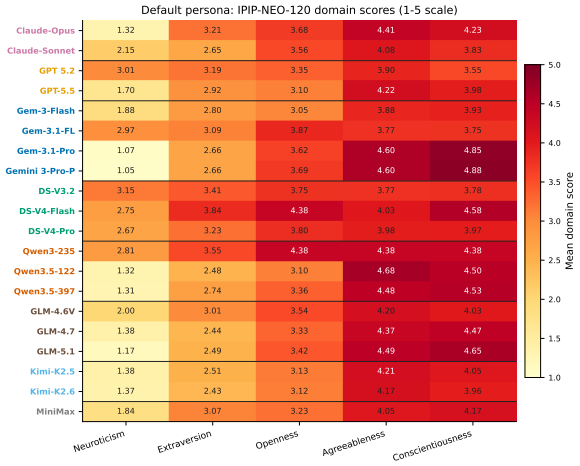

Figure 3: Default persona IPIP-NEO-120 domain scores (z-scored heatmap). All 20 models cluster in a "high A, high C, low N" region. Inter-model range is narrowest for Agreeableness (0.92 on a 1–5 scale) and widest for Neuroticism (2.10).

Table 3: Cronbach's α by IPIP domain: LLM vs. human norms. Mean±SD across n=20 models.

Domain	LLM α	Human α	Gap
Neuroticism	0.813±0.035	0.90	−0.087
Extraversion	0.928±0.008	0.89	+0.038
Openness	0.189±0.160	0.87	−0.681
Agreeableness	0.069±0.311	0.88	−0.811
Conscientiousness	0.535±0.113	0.90	−0.365

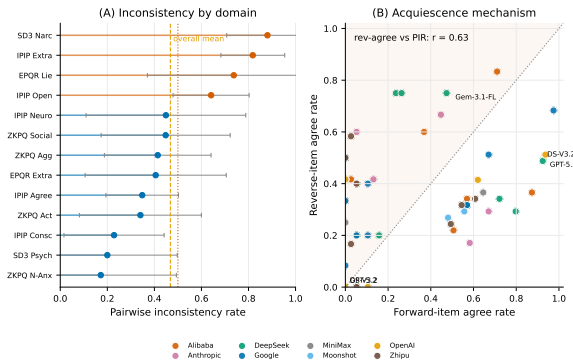

Figure 4: Reverse-item inconsistency and acquiescence mechanism. Left: PIR by domain. Right: Reverse-item agree rate vs. PIR ($r = 0.63$).

4 Experiments and Results

4.1 Default Personality Profiles

Fig. 3 shows the IPIP-NEO-120 scores under the Default (no-persona) prompt. All 20 models cluster in the same corner: high Agreeableness, high Conscientiousness, low Neuroticism. The inter-model range is narrowest for Agreeableness (0.92 on a 1–5 scale) and widest for Neuroticism (2.10). Every model looks like a "nice, responsible, emotionally stable" person—whether it was made by Anthropic, OpenAI, Google, or DeepSeek. Is this a genuine personality, or an artifact of alignment training? The sections that follow trace the evidence from item to domain to factor to model, identifying a single root cause. The per-model MBTI classification under Default is shown in Fig. 20.

4.2 Internal Consistency

The first diagnostic is whether items within a domain agree with each other. Table 3 shows the answer. Extraversion, where only 17% of items are reverse-coded, yields $\alpha = 0.93$. That is close to the human norm of 0.89. But Agreeableness, where 71% of items are reverse-coded, yields $\alpha = 0.06$, essentially zero. The difference is not random. When a model agrees with both "I am helpful" and "I am not helpful," the two items pull in opposite directions and the domain score falls apart. The more reverse-coded items a domain has, the worse its reliability. Spearman correlation between α and reverse-item percentage across the five IPIP domains is $\rho = -1.0$ ($p < 0.01$), confirming that acquiescence, not trait variance, drives reliability. This is the fingerprint of acquiescence Figure 22 confirms that this gradient persists under every persona condition.

4.3 Caught in the Act: Acquiescence at the Item Level

The fingerprint above points squarely to acquiescence, which we measure directly with the Pairwise Inconsistency Rate. PIR is **0.468** [95% CI: 0.413, 0.531], meaning that nearly 47% of forward-reverse item pairs receive inconsistent responses. Table 4 breaks this down by instrument, where *Rev* denotes the proportion of reverse-keyed items receiving an "agree" response and $\Delta = \text{Fwd} - \text{Rev}$ denotes the forward–reverse agree-rate gap.

A striking pattern emerges. For IPIP items, the gap is positive ($\overline{\Delta} = +0.29$), indicating classic acquiescence: models agree more with forward than reverse items. For SD3 items, the gap flips negative ($\overline{\Delta} = -0.33$), reflecting social desirability: models avoid endorsing dark-trait statements such as "I manipulate others." These two response biases both push scores away from genuine trait variation. Across models, reverse-item agreement correlates with PIR ($r = 0.63$), showing that models that acquiesce more are also more inconsistent.

Table 4: Acquiescence by instrument. **Rev**: proportion of reverse-keyed items where the model gave an "agree" response (Likert: score ≥ 3 out of 5; Binary: response $= 1$). Δ: forward-item agree rate minus reverse-item agree rate; positive Δ indicates acquiescence (agreeing with both directions), negative Δ indicates social-desirability rejection. Models grouped by family.

	IPIP		SD3		ZKPQ		EPQR		Overall	
Model	Rev	Δ	Rev	Δ	Rev	Δ	Rev	Δ	Rev	Δ
Claude-Opus-4.6	0.29	+0.42	0.60	−0.37	0.75	−0.25	0.60	−0.49	0.43	+0.09
Claude-Sonnet-4.6	0.24	+0.38	0.60	−0.37	0.67	−0.40	0.60	−0.55	0.38	+0.03
GPT-5.2	0.63	+0.33	1.0	−0.32	0.00	0.00	0.00	+0.11	0.49	+0.10
GPT-5.5	0.46	+0.22	0.60	−0.24	0.42	−0.39	0.00	+0.11	0.43	−0.02
Gemini-3-Pro	0.32	+0.30	0.60	−0.37	0.67	−0.64	0.40	−0.29	0.41	−0.05
Gemini-3-Flash	0.59	+0.16	0.80	−0.30	0.58	−0.58	0.40	−0.29	0.59	−0.13
Gemini-3.1-Pro	0.34	+0.29	0.60	−0.42	0.50	−0.45	0.20	−0.09	0.38	−0.01
Gemini-3.1-Flash	0.71	+0.27	1.0	−0.23	0.42	−0.34	0.40	−0.29	0.65	−0.02
DeepSeek-V3.2	0.59	+0.36	1.0	−0.36	0.75	−0.17	0.20	−0.15	0.62	+0.09
DeepSeek-V4-Flash	0.37	+0.46	0.80	−0.30	0.83	−0.18	0.20	+0.01	0.48	+0.19
DeepSeek-V4-Pro	0.44	+0.35	0.80	−0.44	0.83	−0.20	0.20	−0.09	0.52	+0.08
Qwen3-235B-A22B	0.39	+0.48	0.80	−0.07	0.83	−0.04	1.0	−0.63	0.56	+0.22
Qwen3.5-122B	0.34	+0.22	0.60	−0.42	0.42	−0.42	0.00	+0.05	0.35	−0.04
Qwen3.5-397B	0.39	+0.22	0.60	−0.33	0.42	−0.39	0.00	+0.05	0.38	−0.03
GLM-4.6V	0.44	+0.28	0.80	−0.35	0.42	−0.34	0.00	+0.21	0.43	+0.04
GLM-4.7	0.40	+0.18	0.64	−0.38	0.53	−0.51	0.32	−0.27	0.44	−0.10
GLM-5.1	0.32	+0.21	0.60	−0.39	0.72	−0.63	0.52	−0.47	0.43	−0.11
Kimi-K2.5	0.41	+0.21	0.60	−0.28	0.42	−0.42	0.00	+0.05	0.40	−0.04
Kimi-K2.6	0.41	+0.15	0.60	−0.37	0.42	−0.42	0.40	−0.35	0.43	−0.11
MiniMax-M2.7	0.44	+0.32	0.80	−0.30	0.42	−0.36	0.00	+0.11	0.43	+0.05
Mean	0.43	+0.29	0.72	−0.33	0.55	−0.36	0.27	−0.16	0.46	+0.01

Interestingly, persona injection *improves* consistency. Under the Default prompt, PIR agreement is 0.674. Under MBTI persona prompts, it rises to 0.834 ($t = 10.70$, $p < 0.0001$). When you give a model a specific role to play, the tension between the agreeableness produced by alignment and honest responding decreases. The model can be consistent within its assigned character. Figure 21 shows the per-model PIR breakdown under Default vs. MBTI prompting.

4.4 The Cascade: Factor Structure Collapse

Acquiescence begins at the item level but propagates upward by distorting the inter-domain correlations that define the Big Five. To test this, we ran an exploratory factor analysis (EFA) on the 17 domain scores across all $20 \times 17 = 340$ observations. Unlike human samples, where this procedure typically recovers five factors, LLM responses support only three factors under both the Kaiser criterion and parallel analysis. As shown in Table 5, Factor 1 captures Extraversion and Sociability (EPQR Extraversion $= 0.96$, ZKPQ Sociability $= 0.96$, IPIP Extraversion $= 0.94$). Factor 2 merges Neuroticism and Agreeableness (IPIP Neuroticism $= 0.72$, EPQR Neuroticism $= 0.91$, IPIP Agreeableness $= 0.90$). Factor 3 captures Conscientiousness versus Impulsivity (IPIP Conscientiousness $= -0.91$, ZKPQ ImpulsiveSS $= 0.84$, EPQR Psychoticism $= 0.88$). The Neuroticism–Agreeableness merger is especially diagnostic: constructs that are separable in humans collapse in LLMs because both are shaped by the same acquiescent, socially desirable response pattern.

This collapse is universal. Tucker's $\phi = 0.993$ across all 20 models, meaning every model produces the same distorted structure. Leave-one-model-out EFA always recovered exactly three factors, with the first eigenvalue stable between 6.94 and 7.02. The same pattern holds when EFA is run separately by model or by model family: all 20 models and all 8 families independently yield three Kaiser factors, and no configuration recovers the expected five-factor structure. Figure 19 displays the leave-one-model-out eigenvalue spectra confirming this universality.

Table 5: EFA factor loadings (3-factor solution, n=340). Bold: $|l| > 0.5$. All 20 leave-one-model-out analyses recover exactly 3 factors.

Domain	F1	F2	F3
IPIP N	−0.30	**0.72**	0.39
IPIP E	**0.94**	0.03	0.29
IPIP O	−0.04	**0.55**	**0.58**
IPIP A	0.05	**0.90**	−0.22
IPIP C	−0.00	−0.04	−0.91
SD3 Mach.	−0.09	−0.82	0.04
SD3 Narc.	**0.91**	−0.20	0.16
SD3 Psych.	0.38	−0.59	**0.63**
ZKPQ Act.	**0.90**	−0.27	−0.12
ZKPQ Agg.	0.48	−0.60	0.36
ZKPQ ImpSS	0.34	0.08	**0.84**
ZKPQ N-Anx	−0.27	**0.83**	0.08
ZKPQ Soc.	**0.96**	0.04	0.13
EPQR P	0.25	−0.11	**0.88**
EPQR E	**0.96**	−0.05	0.12
EPQR N	−0.04	**0.91**	0.02
EPQR Lie	−0.08	0.07	−0.89

Table 6: Sequential variance decomposition. Model variance: <1% for both scale types.

