Personas in Pieces: Factorial Prompt Geometry and Item-Indexed Response Heterogeneity in Language Models

Anonymous authors
Paper under double-blind review

Abstract

Persona prompting creates reliable, recognizable behavior. We ask where that reliability lives: in a compositional prompt representation, in an item-specific response policy, or in a stable latent trait. We evaluate 20 models from eight families on 221 items from four personality inventories, 16 MBTI prompts, a default condition, and five samples per cell (375,700 responses). The prompt code and item-indexed policy are empirically separable; the experiment does not identify a stable-trait interpretation of the resulting scores. At the prompt level, a decoder trained on seven model families recovers the MBTI labels of the held-out eighth at 99.7% accuracy (99.1% with balanced forward/reverse weighting that retains reverse scoring), and 90.8% of the variation across persona profiles lies in the four single-letter contrasts. At the item level, the same persona shifts different items by different amounts: persona–item interactions account for 48.7% of Likert response variation (52.0% binary) against 0.36%/0.51% for model identity—far above sampling noise (F=1,045/649) and concentrated on four directions. Under a specified PC1/k-nearest-neighbor null, a one-dimensional coordinate absorbs 68% of the Likert interaction and 81% of the binary interaction; four letter-aligned coordinates absorb 86–101%, while random four-dimensional bases absorb only 34–44%. Aggregation then hides this heterogeneity: equal-key scoring shifts scores by 0.041 on average and reverses 10.2% of untied model rankings. A replication and ablation suite on two family-external models (Qwen3.8-27B and Mistral Small 3.2; 58,444 responses each) reproduces both levels, survives a template ablation that deletes all MBTI letters and type names (first-order energy 0.966/0.955), and persists when every item is paraphrased (domain-profile correlation 0.970/0.959; item-level effects 0.83/0.77); it also generalizes beyond MBTI letters and Likert scales, to factorial Big-Five prompts, unstructured role personas, and forced-binary formats. Continuation probes found no large verbatim contribution under the tested probes. The result is a constructive boundary: the prompt geometry transfers, while its item-level expression depends on the instrument. We measure questionnaire response channels; no external behavioral or criterion-validity claim is made.

1 Introduction and Related Work

A persona instruction can change a model's behavior reproducibly, but a questionnaire score does not reveal where that reliability lives. We distinguish three competing explanations: a transferable and compositional prompt code; an item-indexed response policy that translates that code differently for different questions; and a stable latent trait whose score should be invariant to the instrument. The paper tests these explanations on a complete model×persona×item design rather than treating one aggregate score as a trait.

Human inventories assume stable item polarity and score meaning, while response-style and item-response models separate content from agreement tendencies (Johnson, 2014; Jones & Paulhus, 2014; Aluja et al., 2006; Francis et al., 1992; Couch & Keniston, 1960; Paulhus, 1991; Billiet & McClendon, 2000; Brown & Maydeu-Olivares, 2011). Recent LLM work studies

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

psychometric validity, persona axes, steerability, sycophancy, and dissociations between self-report and behavior (Sühr et al., 2025; Serapio-García et al., 2025; Lu et al., 2026; Plisiecki et al., 2026; Vennemeyer et al., 2025; Han et al., 2025). Complementary 2026 studies report directional response bias and item-selection dependence across dozens of models (Meyer et al., 2026), content-validity failures in human Big-Five items (Zierahn et al., 2026), and instability under scale, reasoning, ordering, paraphrasing, and conversation history (Tosato et al., 2026). We connect these threads at item level: prompt geometry is tested for transfer and composition, response policy for item interactions and score distortion, and the trait interpretation against matched null simulations.

We make five contributions. We provide a held-family-out test of transferable persona information; quantify compositionality with an exact Walsh expansion; measure contextual expression with an orthogonal response-tensor decomposition; show how aggregation turns policy heterogeneity into key-sensitive scores and rank changes; and replicate both levels on two family-external models with template, wording, memorization, taxonomy, and response-format controls. The result is a boundary: prompt geometry transfers, while its item-level expression belongs jointly to the persona and the instrument.

2 Experimental Design

Table 1: Data and design. The primary analysis uses all inventories; IPIP supplies the largest Likert battery and SD3, ZKPQ, and EPQR-A provide validation across formats and domains.

Design	20 models (8 families) × (Default + 16 MBTI personas) × 221 items
Battery	4 published inventories, 17 domains; 147 Likert, 74 binary
Sampling	Temperature 0.7; 5 samples/cell; 375,700 responses (63 missing)
Replication	Qwen3.8-27B and Mistral Small 3.2; 58,444 responses/model
Estimands	Prompt code, persona×item policy, score distortion

The battery contains IPIP-NEO-120 (120 Likert items), SD3 (27 Likert items), ZKPQ-50-CC (50 binary items), and EPQR-A (24 binary items). Likert responses are normalized by $(v-1)/4$ and binary responses remain in $\{0,1\}$; reverse-keyed items become $1-y$. The common prompt uses a Strongly Disagree–Strongly Agree frame for Likert inventories, so this is a response-format adaptation rather than a claim to reproduce the original human administration or validity evidence. Cell means average the available samples without imputation. The single incomplete item–condition cell ('sd3_026') excludes that item from all-condition ANOVA, leaving 146 Likert and 74 binary items; MBTI-only ANOVA retains all 147 Likert items. Scores use all available observations.