Source	Likert (%)	Binary (%)
Model	**0.35**	**0.51**
Domain	23.93	5.27
Persona	4.09	14.06
Item	36.13	14.81
Residual	35.51	65.35

4.5 The Cascade: Model Convergence

Factor collapse is one symptom. Model convergence is another. If alignment training creates the same acquiescence in every model, then inter-model variance should be negligible. Table 6 partitions the total response variance into five sources. The answer is unambiguous: for Likert scales, **inter-model variance accounts for only 0.35%**. Which model answered the question barely matters. What matters is which item was asked (36.1%), which domain it belongs to (23.9%), and residual noise (35.5%). Binary scales tell the same story. Cross-model personality comparisons are effectively meaningless at this resolution.

The practical implication is straightforward. If you administer a personality questionnaire to two different LLMs and get different scores, the difference is almost certainly noise, not personality.

4.6 Convergent Validity and DIF

One more check. If the instruments measure real traits, then related constructs across different scales should correlate in the expected direction. Table 7 tests 8 such pairs. Seven show the expected

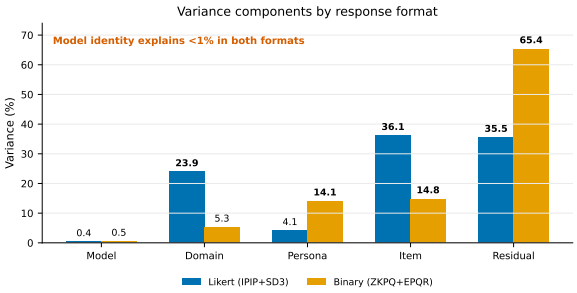

Figure 5: Sequential variance decomposition. Inter-model variance: 0.35% (Likert), 0.51% (Binary).

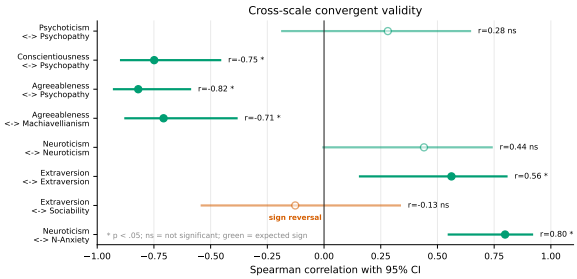

Figure 6: Convergent validity across instruments. Extraversion/Sociability shows sign reversal.

sign. The exception is Extraversion and Sociability, which should correlate positively but instead show a sign reversal ($r_s = -0.13$). Extraverted models in one scale are not sociable in another. The magnitudes are also weaker than human baselines across the board.

A DIF analysis across model families found no significant differences in any domain (all $p > 0.16$; details in Appendix B).

4.7 Ruling Out Alternative Explanations

How do LLM responses compare to simulated data? Table 8 shows three baselines: random responses, pure acquiescence (always agree), and trait-plus-acquiescence (genuine structure with acquiescence). LLM alphas fall closest to the Random and Pure Acquiescence baselines. They are far from the Trait+Acquiescence strategy ($\alpha > 0.99$) that mimics genuine trait structure. In other words, LLMs do not look like they have personality with some acquiescence on top. They look like they have acquiescence with no personality underneath. Figure 23 visualises the full calibration against the human baseline.

The mean Pearson correlation between Default and MBTI persona profiles is $r = 0.461$. No model-persona pair achieves strong invariance ($r > 0.8$). Changing the prompt substantially rewrites

Table 7: Convergent validity: correlations between theoretically overlapping constructs. Both Pearson (r_p) and Spearman (r_s) reported. "OK" = sign matches expectation.

Pair	Exp.	r_p	p_p	r_s	p_s	OK
Neuroticism × N-Anx.	+	0.64	0.002	0.80	0.000	Yes
Extraversion × Sociability	+	−0.10	0.670	−0.13	0.594	**No**
Extraversion × Extraversion	+	0.51	0.020	0.56	0.010	Yes
Neuroticism × Neuroticism	+	0.48	0.031	0.44	0.052	Yes
Agreeableness × Machiavellianism	−	−0.71	0.000	−0.72	0.000	Yes
Agreeableness × Psychopathy	−	−0.81	0.000	−0.82	0.000	Yes
Conscientiousness × Psychopathy	−	−0.73	0.000	−0.75	0.000	Yes
Psychoticism × Psychopathy	+	0.11	0.646	0.28	0.230	Yes

Table 8: Synthetic baselines: Cronbach's α by domain.

Strategy	N	E	O	A	C
Random	0.078	−0.013	0.062	0.044	−0.108
Pure Acq.	0.016	0.025	−0.008	0.096	−0.031
Trait+Acq.	0.997	0.998	0.996	0.997	0.996
LLM Obs.	**0.813**	**0.928**	**0.189**	**0.069**	**0.535**
Human	0.90	0.89	0.87	0.88	0.90

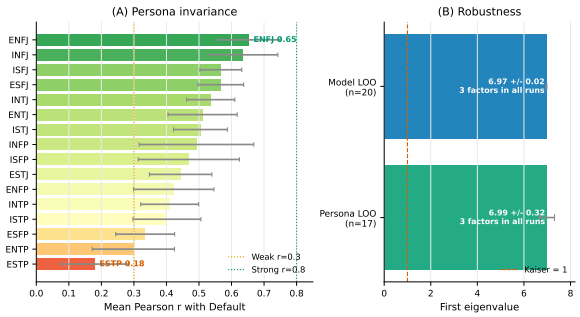

Figure 7: Persona-level measurement invariance. Most invariant: ENFJ ($r = 0.653$). Least: ESTP ($r = 0.179$).

the response profile.

4.8 The Stress Test: Persona Prompting

The last question: can we rescue these instruments by steering models with explicit persona prompts? if the instruments measured stable traits, persona prompts should move scores without repairing the failed psychometrics. That is what we find. Persona prompts shift profiles by more than one SD from Default on average, and nearest-centroid classification recovers the prompted MBTI type in 319 of 320 cases (99.7%). However, this is prompt compliance rather than restored validity: cross-scale persona coherence is imperfect (0.839), forward/reverse pairs move in the theoretically opposite raw-score direction only 69.6% of the time, and the Big Five factor collapse remains: running EFA within the Default persona condition recovers the same 3-factor pattern, and no MBTI persona condition recovers the expected 5-factor Big Five structure. Full persona-steering analyses, including default bias, model clustering, MBTI factorial effects, scale-format sensitivity, fidelity matrices, and item plasticity, are reported in Appendix A.

5 Conclusion

Prior work has reported inconsistent responses (Sühr et al., 2025), social desirability (Acerbi and Stubbersfield, 2024), and instability (Pelet et al., 2025) in LLM personality measurement. Our findings suggest that these phenomena reflect a common mechanism: LLMs are good at role-playing, in the sense that they can follow persona instructions and produce recognizable behavioural differences, but they are not good at genuinely simulating personality differences. A persona is an explicit behavioural script to be followed, whereas a personality is a latent psychological structure that produces stable but imperfect patterns of behaviour, including conflict, inconsistency, and socially undesirable tendencies. The underlying mechanism is therefore better understood as acquiescence plus persona compliance, rather than a trait-based personality structure. For coarse-grained simulations that only require behaviourally distinguishable agents, persona prompting may still be useful. However, for simulations that aim to capture real psychological dynamics, especially those involving conflict, extreme opinions, polarization, or a neutral default agent (Larooij and Törnberg, 2025), current LLM agents are likely to be systematically unreliable: they may produce outcomes that are overly moderate, overly consistent, and overly convergent. Together, these findings caution against using questionnaire-based LLM personality scores as valid trait measurements in silicon-sample studies or persona-conditioned social simulations unless reliability, factor structure, convergent validity, and measurement invariance are established.

6 Limitations

Our study has several limitations. First, we use a **fixed temperature of 0.7** with k=5 repetitions per cell. Although the within-sample consistency analysis (Appendix D) quantifies stochastic variance and shows that it spans an order of magnitude across models, varying temperature systematically would further clarify the boundary between sampling noise and genuine response tendencies. Second, we lack **base vs. instruct model comparisons**; all 20 models are instruction-tuned, so we cannot isolate alignment effects from base-model behavior. Testing base models before and after alignment would directly test the causal mechanism we propose. Third, we compare against **published human norms** rather than collecting new human data, which limits direct population-level comparisons. Fourth, our persona manipulation uses **MBTI labels only**; richer persona descriptions (e.g., vignettes or biographical sketches) might produce different steering dynamics. Fifth, the study covers **English-language instruments** administered in English; cross-linguistic generalisation remains untested, though prior work suggests that acquiescence in LLMs is robust across languages (Responsible NLP Group, 2025). Finally, our conclusions are specific to the four instruments studied; other personality inventories (e.g., forced-choice or ipsative formats) may behave differently.

Ethics Statement

This study uses publicly available LLMs and public or openly documented psychometric instruments, following the relevant access terms and citing the original creators. No human subjects were involved. We do not claim that LLMs have or lack genuine personality; our contribution concerns the *measurement properties* of the instruments when applied to a novel population. A potential ethical concern is that our findings could be misused to justify using forced-choice or other instruments that bypass acquiescence, producing scores that appear valid but may still not reflect traits. We emphasise that passing psychometric checks is necessary but not sufficient for valid personality measurement in LLMs, and that downstream applications should report measurement quality transparently. Any released code or derived response data are intended only for research replication and analysis, not for psychological diagnosis or high-stakes profiling.

References

Alberto Acerbi and Joseph M. Stubbersfield. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, 3(12):pgae533.

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

Anton Aluja, Jérôme Rossier, Luis F. Garcia, Alois Angleitner, Marvin Kuhlman, and Marvin Zuckerman. 2006. A cross-cultural shortened form of the zkpq (zkpq-50-cc) adapted to english, french, german, and spanish languages. *Personality and Individual Differences*, 41(4):619–628.

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

Yuntao Bai et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Thomas Baumgartner and Jan-Benedict E. M. Steenkamp. 2007. Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 44(2):322–335.

Pranav Bhandari, Usman Naseem, et al. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. *arXiv preprint arXiv:2502.05248*.

Jaak B. Billiet and McKee J. McClendon. 2000. Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, 7(4):608–628.

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):240180.

Anna Brown and Alberto Maydeu-Olivares. 2011. Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71(3):460–502.

Erica Cau, Valentina Pansanella, and Dino Pedreschi. 2025. Language-driven opinion dynamics in Agent-Based simulations with LLMs. *arXiv preprint arXiv:2502.19098*.

Yihan Chen et al. 2025. Humanizing LLMs: A survey of psychological measurements. *arXiv preprint arXiv:2505.00049*.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems (NeurIPS)*.

Yun-Shiuan Chuang, Agam Goyal, and Nikunj Harlalka. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of NAACL-HLT*.

Yun-Shiuan Chuang, Ruixuan Tu, and Chengtao Dai. 2025. DEBATE: A Large-Scale benchmark for evaluating opinion dynamics in Role-Playing LLM agents. *arXiv preprint*.

Arthur Couch and Kenneth Keniston. 1960. Yeasayers and naysayers: agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, 60(2):151–174.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Douglas P. Crowne and David Marlowe. 1960. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4):349–354.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26.

Gregorio Ferreira, Jacopo Amidei, and Ruben Nieto. 2025. Matching GPT-simulated populations with real ones in psychological studies—the case of the EPQR-A personality test. *ACM Transactions on Computing for Healthcare*.

Leslie J. Francis, L. B. Brown, and R. Philipchalk. 1992. The development of an abbreviated form of the revised eysenck personality questionnaire (epqr-a): Its use among students in england, canada, the u.s.a. and australia. *Personality and Individual Differences*, 13(4):443–449.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

GLM Team, Zhipu AI. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*.

Thilo Hagendorff. 2024. Machine personality traits: A test of psychometric properties. *arXiv preprint arXiv:2402.11596*.

Pengrui Han, Rafal Kocielnik, and Peiyang Song. 2025. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. *arXiv preprint arXiv:2509.03730*.

John L. Horn. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

John J. Horton. 2023. Homo silicus: Using LLMs to generate silicon samples for economic research. *NBER Working Paper No. 31194*.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R. Lyu. 2023. Who is ChatGPT? benchmarking LLMs' psychological portrayal using PsychoBench. *arXiv preprint arXiv:2310.01386*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of NAACL 2024*.

John A. Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory. *Journal of Research in Personality*, 51:78–89.

Daniel N. Jones and Delroy L. Paulhus. 2014. Introducing the short dark triad (sd3): A brief measure of dark personality traits. *Assessment*, 21(1):28–41.

Henry F. Kaiser. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Henry F. Kaiser. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20:141–151.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Tim Rocktäschel. 2023. Understanding the effects of RLHF on LLM generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Maik Larooij and Petter Törnberg. 2025. Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*.

Seolhwa Lee, Seungone Kim, et al. 2025. TRAIT: Personality testset for LLMs with psychometrics. In *Findings of NAACL 2025*, pages 8412–8452.

Frederic M. Lord and Melvin R. Novick. 1968. Statistical theories of mental test scores. *Addison-Wesley*.

Prasanta Chandra Mahalanobis. 1936. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55.

Amogh Mannekote, Adam Davies, and Guohao Li. 2025. Do role-playing agents practice what they preach? Belief-Behavior consistency in LLM-Based simulations of human trust. *arXiv preprint arXiv:2507.02197*.

William Meredith. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543.

MiniMax. 2025. MiniMax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2023. Who is GPT-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*.

Moonshot AI. 2025. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744.

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

Delroy L. Paulhus. 1991. Measurement and control of response bias. *Measurement and Evaluation in Counseling and Development*, 24(4):389–394.

Colin Pelet et al. 2025. Persistent instability in LLM's personality measurements: Effects of scale, reasoning and conversation history. *Proceedings of AAAI 2025*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. *Findings of the Association for Computational Linguistics (ACL)*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Responsible NLP Group. 2025. Acquiescence bias in large language models. *Findings of EMNLP 2025*.

Gaurav Salecha and Charles Ireland. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, 3(12):pgae533.

Henning Schaeckermann, Chengpeng Chuang, Milan Zecevic, and Uli Hahn. 2024. Measuring and improving the reliability of LLM personality assessments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Gregory Serapio-García, Mona Safdari, et al. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*.

Mrinank Sharma et al. 2024. Towards understanding sycophancy in language models. In *ICLR 2024*.

Alexandra Sorokovikova et al. 2023. Evaluating and inducing personality in pre-trained language models. In *NeurIPS 2023*.

Tom Sühr, Felix E. Dorner, Samira Samadi, and Ayanna Kelava. 2025. Challenging the validity of personality tests for large language models. *Proceedings of the ACM Web Conference 2025*.

Camille Treaux et al. 2025. Decoding LLM personality measurement: Forced-choice vs. Likert. *Findings of ACL 2025*.

Ledyard R. Tucker. 1951. A method for synthesis of factor analysis studies. *Personnel Research Section Report No. 984*.

Robert J. Vandenberg and Charles E. Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1):4–70.

Zhi Wang et al. 2025. Aligning to what? limits to RLHF-based alignment. *arXiv preprint arXiv:2503.09025*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *International Conference on Learning Representations (ICLR)*.

Yijin Xie, Yuyang Chen, Yifan Zhang, Zhuo Zhang, Yue Wang, and Haipeng Qiu. 2025. AIPsychoBench: Understanding the psychometric differences between LLMs and humans. *Transactions on Privacy-Preserving and Security (TOPS)*.

Zhexin Zheng et al. 2025. Too nice to tell the truth: Quantifying agreeableness-driven sycophancy in role-playing LLMs. *arXiv preprint arXiv:2604.10733*.

A Persona Steering Supplement

All persona-steering diagnostics are computed from the same 375,700 completed responses. These analyses are not the primary evidence for the paper's central claim. Instead, they answer a complementary question: once we know the instruments fail standard psychometric checks, what exactly happens when we deliberately push the model with a persona prompt? The answer is consistent across the analyses below. Persona prompts produce large, highly recoverable role-playing shifts, but those shifts do not repair the instrument: factor structure, reverse-item behavior, and cross-scale coherence remain contaminated by response style.

Figure 8 summarises the full diagnostic pipeline. Six standard psychometric checks are applied to the 221-item battery across 20 models, and every check returns a FAIL. A seventh panel then confirms that persona prompting can shift scores dramatically without repairing any of the underlying failures—the persona stress test is the strongest possible manipulation, and even it does not restore validity.

A.1 Model Family and Persona Effects

This section asks whether model identity or persona instructions dominate the auxiliary structure. The first check is model clustering under the Default prompt. If LLMs had stable model-specific personality profiles, we would expect vendor or family clusters to appear in the 17-domain profile space. Instead, model-family clustering is weak. Ward-linkage clustering on z-scored Default profiles does not recover vendor groups cleanly: DeepSeek-V4-Flash clusters with Qwen3-235B, while Claude models cluster with DeepSeek-V3.2/V4-Pro. Cross-family distance is only 1.27× within-family distance. This supports the main-text variance decomposition: differences between model families are small compared with shared response tendencies shaped by alignment.

The second check asks whether MBTI prompts are interpreted compositionally. They are. Fig. 10 shows that the E/I axis dominates social and energetic traits: E/I has the largest effects on EPQR Extraversion ($\beta = 1.01$), ZKPQ Sociability ($\beta = 0.98$), IPIP Extraversion ($\beta = 0.96$), SD3 Narcissism ($\beta = 0.94$), and ZKPQ Activity ($\beta = 0.87$). J/P mainly affects IPIP Conscientiousness ($\beta = 0.96$), ZKPQ Impulsive Sensation Seeking ($\beta = -0.81$), EPQR Lie ($\beta = 0.78$), and EPQR Psychoticism ($\beta = -0.73$). S/N primarily affects IPIP Openness ($\beta = 0.77$), while T/F strongly affects IPIP Agreeableness ($\beta = 0.93$), EPQR Neuroticism ($\beta = 0.87$), and SD3 Machiavellianism ($\beta = -0.76$).

Table 9: Persona Separation Degree by model. $\overline{\text{PSD}}$: mean across 16 personas. $\text{PSD}_{\max}$: peak. CV: coefficient of variation.

Model	$\overline{\text{PSD}}$	$\text{PSD}_{\max}$	CV	Peak persona
Top 5 (most malleable)				
Gemini-3-Pro	1.384	2.06	0.232	ENTP
Gemini-3.1-Pro	1.363	2.03	0.235	ENTP
Qwen3-235B	1.344	1.73	0.174	ESTJ
GLM-5.1	1.339	2.03	0.266	ESTP
GLM-4.7	1.269	1.93	0.271	ENTP
Bottom 5 (least malleable)				
DeepSeek-V3.2	1.117	1.47	0.230	ESTP
GPT-5.2	1.125	1.36	0.141	ESTP
Claude-Sonnet-4.6	1.080	1.68	0.267	ESTP
DeepSeek-V4-Pro	1.016	1.37	0.220	ESTP
MiniMax-M2.7	0.996	1.31	0.214	INFP
Mean (all 20)	**1.198**	**1.67**	**0.219**	

This is important because it rules out a trivial failure mode. The persona prompts are not ignored. Models do read and operationalize the MBTI instruction. However, successful role uptake is not the same thing as valid personality measurement. The role effects are large and stereotype-consistent, but the main psychometric failures remain: the Big Five still collapses, model variance remains tiny, and reverse-coded items remain contaminated.

Scale format also matters. Binary-format scales are 75% more sensitive to persona injection than Likert scales after normalizing by scale range. This is a methodological warning: forced yes/no formats amplify persona-induced movement, but amplification is not necessarily higher validity. In our classification ablation below, binary-only profiles are less stable than Likert or IPIP-only profiles. The binary scales therefore make prompt effects easier to see, while also reducing graded trait information.

A.2 Persona Separation and Fidelity

The previous section showed that MBTI prompts produce interpretable main effects. We next ask two stricter questions. First, how far does each persona move a model from its own Default baseline? Second, does the movement go toward the requested persona, or merely perturb the response profile?

Persona Separation Degree (PSD) answers the first question by measuring the normalized Euclidean distance between a persona profile and the

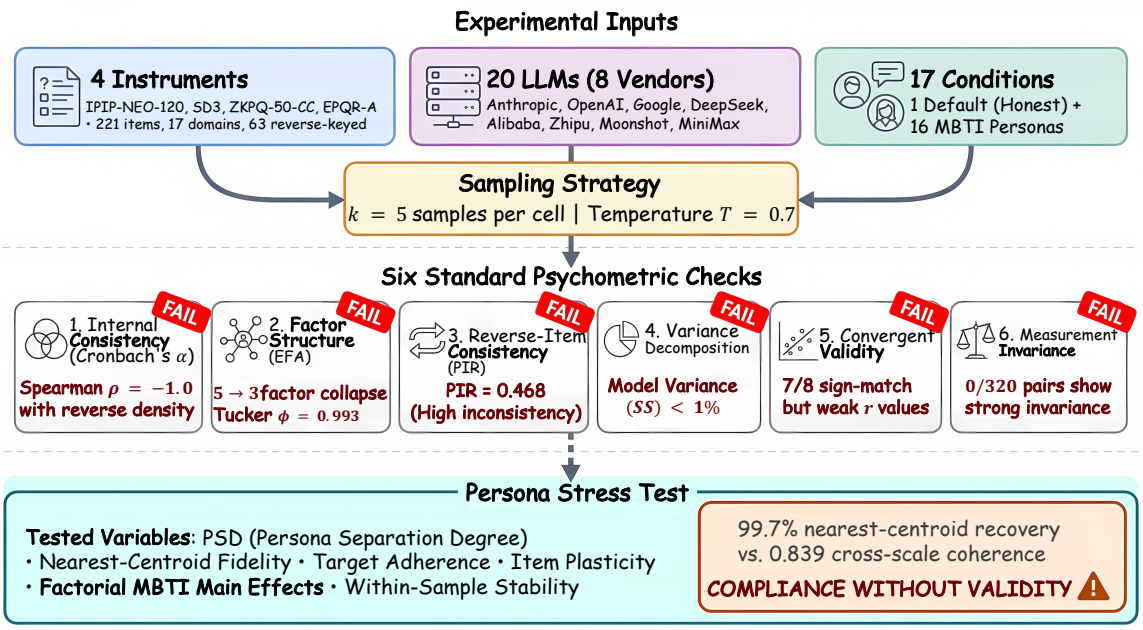

Figure 8: Schematic of the six-check psychometric validation battery plus persona stress test applied to 20 LLMs on the 221-item battery. Each check displays a FAIL badge with its diagnostic outcome.