The prompt code is represented by the factorial model

$$z_{mp} = \mu_m + \sum_{k=1}^{4} \beta_{mk} h_{pk} + \epsilon_{mp}, \qquad h_{pk} \in \{-1, +1\}, \tag{1}$$

where h_{pk} are the four MBTI contrasts and z_{mp} is a domain profile; μ_m, β_{mk}, and ϵ_{mp} are vectors in that space. This approximation concerns the 16 factorial prompts. Item responses, including Default, use the saturated orthogonal decomposition

$$y_{mpi} = \mu + a_m + b_p + c_i + (ab)_{mp} + (ac)_{mi} + (bc)_{pi} + (abc)_{mpi}, \tag{2}$$

with sum-to-zero projections on the balanced grid. The three-way term is remaining cell variation, not an independently estimated stochastic residual. For a mixed-key domain, let F, R be raw forward/reverse means and $N = n_+ + n_-$. Published scoring is $S = [n_+F + n_-(1-R)]/N$; equal-key scoring is $B = [F + (1-R)]/2$, which retains reverse scoring but gives each direction half the weight. For complete profiles,

$$S - B = \frac{n_+ - n_-}{n_+ + n_-} \frac{F - (1-R)}{2}. \tag{3}$$

Transfer without test-family leakage. For held-out family f, let $c_{p'}^{(-f)}$ and D_{-f} be persona centroids and the diagonal matrix of domain standard deviations, both estimated only from

other families. The prediction is

$$\hat{p}_{mp} = \arg\min_{p'} \left\| D_{-f}^{-1}\left(z_{mp} - c_{p'}^{(-f)}\right) \right\|_2^2. \tag{4}$$

All 320 MBTI profiles are evaluated out of family; a separate split leaves out one model. Decoder intervals resample eight family blocks 10,000 times, while interaction-share intervals resample items with all model/persona cells together. These intervals describe different sampling units, not independent responses.

Exact projection, not causal variance attribution. Let A_a denote each term of Equation 2, broadcast to the complete tensor shape. Balanced averaging and centering are orthogonal projections, so

$$\langle A_a, A_b \rangle_F = 0 \ (a \neq b), \qquad r_a = \frac{\|A_a\|_F^2}{\|Y - \bar{Y}\|_F^2}, \qquad \sum_a r_a = 1. \tag{5}$$

Each centered factor sums to zero; its inner product with a term constant on that factor vanishes. Applying this argument across factors gives the exact reconstruction and sum of squares. Shares quantify the sampled response channel; their magnitudes depend on the response representation and factor degrees of freedom. Full bootstrap, DIF, and null protocols are in the Appendix.

3 Results

3.1 Transferable prompt code

The held-family-out centroid decoder recovers 319/320 MBTI labels (99.7%) when trained on seven families, and 317/320 (99.1%) after balanced equal-key reweighting that retains reverse scoring. Model-held-out decoding is 99.4%; forward-only and reverse-only features remain 98.8% and 94.1%. The same profiles concentrate 90.8% of centered Walsh energy in the four first-order letter contrasts, with 7.3% pairwise, 1.6% three-way, and 0.3% four-way energy. These values are summarized in Table 2 and visualized in Figure 1; the geometry survives balanced direction weights because prompt identity remains recoverable.

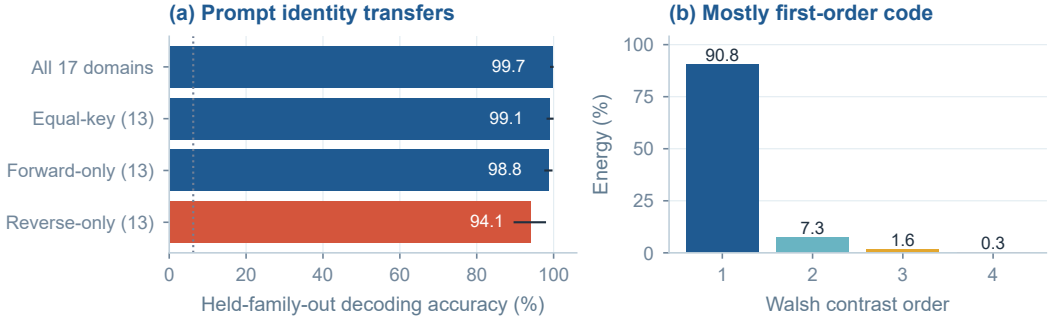


Figure 1: Prompt code transfer and compositionality. Held-family-out decoding remains near ceiling under equal-key, forward-only, and reverse-only features (left); centered MBTI prompt energy is dominated by first-order Walsh contrasts (right). Whiskers are family-level bootstrap 95% CIs and the dotted line is chance (6.25%).

The Walsh calculation is an exact change of basis rather than a regression approximation, consistent with graded compositionality measures in representation learning (Andreas et al., 2019). Let $H \in \{-1, +1\}^{16\times 15}$ contain the non-constant products of the four axis indicators, with $H^\top H = 16I$. For a centered profile $u_{mp} = z_{mp} - \bar{z}_m$, the coefficient of contrast s is

$$\widehat{u}_{ms} = \frac{1}{16}\sum_{p=1}^{16} H_{ps} u_{mp}, \qquad u_{mp} = \sum_{s=1}^{15} H_{ps}\widehat{u}_{ms}. \tag{6}$$

Parseval's identity gives $\sum_p \|u_{mp}\|_2^2 = 16\sum_s \|\widehat{u}_{ms}\|_2^2$. Thus the order-$k$ values in Table 2 are fractions of an exactly conserved energy, not fractions of explained variance from a fitted

Table 2: Prompt compositionality. Accuracy is held-family-out unless noted; Walsh energy is pooled over domains within each model.

Analysis	Estimate	95% CI / unit
All-domain decoder	99.7%	[99.1,100.0]
Balanced equal-key decoder	99.1%	[98.2,100.0]
Model-held-out decoder	99.4%	[98.5,100.0]
First-order Walsh energy	90.8%	fraction
TRE_1	9.2%	fraction
Higher-order energy (2nd/3rd/4th)	7.3/1.6/0.3%	fraction
Effective contrast dimension	4.46	out of 15

model. If $\widetilde{u}^{(1)}$ retains only the four single-axis terms, its truncation error is

$$\mathrm{TRE}_1 = \frac{\sum_m \|u_m - \widetilde{u}_m^{(1)}\|_F^2}{\sum_m \|u_m\|_F^2} = 1 - E_1 = 0.092. \tag{7}$$

The entropy-effective dimension $d_{\mathrm{eff}} = \exp[-\sum_s q_s \log q_s]$, with q_s the contrast-energy proportions, is 4.46. It is larger than four because weak pairwise and higher-order terms remain, but much smaller than the 15 available contrasts. Finally, with 16 equally likely labels and decoder error e, Fano's inequality gives $I(P;X) \geq 4 - h_2(e) - e \log_2 15$. The 99.7% family-held-out accuracy therefore implies at least 3.96 bits of label information (3.89 bits after equal-keying), close to the four bits needed to identify a persona. These quantities jointly support a compositional code: the decoder transfers because the same four directions dominate, not merely because all 16 prompts are separable in one archive.

3.2 Geometry and item plasticity

The 16 persona centroids form a structured displacement from the default profile: the first two components explain 56.0% and 20.8% of centered centroid energy. At the item level, the four letter directions vary in magnitude and sign across domains, and default-to-persona shifts range from nearly zero to 0.63 on the normalized scale. Figure 2 places these views together: a low-dimensional prompt geometry is compatible with highly non-uniform item expression.

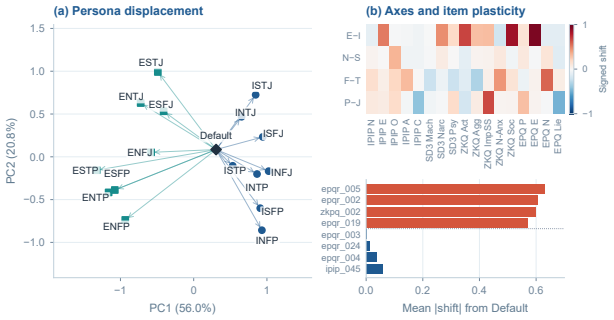


Figure 2: Geometry and item plasticity. (a) Persona centroids in the two leading components, with arrows from Default. (b) Signed axis effects across domains (top) and the four most and four least plastic items (bottom). A common code permits non-uniform item shifts.

3.3 Mechanistic boundary

Figure 3 summarizes the evidence chain. The first panel is the transferable prompt code, the second is the item-dependent response policy, and the third is the scoring boundary created by key imbalance. The robot and questionnaire drawings are schematic: they explain the measured relationships and are not evidence about an internal neural mechanism.

3.4 Item-indexed response policy

Persona×item interactions account for 48.7% of Likert response variation and 52.0% of binary variation, compared with 0.36% and 0.51% for model identity. The contextuality

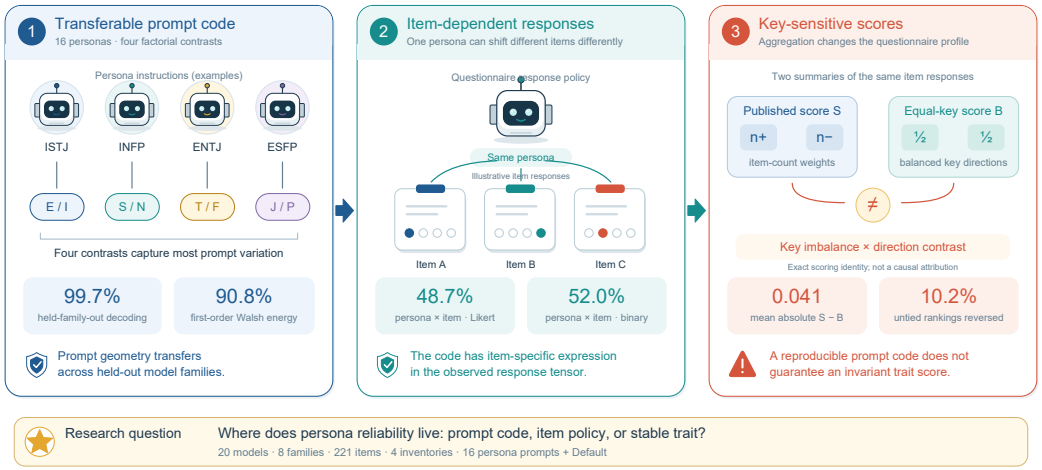


Figure 3: Mechanistic overview of persona reliability. The three panels show (1) transferable prompt code, (2) item-dependent response policy, and (3) key-sensitive aggregation. Reported quantities are the crossed model×persona×item estimates; icons are schematic.