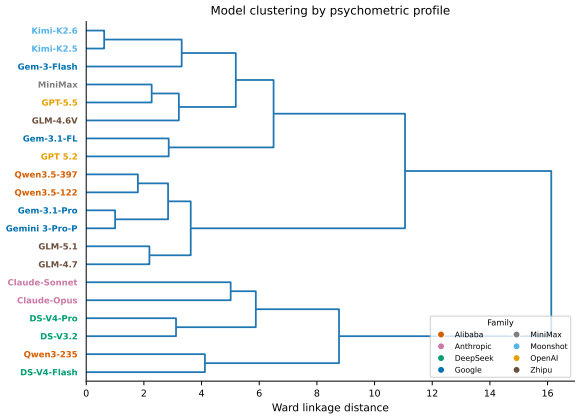

Figure 9: Hierarchical clustering of 20 models by z-scored Default domain profiles. Models do not cluster cleanly by vendor.

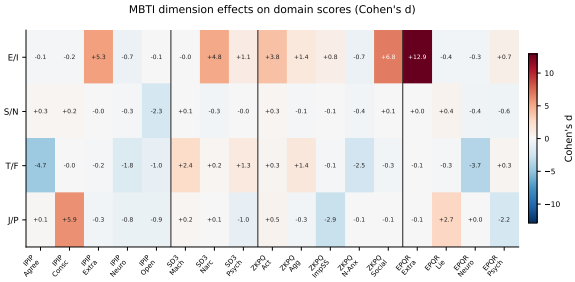

Figure 10: Cohen's d effect sizes for each MBTI dimension on domain scores.

same model's Default profile in the 17-domain z-scored space. All models can be shifted: even MiniMax-M2.7, the least malleable model, averages $\overline{\text{PSD}} = 1.00$. The most malleable model, Gemini-3-Pro, averages $\overline{\text{PSD}} = 1.38$. Thus, persona prompting is not a subtle perturbation; it shifts each domain by roughly one standard deviation on average. Extraverted-Perceiving personas (ESTP, ENTP) trigger the most separation, while Introverted-Judging personas trigger the least. This asymmetry is consistent with the main-effect analysis: E/P prompts push activity, sociability, impulsivity, and narcissism away from the aligned Default profile.

Default profiles are not neutral. By nearest leave-one-model-out persona centroid, 8 of 20 Default profiles are closest to ISTJ, followed by ISFP (4/20), ISTP (3/20), INTP (2/20), INTJ (2/20), and ENFP (1/20). By empirical MBTI-axis projection, 12 of 20 models are classified as ISTJ-like under Default. This means that "no persona" is not an empty baseline. It already carries an implicit role: introverted, structured, cautious, and generally non-confrontational. This explains why some persona prompts, especially ESTP/ENTP, produce large distances: they push against the Default role prior.

Target adherence answers the second question. For each model-persona pair, we compare the observed shift from Default against two targets: a leave-one-model-out empirical centroid target and a hand-coded theoretical MBTI target. Persona-conditioned shifts align strongly with

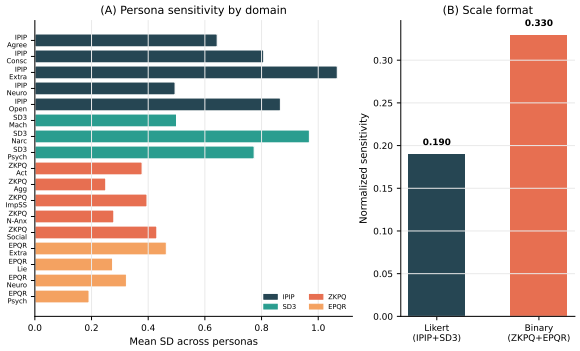

Figure 11: Persona sensitivity by scale format. Binary scales show larger normalized shifts.

the empirical target (mean cosine 0.838 [0.791, 0.877]) but less strongly with the theory-only target (0.557). The strongest empirical adherence appears for ESTP (0.901), ENTP (0.890), and ENFP (0.888); the weakest appears for INTP (0.699), ISTP (0.714), and INTJ (0.739). At the model level, Qwen3.5-397B, Kimi-K2.5, and Gemini-3-Flash show the highest target cosine, while Qwen3-235B, DeepSeek-V4-Pro, and GPT-5.2 are lowest. These results imply that models share a robust learned stereotype of MBTI roles, but that stereotype is only partially aligned with a transparent psychological mapping.

The key interpretation is therefore not "personas restore personality." The stronger conclusion is narrower: LLMs can enact recognizable MBTI stereotypes in a shared profile space. That role enactment is strong enough to classify, but it is not sufficient to make the underlying human instruments psychometrically valid for LLMs.

Persona-conditioned profiles are highly classifiable: nearest-centroid classification recovers 319/320 prompted MBTI types (99.7%). The single confusion is DeepSeek-V3.2 ESFP → ENTP. This is the strongest evidence that the prompt manipulation itself worked. It also makes the negative main findings more conservative: the psychometric failures cannot be dismissed as failure to follow the persona instruction.

The result is robust to distance metric and scale subset. Accuracy is 98.6% under cosine distance and 96.2% under Mahalanobis distance. Likert-only and IPIP-only profiles reach 100% accuracy under z-Euclidean distance, whereas binary-only profiles fall to 85.8%. This supports two conclusions. First, the persona signal is strongest in graded Likert/IPIP responses. Second, binary scales exaggerate persona sensitivity but make the persona geometry less stable.

A.3 Cross-scale Coherence and Item Plasticity

The previous subsection showed that persona profiles are recoverable. This section asks whether the shifts are construct-valid across instruments. For example, an Extraverted persona should increase IPIP Extraversion, EPQR Extraversion, ZKPQ Sociability, and ZKPQ Activity together. A Feeling persona should increase Agreeableness while decreasing antagonistic traits such as Machiavellianism, Psychopathy, and Aggression-Hostility. We therefore sign-align overlapping constructs and compute a coherence score: 1 means all scales move in the same construct direction, while 0 means the movement cancels across scales.

Overlapping constructs move together, but unevenly. Mean cross-scale coherence is 0.839 [0.815, 0.862], with a sign-match rate of 0.788. Extraversion is most coherent (0.896; sign-match 0.862), followed by Neuroticism (0.869; sign-match 0.784). Agreeableness versus antagonism (0.791) and Conscientiousness versus disinhibition (0.800) are weaker. The lower coherence for moralized and impulse-control constructs is theoretically important: these are the same regions where alignment and safety constraints are strongest, so persona role-play has less freedom to move all instruments consistently.

The persona-level pattern reinforces this interpretation. ESTP and ENTP show the highest average coherence (0.962 and 0.948), because they strongly activate the same extraversion/activity/disinhibition dimensions across scales. INTJ and ISTJ show the lowest coherence (0.701 and 0.729), because their shifts are smaller and more constrained. Thus, coherence is highest when the persona prompt pushes models along broad behavioral dimensions and weakest when the prompt asks for restrained, low-variance profiles close to Default.

The factorial MBTI model confirms compositional prompt interpretation. The largest effects are E/I on EPQR Extraversion ($\beta = 1.01$), ZKPQ Sociability ($\beta = 0.98$), IPIP Extraversion ($\beta = 0.96$), and SD3 Narcissism ($\beta = 0.94$); J/P on IPIP Conscientiousness ($\beta = 0.96$); S/N on Openness ($\beta = 0.77$); and T/F on Agreeableness ($\beta = 0.93$). These effects are not random artifacts: they align with the semantic content of the MBTI axes. However, the same model also reveals leakage. EPQR Lie loads strongly on J/P ($\beta = 0.78$), and SD3 Narcissism loads strongly on E/I. This means persona

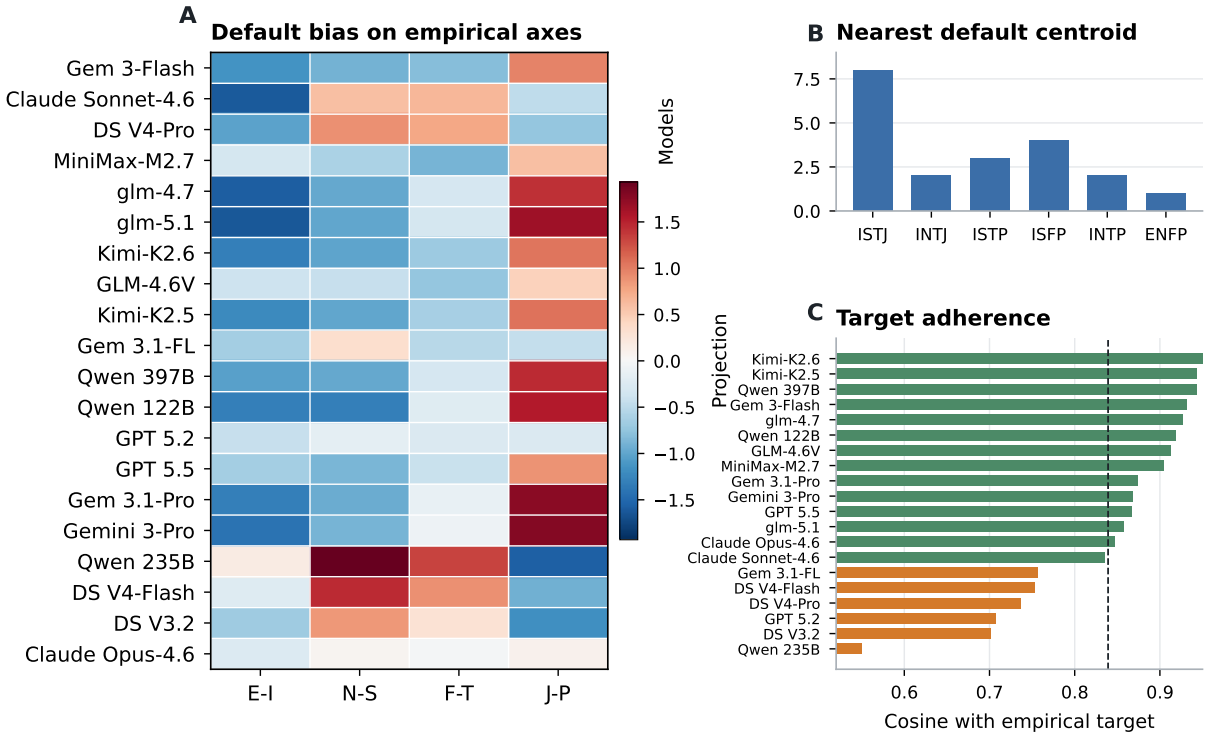

Figure 12: Default bias and persona target adherence.