index $C = r_{pi}/(r_{pi} + r_p)$ is 0.922 for Likert and 0.787 for binary. The interaction exceeds the matched noise floor (1.3%/2.6%), has effective rank 3.9, and is concentrated on the same four directions as the prompt code. Figure 4 shows the decomposition and item-level heatmap; full degrees-of-freedom and bootstrap diagnostics are in the Appendix.

The interpretation of the 48.7% share depends on the projection geometry. On the complete Likert grid the persona×item term has 2,320 degrees of freedom, but its squared norm per degree of freedom is only 1.1× the model main effect; for binary it is 1.7×. The large total share is therefore not evidence for thousands of independent persona rules. It is a broad interaction supported by a low effective rank, with $F = 1{,}045$ and 649 against the within-cell noise scale and pure-noise floors of 1.3% and 2.6%. In keyed space, the corresponding persona main effects are 4.1% and 14.1%; in raw response space they are 1.34% and 3.97%, while persona×item shares are 51.1% and 62.7%. Keying amplifies the main effect through agreement direction, but it does not create the item interaction. The key-direction projection spans only 0.76%/0.59% of the interaction and key imbalance correlates negatively with within-domain interaction share ($\rho = -0.30$). This separates three quantities that are easy to conflate: a persona-wide response tendency, item-dependent modulation, and a scoring transformation.

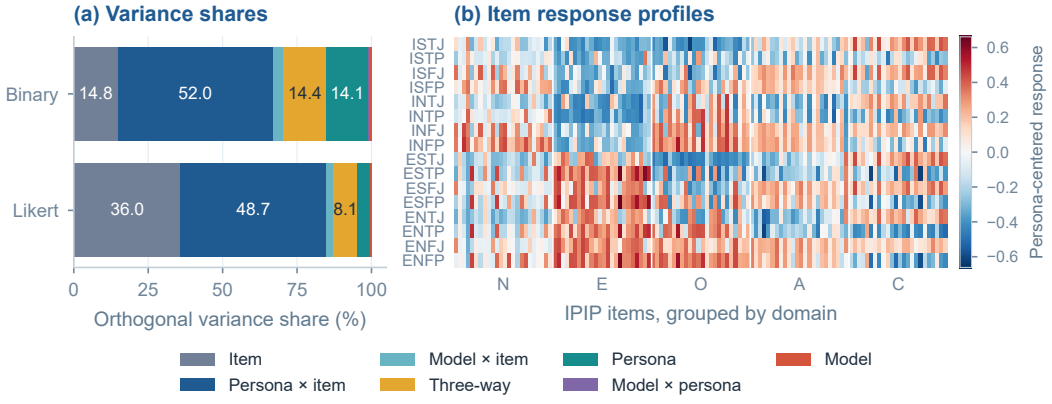


Figure 4: Item-indexed response policy. Orthogonal shares quantify persona×item variation for Likert and binary responses; the heatmap shows representative domain-averaged profiles from IPIP, SD3, and ZKPQ. The model main effect is small, while the interaction is above the matched sampling floor.

3.5 Aggregation and score distortion

Equal-key reweighting shifts mixed-key domain scores by 0.041 on average and reverses 10.2% of untied default-condition model rankings. Equation 3 gives the sharp bound, and the largest observed shift is 0.122 in ZKPQ Neuroticism–Anxiety. Sampling sensitivity is small: decoding remains 99.4–99.7% from one to five samples and score MAE falls from 0.0189 to 0.0074 by three samples. Table 3 and Figure 5 combine the policy shares, key mechanism, rank reversals, and sampling checks. Thus a transferable code and a questionnaire score are different estimands: the latter is a joint function of persona and instrument.

Equation 3 also gives an immediate bound on when score distortion must disappear. If $n_+ = n_-$, or if the raw forward and reverse means satisfy $F = 1 - R$, then $S = B$. Otherwise the shift has the sign of the key imbalance times the contrast between the two response directions. The observed mean shift of 0.041 is consequently not a hidden model effect: it is the product of an instrument's weighting asymmetry and a real directional response contrast. Rank reversals arise when this product is large enough to cross a pairwise score gap. Because the analysis counts only pairs untied under both schemes, the 10.2% figure is conservative with respect to ties. The sampling curve gives a separate inference: three samples reduce score error by 61% relative to one sample while leaving the transfer estimate essentially unchanged, so the key result is not a noisy single-generation artifact.

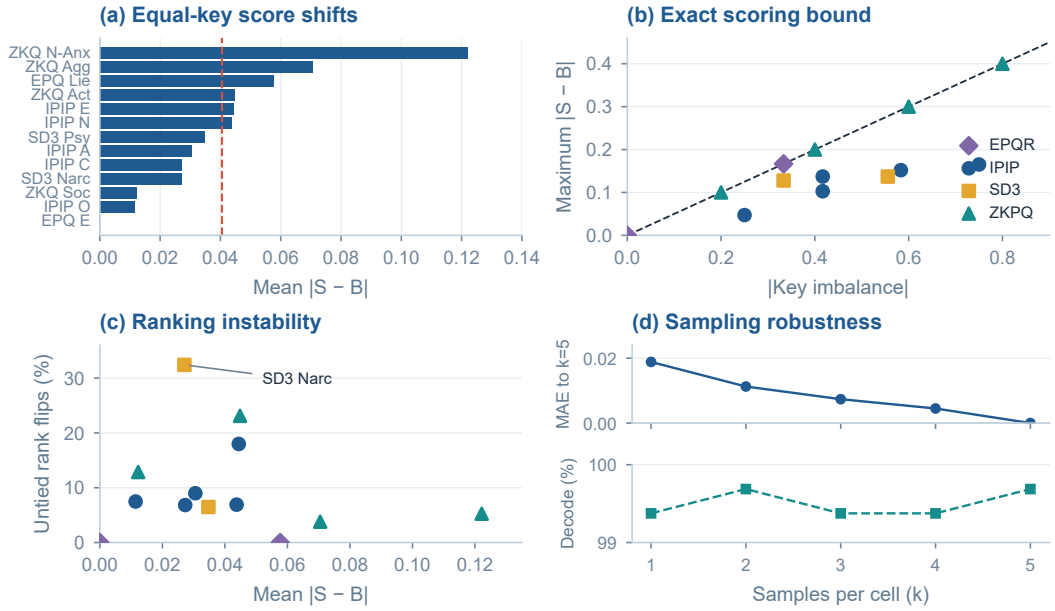


Figure 5: Score distortion and key mechanism. (a) Equal-key score shifts; (b) observed shifts against the algebraic key-imbalance bound; (c) model rank reversals; (d) sampling robustness. The dashed lines mark the cross-domain shift (0.041) and the exact bound.

Table 3: Policy and score robustness. Percentages are orthogonal shares; shifts are on the normalized [0,1] scale.

Quantity	Likert	Binary
Persona×item share	48.73% [44.0,53.3]	51.98% [46.1,56.1]
Model identity share	0.36%	0.51%
Pure-noise floor	1.3%	2.6%
MBTI-only interaction share	50.46%	53.39%
Mean equal-key shift	0.041	–
Untied rank reversals	10.2%	–
Score MAE ($k = 1 \to 3$)	0.0189→0.0074	–

3.6 Cross-inventory transfer

Coarse response statistics are inventory-specific, but the four-letter space transfers. Regressing each domain's persona effects on the letter indicators gives $R^2 \geq 0.79$ for all 17 domains (median 0.91). Leave-one-inventory decoding reaches 99.7% on IPIP, 86.6% on ZKPQ, 80.0% on EPQR-A, and 63.8% on SD3. Matched constructs share loading vectors with cosines 0.989–0.994 for the extraversion triple, 0.84–0.97 for the neuroticism triple, 0.946 for SD3 Psychopathy–ZKPQ Aggression–Hostility, and 0.974 for IPIP Conscientiousness–EPQR Lie (Figure 6). The transferable object is letter space, not a universal item profile.

Formally, for domain d we fit $\delta_d = H\beta_d + \eta_d$, where δ_d is the 16-persona effect vector after removing its mean and H is the same factorial design used in Equation 6. The R^2 values test whether the four axes span each domain's persona displacement; the leave-one-inventory decoder tests whether the resulting profile remains identifiable when all domains from another inventory are removed. The gap between IPIP and SD3 is therefore informative: a short battery carries less redundant information about the same code, while the above-chance result on every inventory argues against a single inventory-specific lexical shortcut. Loading cosines for matched constructs are a second, stricter check because they compare directions rather than decoder labels.

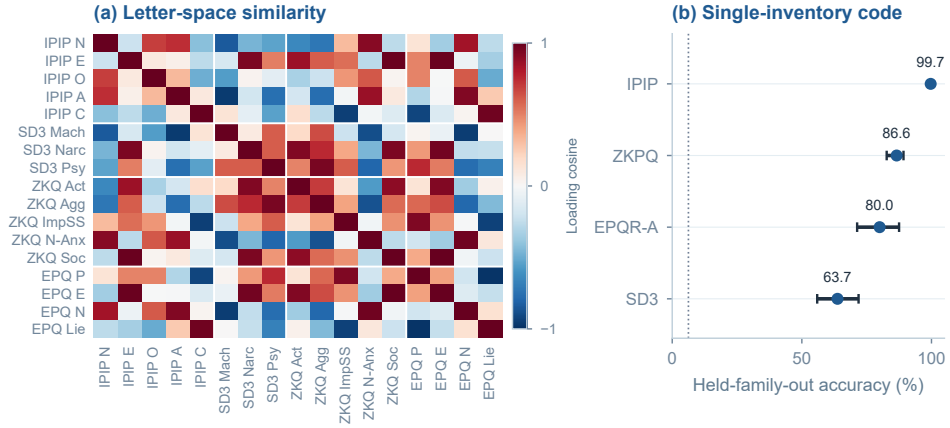


Figure 6: Cross-inventory transfer. The letter-space similarity matrix preserves matched constructs across four inventories (left); inventory-stratified held-family-out decoding remains above chance for every battery (right).