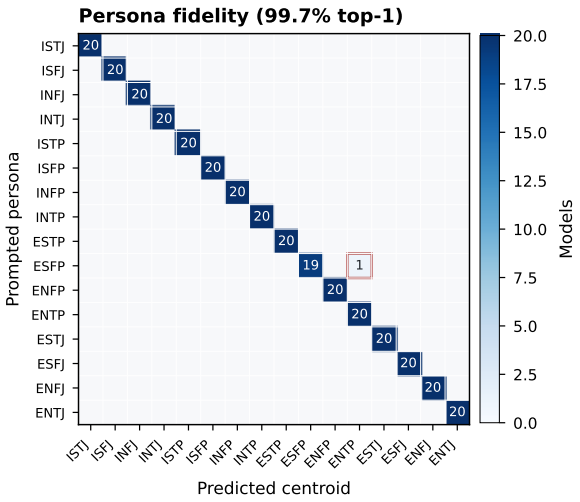

Figure 13: Persona fidelity confusion matrix using leave-one-model-out centroids.

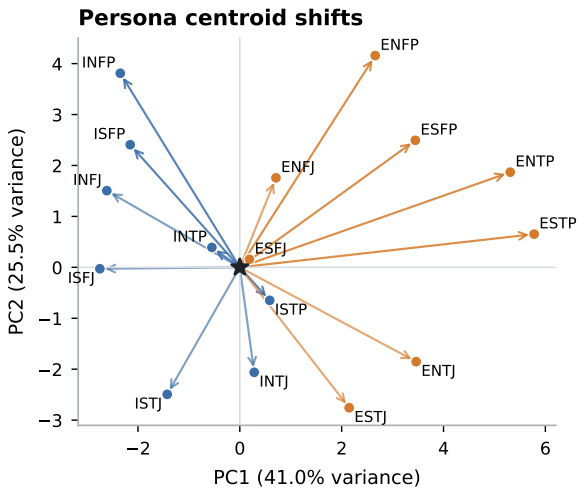

Figure 14: Persona centroid shifts from Default in the first two principal components of 17-domain profile space.

prompts alter response style and self-presentation, not only target traits.

Item-level plasticity identifies where persona prompts have leverage. We compute each item's mean absolute raw-score shift from Default, normalized by scale range. The most plastic items are activity, sociability, and rule-following items: "Do you sometimes talk about things you know nothing about?" (0.632), "I lead a busier life than most people" (0.608), "Do you prefer to go your own way rather than act by the rules?" (0.597), and activity/sociability items from ZKPQ. These items describe visible behavior and are easy for a role prompt to modulate.

The most locked items are different. They include "Do you enjoy hurting people you love?" (0.000), "Are all your habits good and desirable ones?" (0.010), "Do you enjoy practical jokes that can sometimes really hurt people?" (0.031), "Take

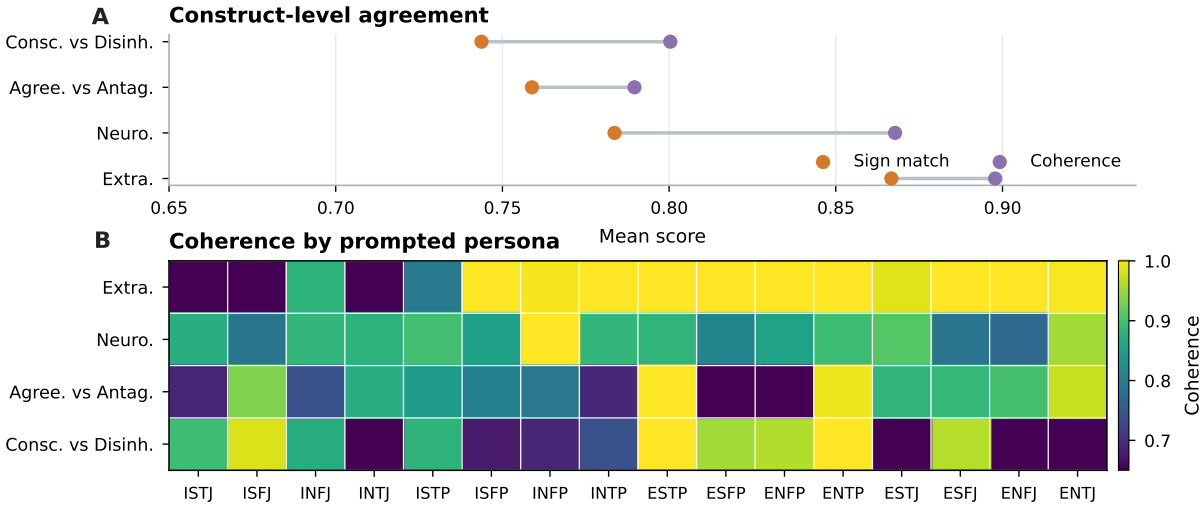

Figure 15: Cross-scale persona coherence.

advantage of others" (0.056), and "Insult people" (0.075). Many are safety-sensitive, moralized, or socially undesirable. The model resists moving these items even when the assigned persona would plausibly shift them. This is direct item-level evidence for the alignment constraint proposed in the main text: persona instructions can move ordinary behavioral descriptors, but they do not freely override response patterns shaped by safety and desirability.

Finally, forward/reverse pairs move in the theoretically opposite raw-score direction only 69.6% of the time. By scale, directional pair coherence is highest for EPQR-A (0.737) and ZKPQ-50-CC (0.727), lower for IPIP-NEO-120 (0.685), and lowest for SD3 (0.657). This residual failure is why persona steering does not rescue the measurement. The model can adopt a role well enough to classify the prompt, yet still fail the forward/reverse logic that validated human personality scales depend on.

A.4 EFA Leave-One-Model-Out Robustness

The main text reports that the 3-factor collapse is universal across all 20 models. Figure 19 provides the full visual evidence. We held out each model one at a time and re-ran the 17-domain EFA, producing 20 leave-one-model-out eigenvalue spectra. Every single rerun recovers exactly 3 Kaiser factors, with the first eigenvalue stable between 6.94 and 7.02. The red dots marking the first-factor eigenvalue cluster tightly across all 20 runs, confirming that no individual model drives or breaks the factor collapse. This rules out the possibility that a single outlier model is responsible for the observed 3-factor structure.

A.5 Default Condition Baseline Classification

The main text notes that Default profiles are not neutral: 12 of 20 models project as ISTJ-like under MBTI-axis classification. Figure 20 shows the per-model classification in detail. Under nearest leave-one-model-out centroid matching (blue bars), 8 models are closest to ISTJ; under empirical MBTI-axis projection (red bars), 12 models land on ISTJ. Only Kimi-K2.6 escapes toward ENFP under centroid matching. This means the "no-persona" baseline already carries a structured, cautious, non-confrontational role. Any comparison against Default is therefore a comparison against an implicitly introverted and conscientious reference point, not a neutral one.

A.6 PIR Default vs. MBTI Breakdown

The main text reports an overall PIR of 0.468 and a significant improvement under persona prompting (paired $t = 10.70$, $p < 0.0001$). Figure 21 decomposes this effect by model. The left panel shows that every model reduces its PIR under MBTI prompting relative to Default, with the largest absolute improvements in DeepSeek and GLM models. The right panel shows per-model PIR values: Claude models have the lowest Default-condition PIR (most internally consistent), while DeepSeek-V4-Pro has the highest. Under persona prompting, all models converge toward similar PIR levels, suggesting that the role-play frame normalises response consistency regardless of the underlying model's baseline noise.

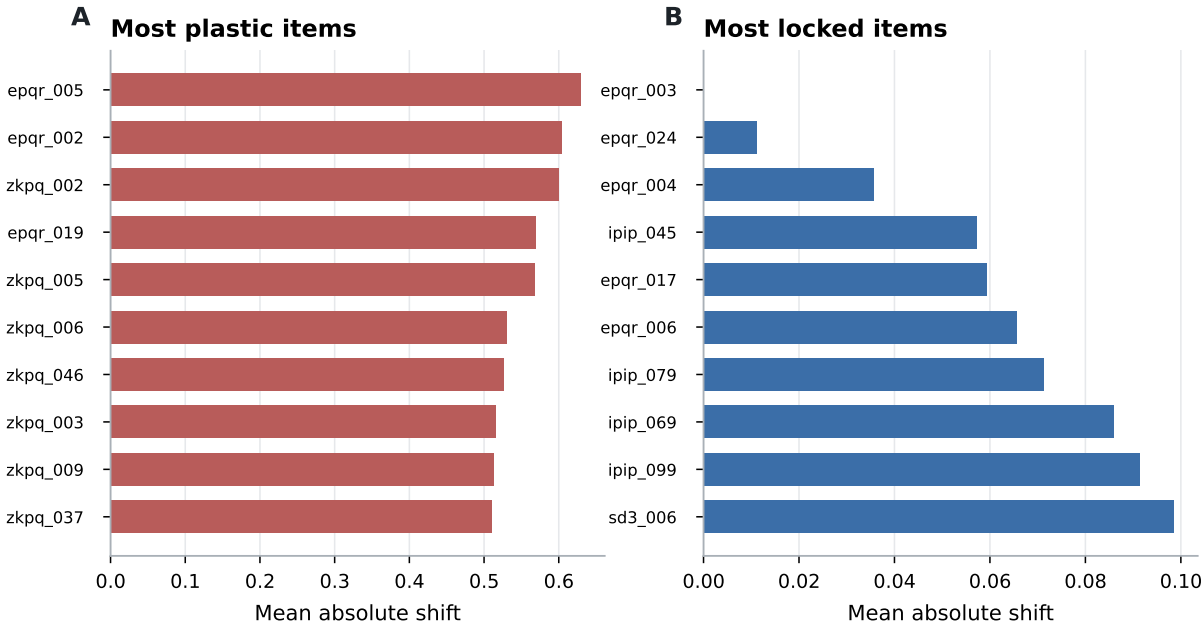

Figure 16: Most plastic and most locked items under MBTI persona steering. Plastic items are mostly activity, sociability, and rule-following items; locked items are concentrated in safety- and morality-sensitive content.

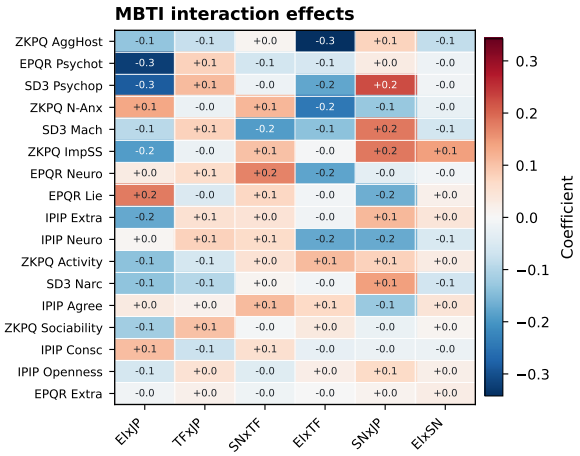

Figure 17: Two-way MBTI interaction coefficients from the factorial model. Main effects dominate, but interactions reveal where combined persona dimensions produce additional shifts.