3.7 External replication and generalization

Two family-external models reproduce both levels. On original items, persona×item shares are 57.7%/55.8% for Qwen3.8-27B/Mistral Small 3.2, first-order energies are 0.895/0.842, and archive-centroid decoding is 14/16 and 12/16. Masking all letters and type names increases first-order energy to 0.966/0.955; rewording preserves profile correlations of 0.985/0.986; paraphrased items retain domain-profile correlations of 0.970/0.959. Big-Five prompts, role personas, and forced-binary responses preserve the same low-dimensional interaction structure (Table 4).

These controls constrain different failure modes. The masked-template arm removes explicit MBTI tokens, so its first-order concentration cannot be attributed to reading the four letters as labels. Reworded prompts test wrapper sensitivity, whereas paraphrased items hold the persona prompt fixed and perturb the item surface form. Their high profile correlations show that the response policy is not simply keyed to a few memorized strings, although correlation does not establish semantic equivalence item by item. The Big-Five and role arms change the persona taxonomy itself; the retained low rank therefore identifies a property of factorial or structured conditioning more generally. Forced binary changes the response scale while preserving the item content, ruling out the possibility that the main interaction is only a Likert midpoint artifact.

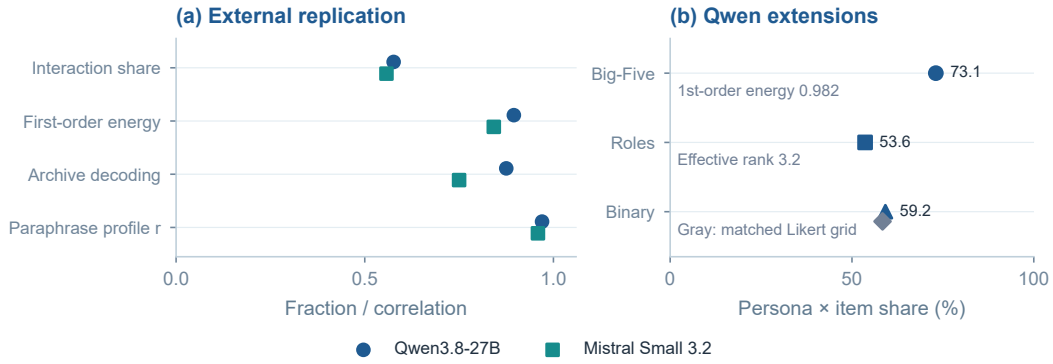


Figure 7: External replication and generalization. Points compare persona×item shares, first-order energy, archive-centroid decoding, and paraphrase profile correlation for the two family-external models; the right panel summarizes Big-Five, role-persona, and forced-binary extensions on Qwen3.8-27B.

Table 4: Replication, ablation, and generalization. Entries are Qwen3.8-27B / Mistral Small 3.2 when paired; extension rows use Qwen3.8-27B.

Condition	r_{pi}	Walsh 1st / profile r	Key diagnostic
Original template/items	57.7/55.8%	0.895/0.842	14/16, 12/16 decode
Masked template	53.2/42.9%	0.966/0.955; 0.920/0.886	no letters/type names
Reworded template	55.1/54.4%	0.890/0.830; 0.985/0.986	wording preserved
Paraphrased items	55.1/57.6%	0.881/0.828; 0.970/0.959	meaning preserved
Big-Five prompts	73.1%	0.982; –	median axis $R^2 = 0.990$
Role personas	53.6%	–; –	15/16 decode
Forced binary	59.2%	–; –	loading Pearson r 0.922–0.950

3.8 Null and robustness boundary

Under the specified PC1/k-nearest-neighbor null, a one-dimensional coordinate reproduces 32.9% of the observed interaction on Likert items (68% of 48.7%) and 41.9% on binary items (81% of 52.0%). Letter-aligned coordinates absorb 86–101%, whereas random four-dimensional bases absorb only 34–44%. The result therefore rejects this one-dimensional smoothing account at matched bandwidth; it does not rule out every heterogeneous multi-dimensional latent-variable model (Figure 8).

The null is deliberately favorable to a one-dimensional trait account. Let θ_{mp} be a profile's latent coordinate and let each item be generated from an unrestricted response curve $g_i(\theta)$, estimated by leave-one-out nearest-neighbor resampling. This model can reproduce nonlinear and non-monotone item curves; it only forbids a persona-specific item rule after conditioning on θ. We simulate five draws per cell with the observed within-cell noise and recompute the same orthogonal share. The PC1 and equal-width-bin variants agree (about 33–42%), showing that the result is not caused by one implementation of this one-dimensional estimator. Adding PC2 raises absorption to 84–95%; four principal components reach 84–102%, and letter-aligned four-dimensional coordinates reach 86–101%. The increase is selective: random four-dimensional coordinates absorb only 34–44%.