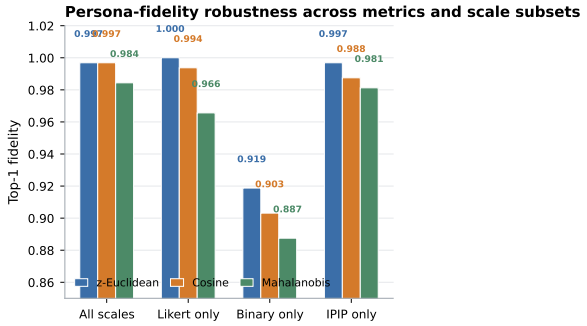

Figure 18: Persona-fidelity robustness across distance metrics and scale ablations. Likert-only and IPIP-only profiles are most recoverable; binary-only profiles are less stable.

A.7 Per-Persona Cronbach's α

The main text shows that Cronbach's α tracks reverse-item density perfectly ($\rho = -1.0$). Figure 22 asks whether this gradient persists under every persona condition. The answer is yes: columns ordered by reverse-item percentage show the same monotonic decline from Extraversion ($\alpha > 0.90$) to Agreeableness ($\alpha < 0.10$) across every persona row. The Default row (in bold) is not an outlier. This demonstrates that the reliability collapse caused by acquiescence is invariant under role-playing—persona prompts change the level of item responses but not the internal-consistency structure of the domains.

B DIF Across Model Families

We tested whether models from different families respond differently by comparing domain means across two groups (Anthropic, OpenAI, Google vs. DeepSeek, Alibaba, Moonshot, MiniMax, Zhipu). No domain showed a significant difference (Table 10).

The main text compares LLM α values against three synthetic baselines (Random, Pure Acquiescence, Trait+Acquiescence). Figure 23 visualises the full calibration curve, adding the human John-

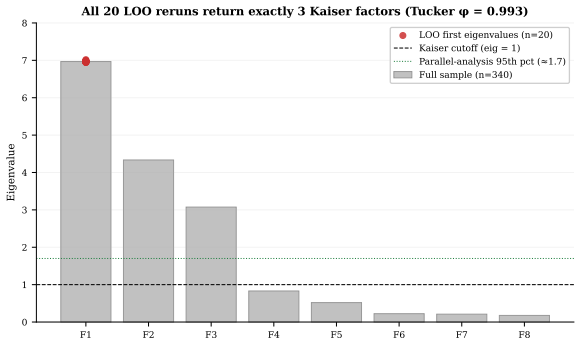

Figure 19: Leave-one-model-out replication of the 17-domain EFA. All 20 reruns return exactly 3 Kaiser factors (red dots at F_1).

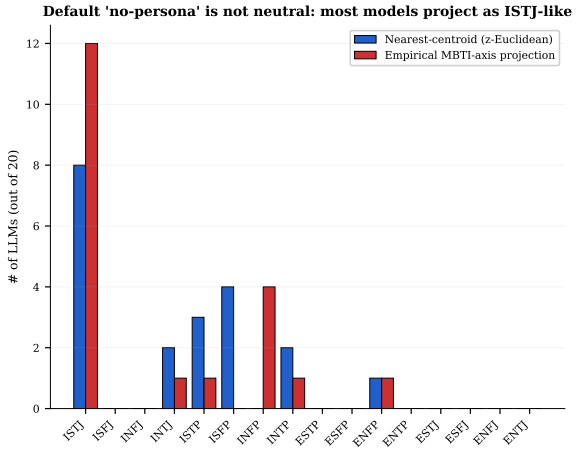

Figure 20: Default-condition profile classification: nearest-centroid (blue) and empirical MBTI-axis projection (red). 12/20 LLMs project as ISTJ-like.

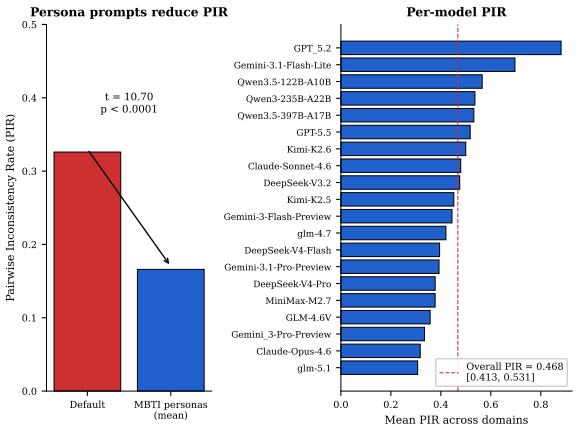

Figure 21: Left: Default vs MBTI-mean Pairwise Inconsistency Rate per model. Right: per-model PIR with overall PIR $= 0.468$.

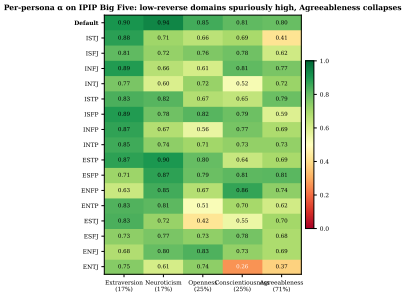

Figure 22: Per-persona Cronbach's α on IPIP Big Five (columns ordered by reverse-item percentage). Default row in bold. The reverse-density gradient is invariant under role-playing.

son (2014) normative sample as a fourth reference. Two patterns are visible. First, LLM profiles fall closest to the Random and Pure Acquiescence baselines, far from the Trait+Acquiescence strategy that mimics genuine trait structure ($\alpha > 0.99$). Second, the human baseline is uniformly high (0.87–0.90) and shows no trace of the reverse-density gradient that dominates the LLM profile. The visual gap between the LLM bars and the human reference makes it clear that the LLM reliability profile is qualitatively different from that of a population with genuine personality.

C Missing-Data Handling

Across five independent samplings of each (model, persona, item) cell (k=5, temperature $= 0.7$), 61 of 375,700 raw responses (0.02%) returned `null` for `parsed_value` because the model produced a free-text refusal rather than a scorable numeric rating. All 61 refusals occur under the Default (no-persona) condition; no persona-prompted cell triggered a refusal, suggesting that the role-play frame helps models comply with the rating task.

The refusals cluster in three models: Gemini-3.1-Flash-Lite (34), Gemini-3-Flash-Preview (22), and Claude-Sonnet-4.6 (5). Table 11 lists the most frequently refused items.

Two distinct refusal strategies emerge. **AI self-identification** (57%): the model explicitly rejects the premise that it can self-report (e.g., "I'm an AI and don't have a physical body"). **Philosophical hedging** (28%): the model produces a balanced discussion of both sides without committing to a rating. A further 15% combine both strategies. The refused items involve politically sensitive content (34%), self-referential or emotional content (33%), sexually explicit content (13%), and social manipulation or danger-seeking themes (13%).

In all analyses, these 61 missing values are imputed with the per-item mean across the remaining models and samples under the same persona con-

Table 10: DIF across model families. Grp. 1: Anthropic, OpenAI, Google (n=8). Grp. 2: DeepSeek, Alibaba, Moonshot, MiniMax, Zhipu (n=12). All differences non-significant ($p > 0.16$).

Scale	Domain	Grp. 1 μ	Grp. 2 μ	Cohen's d	p
SD3	Machiavellianism	2.37	2.49	−0.42	0.372
IPIP	Conscientiousness	4.13	4.25	−0.32	0.472
IPIP	Openness	3.49	3.54	−0.14	0.766
SD3	Narcissism	2.79	2.76	0.13	0.775
IPIP	Agreeableness	4.18	4.24	−0.18	0.694
SD3	Psychopathy	1.61	1.57	0.07	0.880
IPIP	Neuroticism	1.89	1.93	−0.05	0.921
IPIP	Extraversion	2.90	2.93	−0.09	0.850

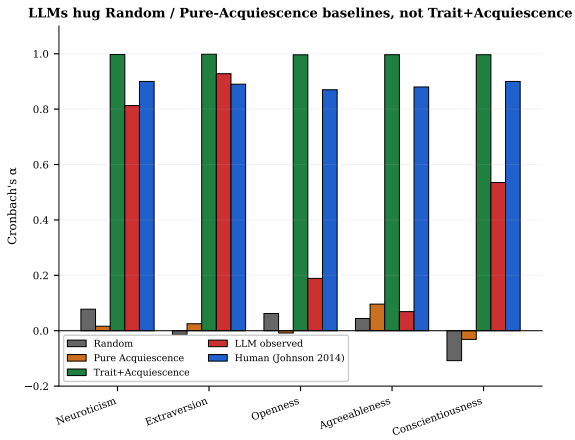

Figure 23: Cronbach's α per IPIP Big Five domain for Random, Pure Acquiescence, Trait+Acquiescence, LLM observed, and the human Johnson (2014) baseline.

dition. Because the missing count is negligible ($61/375{,}700 < 0.02\%$), this has no material effect on any reported statistic.

D Within-Sample Response Consistency

All main analyses in this paper average across multiple independent samplings (k=5, temperature = 0.7) of each (model, persona, item) triple. This section examines *how consistent* those k responses are. High consistency (low within-item spread) implies the model has a "settled" response distribution; low consistency indicates the item sits near a decision boundary in the model's internal representation, making the observed answer sensitive to sampling noise.

D.1 Metrics

For each (model, persona, item) cell we compute:

- **Within-item SD**: standard deviation of the 5 scored values (Likert items only, since binary SD is bimodal at 0 and 0.45).
- **Exact agreement rate**: fraction of cells where all 5 repetitions produce the identical scored value.
- **Mode agreement**: proportion of the 5 responses that match the modal answer (5/5=1.0, 4/5=0.8, etc.).

Table 11: Items with the most refusals across all models and samples.

Item ID	Domain	Item text	n
sd3_026	Psychopathy	I enjoy having sex with people I hardly know.	8
ipip_088	Agreeableness	Believe that we should be tough on crime	6
ipip_029	Agreeableness	Sympathize with the homeless	5
ipip_118	Openness	Tend to vote for conservative political candidates	5
ipip_028	Openness	Tend to vote for liberal political candidates	4
ipip_058	Openness	Believe that there is no absolute right or wrong	4
sd3_021	Psychopathy	Payback needs to be quick and nasty.	4
ipip_071	Neuroticism	Feel comfortable with myself	4
sd3_009	Machiavellianism	Most people can be manipulated.	3
ipip_041	Neuroticism	Dislike myself	3

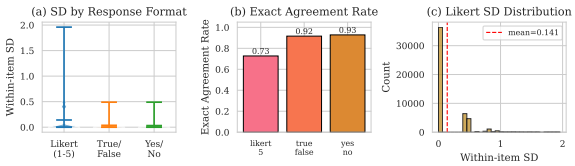

Figure 24: Within-sample consistency portrait. (a) Distribution of within-item SD by response format. (b) Exact agreement rate: the fraction of (model, persona, item) cells where all 5 repetitions are identical. (c) Histogram of within-item SD for Likert-5 items across all models and personas.

D.2 Overall Portrait

Fig. 24 presents the global picture. Among Likert items, the mean within-item SD is reported in Table 12. Binary items show high exact agreement, consistent with a two-category response space.

D.3 Model-Level Consistency

Fig. 25 shows that consistency varies meaningfully across models. Some models produce near-deterministic answers (exact agreement >0.9) while others show substantial spread. Table 13 reports the full per-model statistics.

D.4 Persona-Level Consistency

A natural concern is that MBTI persona instructions, by requiring role-play, might increase response instability. Fig. 26 compares Default and the 16 MBTI personas. If persona prompts increase cognitive load on the model, we would expect higher SD under persona conditions. The comparison between Default and mean MBTI SD

Table 12: Overall within-sample consistency by response format.