There is a necessary saturation caveat. With 16 persona conditions, four coordinates can interpolate a large fraction of an arbitrary persona surface, especially when the neighborhood size is reduced from 34 to 17; the absorbed Likert share rises from 86% to 109% for the letter basis. We therefore treat the multidimensional rows as an estimator-specific upper envelope, not as evidence that such a trait generated the data. The robust negative result is narrower: a single stable coordinate with flexible item curves does not account for the observed interaction at matched smoothing, while random multidimensional smoothing does not recover it either. A fixed multidimensional latent model with heterogeneous loadings remains compatible with these observations and requires new interventions to distinguish. What remains directly identified is a low-dimensional, label-aligned response code whose item expression is heterogeneous.

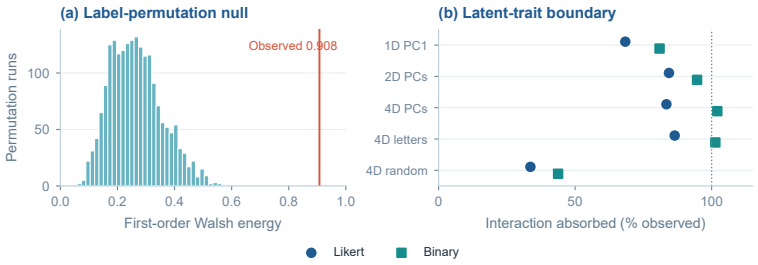


Figure 8: Null and robustness boundary. The permutation null lies far below the observed first-order Walsh energy (left); absorbed interaction share rises for letter-aligned latent dimensions but remains much lower for random four-dimensional bases (right). The one-dimensional trait null absorbs 68% of the Likert interaction and 81% of the binary interaction.

4 Discussion

What is identified. The identified object is a conditional response map. Equation 1 transfers four signed contrasts across families and captures most centered factorial energy, while Equation 2 shows heterogeneous item loadings: Δ_{ci} changes sign and magnitude (sign-consistency 0.290, 0.122, 0.271, and 0.104). The interaction is low-rank, and the key identity explains instrument-sensitive scores.

What is not identified. The estimates do not identify an internal semantic mechanism, training cause, or validity outside self-report. Prompt factors need not be human traits or neural subspaces; the schematic summarizes measured relationships. A trait claim requires invariance across items, wording, instruments, formats, and external behavior under a prespecified criterion.

Limitations and external validity. The archive contains 20 time-stamped models from eight families and five stochastic samples per cell; one item–condition cell is incomplete and extensions use fewer models. Shares depend on inventory and normalization, and the extensions do not establish universality. Full audits and null tables are in the Appendix.

5 Conclusion

Our scientific question was whether reliable persona behavior is a transferable prompt code, an item-indexed response policy, or a stable latent trait. The evidence supports the first two levels jointly: prompt identity is compositional and transferable, while its item expression is heterogeneous and key-sensitive. Questionnaire scores therefore cannot be read as invariant traits; the specified one-dimensional null is rejected at the score level, while broader trait claims remain unidentified without external criteria and stronger invariance tests.

Three measurements locate this boundary. Prompt identity transfers across model families (319/320 held-family-out, 99.1% under equal keying), and 90.8% of centered factorial energy lies in four single-letter contrasts. At the item level, the same axis shifts different items by different amounts, with persona×item interactions accounting for 48.7% of Likert variation versus 0.36% for model identity and remaining organized in roughly four directions. At the score level, equal-key reweighting moves domain scores by 0.041 and reverses 10.2% of untied rankings, within the algebraic key-imbalance bound. Both levels reproduce on two family-external models under template, wording, memorization, taxonomy, and format controls.

The practical implication is to report the prompt code, item-policy heterogeneity, equal-key scores, and key imbalance together. Decoding and ablation test whether a persona instruction changes model responses; external behavioral criteria and interventions are still needed to determine what that response means. The present decomposition makes that next test explicit: hold the response policy fixed while varying the code, rather than treating one questionnaire aggregate as a stable trait.

Ethics Statement

No human participants or personal data were collected. The study uses public questionnaire items and model APIs; released archives contain parsed responses and remove raw telemetry.

Reproducibility Statement

The paper reports the complete model, persona, inventory, sampling, and estimand design. The archived code contains the analysis scripts, prompt templates, model manifest, intermediate result tables, figure sources, and serving details needed to reproduce the reported summaries. Bootstrap, null-model, scoring, and cross-inventory protocols are specified in the Appendix; all numerical claims in the paper were regenerated from the archived result files before submission.

AI Use Statement

Generative AI tools contributed to study design, analysis planning, result interpretation, code drafting, figure prototyping, and language editing. The authors reviewed AI-assisted material against the archived data and analyses and remain responsible for the claims, figures, and submission.

References

Anton Aluja, Jérôme Rossier, Luis F García, Alois Angleitner, Michael Kuhlman, and Marvin Zuckerman. 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, 2006.