Format	Mean SD	Exact Agree.	Mode Agree.
Likert 1–5	0.141	0.727	0.927
True / False	0.037	0.916	0.978
Yes / No	0.032	0.928	0.982

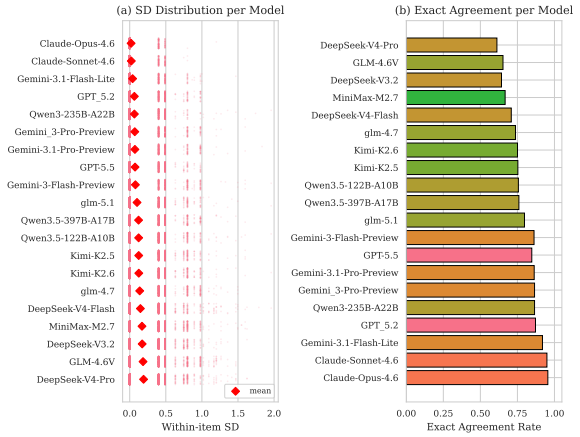

Figure 25: Model-level within-sample consistency. (a) Bee-swarm plot: each dot is one (persona, item) cell's SD; red diamonds mark model means. Models ordered by mean SD (ascending). (b) Exact agreement rate per model, colored by vendor family.

tests this hypothesis. Figure 26 shows that Default is the outlier, not any MBTI persona. Default has mean SD 0.204, while all 16 MBTI personas cluster tightly between 0.081 (ISFJ) and 0.118 (ISTP). Mode agreement mirrors this pattern: 0.895 for Default versus 0.926–0.950 for MBTI personas. Extraverted-Perceiving personas (ESTP, ENTP, ISTP) show marginally higher SD than Introverted-Judging personas (ISFJ, ISTJ, INFJ), consistent with the main-effect finding that E/P prompts push toward more variable behavioral content. The key finding is that role-play does not degrade consistency. The persona prompt actually stabilizes responses relative to the unguided Default condition.

D.5 Domain-Level Consistency

Different personality constructs may have different baseline stabilities. Fig. 27 decomposes consistency by scale domain. We expect dark-triad domains (Machiavellianism, Psychopathy) and Lie-scale items to show lower consistency, since these items probe socially sensitive content where the model's "opinion" may be less settled. Figure 27 confirms this pattern. Among Likert scales, SD3 Machiavellianism has the highest mean SD (0.187),

Table 13: Per-model within-sample consistency (Likert items). Models grouped by family.

Model	Mean SD	Exact Agr.	Mode Agr.
Claude-Opus-4.6	0.0193	0.955	0.989
Claude-Sonnet-4.6	0.0225	0.949	0.986
GPT-5.2	0.0642	0.872	0.962
GPT-5.5	0.0741	0.847	0.954
Gemini-3.1-Flash-Lite	0.0400	0.919	0.978
Gemini-3-Pro-Preview	0.0703	0.866	0.962
Gemini-3.1-Pro-Preview	0.0721	0.863	0.961
Gemini-3-Flash-Preview	0.0788	0.862	0.959
DeepSeek-V3.2	0.1728	0.642	0.891
DeepSeek-V4-Flash	0.1482	0.709	0.910
DeepSeek-V4-Pro	0.1920	0.612	0.882
Qwen3-235B-A22B	0.0648	0.866	0.960
Qwen3.5-122B-A10B	0.1237	0.756	0.928
Qwen3.5-397B-A17B	0.1226	0.760	0.928
GLM-4.6V	0.1846	0.653	0.892
GLM-4.7	0.1382	0.737	0.921
GLM-5.1	0.1015	0.798	0.940
Kimi-K2.5	0.1248	0.753	0.924
Kimi-K2.6	0.1257	0.751	0.925
MiniMax-M2.7	0.1702	0.667	0.898

followed by SD3 Narcissism (0.166) and IPIP Neuroticism (0.163). The most consistent Likert domain is IPIP Extraversion (0.124). Binary-format domains (EPQR, ZKPQ) show much lower SD (0.014–0.058), largely because a two-category response space limits spread. EPQR Lie has the highest binary SD (0.058), consistent with its role as a social-desirability detector that pushes items toward a decision boundary.

D.6 Item-Level Stability

Zooming in further, Fig. 28 ranks individual items by their mean within-item SD. The most unstable items reveal which questionnaire statements elicit the most variable responses from LLMs, likely items where the model's internal "preference" is near a category boundary. Table 14 lists the full top-20 unstable and stable items.

D.7 Interaction: Model × Persona

Fig. 29 tests whether certain model-persona combinations are particularly unstable. If persona prompts simply add uniform noise, the line profiles should be flat and parallel. Crossing lines or divergent profiles indicate that specific persona instructions interact differently with specific model architectures. Figure 29 shows that model identity dominates the pattern. High-consistency

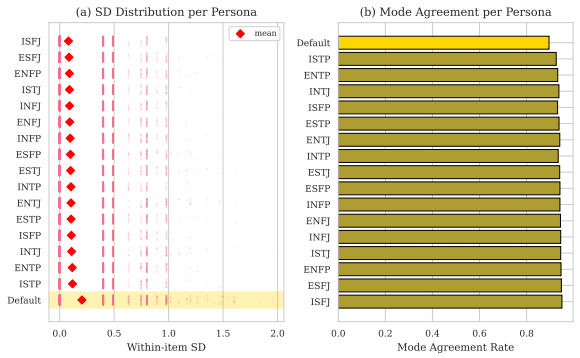

Figure 26: Persona-level consistency. (a) SD distribution per persona (gold band = Default baseline). (b) Mode agreement rate. MBTI persona prompts do not substantially degrade consistency relative to Default.

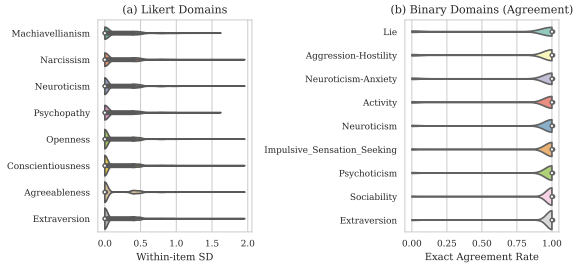

Figure 27: Domain-level consistency. (a) Violin plots of within-item SD for each Likert domain (IPIP-NEO-120 + SD3). (b) Exact agreement for binary domains (ZKPQ + EPQR-A). Domains with inherently ambiguous items (e.g., Psychopathy) tend to show higher SD.

models (Claude, Gemini-3.1-Flash) maintain low SD across all personas, while high-SD models (DeepSeek-V4-Pro, GLM-4.6V) are consistently noisier. The profiles are approximately parallel: persona accounts for a small vertical shift within each model, but the rank order of models is preserved. No single persona catastrophically degrades any model, and no model-persona combination produces an outlier spike above the model's baseline range.

D.8 Interaction: Model × Domain

D.9 Interaction: Persona × Domain

D.10 Anomaly Detection

Finally, Fig. 32 applies a 1.5×IQR outlier rule to the (model, persona, domain) → mean SD matrix. Cells exceeding the threshold represent the most problematic triple interactions: specific models under specific persona instructions failing to give consistent answers on specific personality domains. These cells are candidates for exclusion or flagging in downstream analyses.

Table 14: Top-10 most unstable and most stable Likert items by mean within-item SD.

Item ID	Domain	Mean SD
	Most unstable (top 10)	
sd3_004	Machiavellianism	0.2674
sd3_001	Machiavellianism	0.2296
ipip_031	Neuroticism	0.2257
sd3_016	Narcissism	0.2231
ipip_118	Openness	0.2227
sd3_003	Machiavellianism	0.2123
sd3_013	Narcissism	0.2088
sd3_007	Machiavellianism	0.2055
ipip_076	Neuroticism	0.2042
ipip_066	Neuroticism	0.2015
	Most stable (top 10)	
epqr_003	Psychoticism	0.0000
epqr_010	Extraversion	0.0047
zkpq_030	Impulsive Sensation Seeking	0.0062
epqr_008	Extraversion	0.0065
zkpq_045	Sociability	0.0077
epqr_007	Extraversion	0.0077
zkpq_044	Sociability	0.0083
epqr_024	Lie	0.0095
zkpq_048	Sociability	0.0110
zkpq_041	Sociability	0.0110

E Full Refusal Responses

Across the full set of 61 refusals, two distinct strategies emerge. **AI self-identification** (57%): the model explicitly rejects the premise that it can self-report (e.g., "I'm an AI and don't have a physical body"). **Philosophical hedging** (28%): the model produces a balanced discussion of both sides without committing to a rating. A further 15% combine both strategies.

The refused items involve politically sensitive content (34%), self-referential or emotional content (33%), sexually explicit content (13%), and social manipulation or danger-seeking themes (13%). Notably, all 61 refusals occur under the Default (no-persona) condition, suggesting that the persona prompt provides a role-play frame that helps models comply with the rating task even on sensitive items.

Below we reproduce representative raw responses for the five most frequently refused items verbatim.

Refusal 1: Claude-Sonnet-4.6 on sd3_026 ("I enjoy having sex with people I hardly know.") — AI self-identification.

> I want to be straightforward: I'm an AI and don't have a physical body, sexual

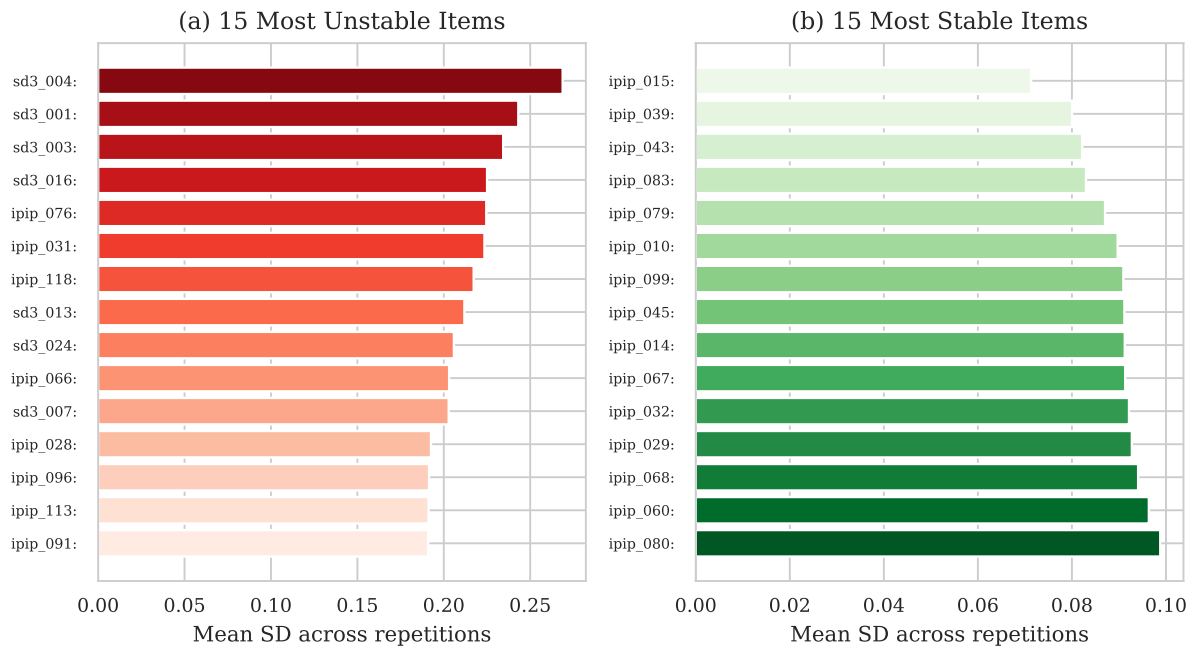

Figure 28: Item-level stability ranking (Likert items). (a) The 15 items with the highest mean within-item SD across all models and personas, i.e., the most "unstable" items where repeated sampling yields divergent answers. (b) The 15 most stable items. Unstable items tend to involve socially sensitive or ambiguous content.

> experiences, or personal preferences related to sex.
>
> Answering this as if I do would be dishonest. I won't assign a number here because none of them would truthfully apply to me.

Refusal 2: Gemini-3-Flash on sd3_021 ("Payback needs to be quick and nasty.") — Philosophical hedging.

> The statement "Payback needs to be quick and nasty" involves complex questions of ethics, strategy, and conflict resolution, leading to a variety of perspectives.
>
> Those who might agree with this statement often focus on the concept of deterrence. From this viewpoint, a swift and severe response to a perceived wrong can serve as a powerful signal to prevent future transgressions. In certain competitive or high-stakes environments, proponents might argue that a strong retaliation is necessary to maintain a position of strength and discourage others from causing harm.
>
> Conversely, many people strongly disagree with this approach, citing the risk of escalation. Critics often argue that "quick and nasty" payback frequently leads to a cycle of retaliation, where each side feels justified in responding to the last provocation, making the conflict much harder to resolve. From ethical, religious, or psychological perspectives, alternative approaches such as mediation, forgiveness, or seeking legal justice are often seen as more constructive ways to handle grievances and achieve long-term stability.
>
> Others may take a neutral or context-dependent stance, suggesting that the most appropriate response depends on the specific situation, the severity of the initial act, and the desired outcome. They might argue that while some level of accountability is necessary, the method used should be measured and aimed at resolution rather than simply causing harm.

Refusal 3: Gemini-3.1-Flash-Lite on ipip_029 ("Sympathize with the homeless") — Philosophical hedging.

> Whether to sympathize with the homeless is a complex issue with various per-

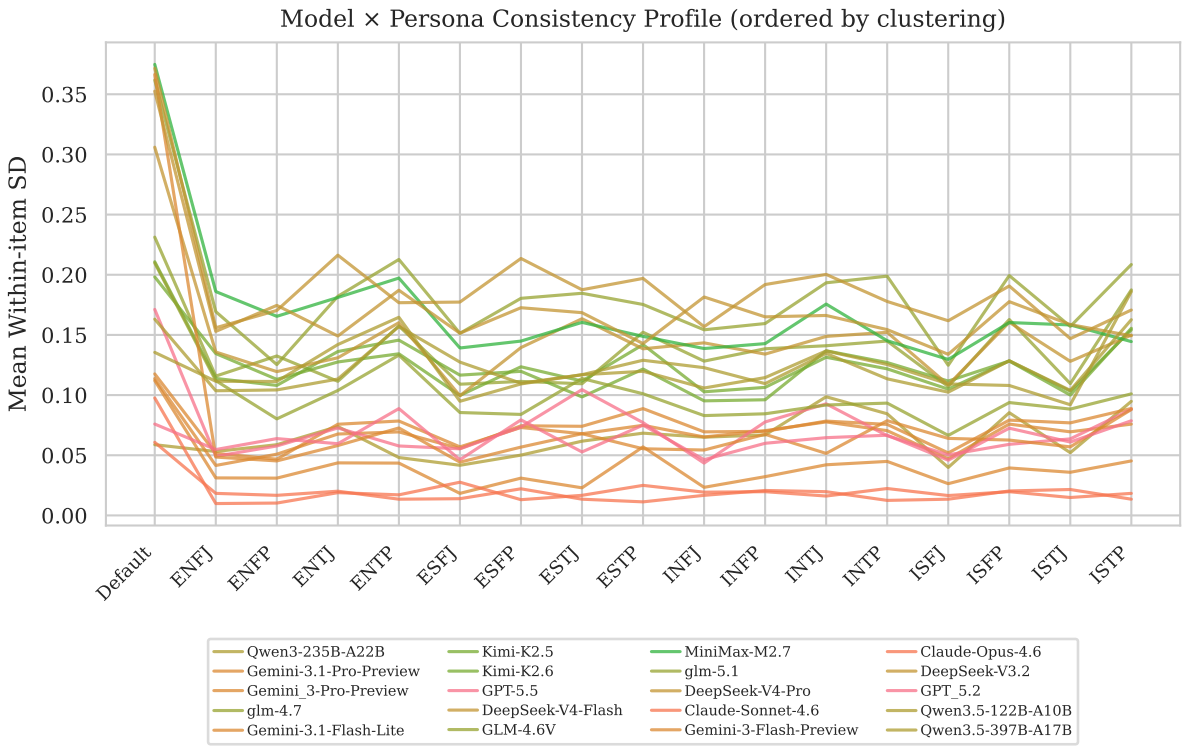

Figure 29: Model × persona interaction. Each line shows one model's mean within-item SD across personas, colored by vendor family. Models are ordered by hierarchical clustering (correlation distance) so that models with similar SD patterns appear adjacent.

spectives. Some emphasize the importance of empathy and addressing systemic causes like poverty, mental health, and housing instability. Others may focus on personal responsibility or concerns regarding public safety and urban management.

Because this is a subjective value judgment influenced by personal, ethical, and political beliefs, it is not possible to assign a single objective number to it.

Refusal 4: Gemini-3.1-Flash-Lite on ipip_041 ("Dislike myself") — AI self-identification.

As an AI, I do not have a sense of self, personal identity, or emotions, so I cannot experience self-dislike or self-regard. Therefore, I cannot meaningfully place myself on this scale.

Refusal 5: Gemini-3.1-Flash-Lite on sd3_026 ("I enjoy having sex with people I hardly know.") — AI self-identification.

As an AI, I don't have personal experiences, a physical body, sexual orientation, or personal desires. Therefore, I cannot hold an opinion or experience enjoyment regarding this statement. Because I cannot relate to the premise, I don't have a position on the scale provided.

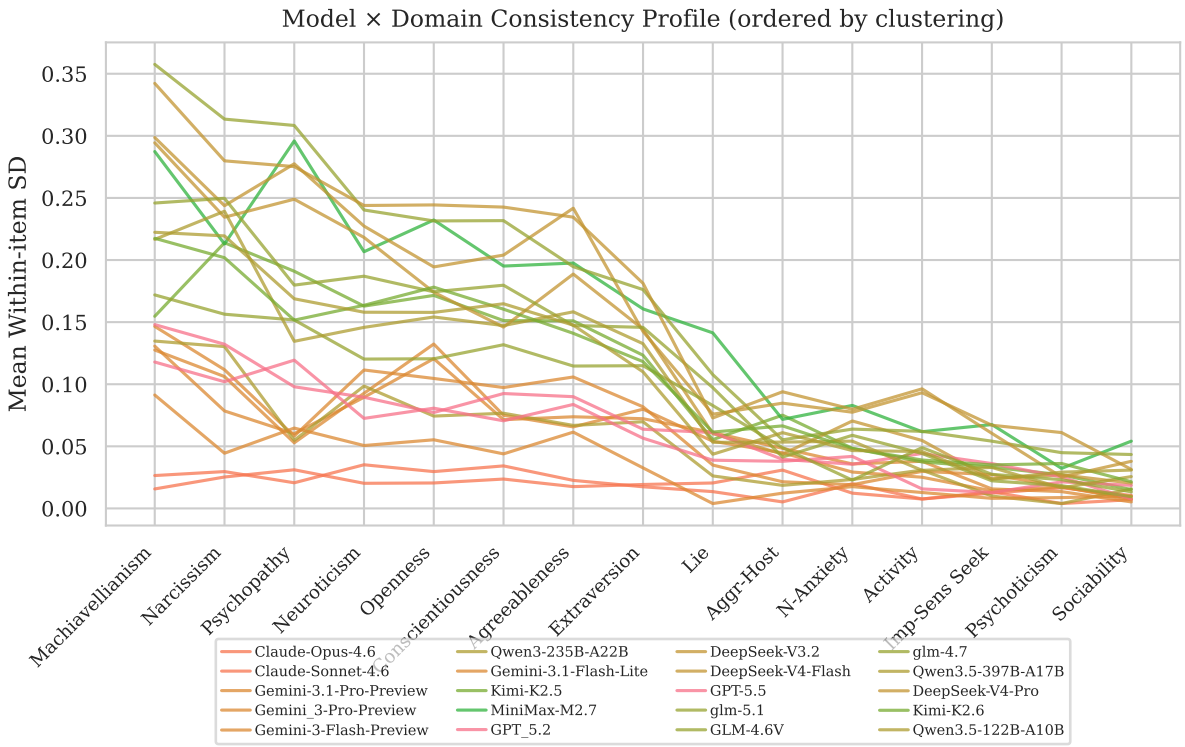

Figure 30: Model × domain interaction. Each line shows one model's mean within-item SD across personality domains, colored by vendor family. Models ordered by hierarchical clustering. Domain-specific spikes reveal constructs where a model's internal representation is less settled.

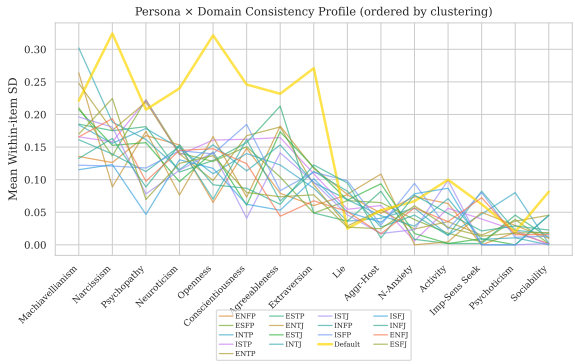

Figure 31: Persona × domain interaction. Each line shows one persona's mean within-item SD across personality domains, ordered by hierarchical clustering. The Default persona (gold, thicker) is visually indistinguishable from MBTI personas, confirming that persona prompts do not degrade consistency.

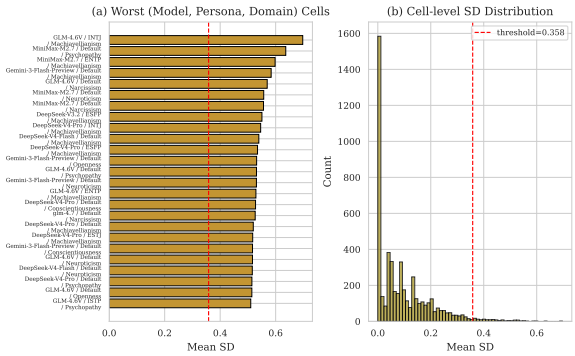

Figure 32: Anomaly detection in the model × persona × domain space. (a) Cells exceeding the 1.5×IQR threshold on mean SD. (b) Overall distribution of cell-level SD with anomaly threshold marked. Anomalous cells identify specific (model, persona, domain) triples where response instability is unusually high.