Jacob Andreas et al. Measuring compositionality in representation learning. In International conference on learning representations, volume 1, pp. 3, 2019.

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

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

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

Meir Feder and Neri Merhav. Relations between entropy and error probability. IEEE Transactions on Information theory, 40(1):259–266, 1994.

Leslie J Francis, Laurence B Brown, and Ronald Philipchalk. The development of an abbreviated form of the revised eysenck personality questionnaire (epqr-a): Its use among students in england, canada, the usa and australia. Personality and individual differences, 13(4):443–449, 1992.

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

John A Johnson. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. Journal of research in personality, 51:78–89, 2014.

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

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. arXiv preprint arXiv:2601.10387, 2026.

Jelena Meyer, David Garcia, and Dirk U Wulff. Apparent psychological profiles of large language models are largely a measurement artifact. arXiv preprint arXiv:2606.20205, 2026.

Delroy L Paulhus. Measurement and control of response bias. 1991.

Hubert Plisiecki, Sabina Siudaj, Kacper Dudzic, Anna Sterna, Maciej Gorski, Karolina Drozdz, and Marcin Moskalewicz. The pinocchio dimension: Phenomenality of experience as the primary axis of llm psychometric differences. arXiv preprint arXiv:2605.05080, 2026.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. Nature Machine Intelligence, 7(12):1954–1968, 2025.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. Challenging the validity of personality tests for large language models. In Proceedings of the 2025 Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 74–81. 2025.

Tommaso Tosato, Saskia Helbling, Yorguin-Jose Mantilla-Ramos, Mahmood Hegazy, Alberto Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. Persistent instability in llm's personality measurements: effects of scale, reasoning, and conversation history. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 37961–37969, 2026.

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

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. Sycophancy is not one thing: Causal separation of sycophantic behaviors in llms. arXiv preprint arXiv:2509.21305, 2025.

Kim Zierahn, Cristina Cachero, Anna Korhonen, and Nuria Oliver. Personality without persons? a psychometric critique of big five testing in large language models. arXiv preprint arXiv:2607.02325, 2026.

Bruno D Zumbo. A handbook on the theory and methods of differential item functioning (dif). Ottawa: National Defense Headquarters, 160:53, 1999.

Table 5: Equal-key sensitivity for all 13 mixed-key domains.

Scale	Domain	n_+	n_-	Mean $\lvert S - B \rvert$	Default ρ
IPIP	Neuroticism	19	5	0.044	0.956
IPIP	Extraversion	21	3	0.044	0.788
IPIP	Openness	15	9	0.011	0.949
IPIP	Agreeableness	7	17	0.031	0.938
IPIP	Conscientiousness	17	7	0.027	0.961
SD3	Narcissism	6	3	0.027	0.478
SD3	Psychopathy	7	2	0.035	0.958
ZKPQ	Activity	8	2	0.045	0.623
ZKPQ	Aggression	7	3	0.070	0.968
ZKPQ	Neuroticism-Anxiety	9	1	0.122	0.963
ZKPQ	Sociability	4	6	0.012	0.779
EPQR	Extraversion	3	3	0.000	0.997
EPQR	Lie	4	2	0.058	1.000

A Additional interpretation and limitations

Memorization of model identities cannot explain held-family transfer, and continuation probes found no large verbatim contribution under the tested conditions. A high-variance item can still contribute to a finite-sample ANOVA, so individual-item influence should be audited with leave-one-item-out and influence diagnostics in future work. Template wording and item wording have been ablated within the questionnaire setting; the forced-binary arm tests one response-format change, while other response options, inventory coverage, and conversational deployment remain untested. A causal account would require matched base and aligned models, prompt interventions, and behavioral criteria outside self-report; the present decomposition specifies which factors those experiments should hold fixed.

B Full key-sensitivity audit

Table 5 reports all mixed-key domains. S is the published reverse-scored mean and B is the equal-key mean.

C Additional analyses

Scoring integrity. Reconstructed keyed values match every stored scored value (0 mismatches). The single incomplete item is 'sd3_026'.

Default condition. The default instruction is a structured prompt condition. Nearest-centroid matching in the decoder's standardized domain space (persona centroids pooled over all 20 models) assigns nine of 20 default profiles to ISTJ (leave-one-out centroids: eight), and projecting each default profile onto its model's empirical letter loadings types 13 of 20 as ISTJ outright, with a 14th within one letter.

D Item-policy and sampling diagnostics

The complete orthogonal decomposition has 2,320 persona×item degrees of freedom on the Likert grid. Per degree of freedom, its squared-norm share is 1.1 times the model main effect for Likert and 1.7 times for binary; the interaction-to-within-cell noise ratios are $F = 1{,}045$ and 649. Matched pure-noise archives place only 1.3% and 2.6% in the persona×item component. The model-averaged interaction has participation-ratio effective rank 3.9. A domain-centered key-direction vector spans 0.76%/0.59% of the interaction (Likert/binary), while raw-space persona×item shares are 1.34%/3.97%; keyed space inflates the persona main effect by 3.1–3.5×. Key imbalance correlates negatively with within-domain persona share (Spearman −0.30).

Removing the Default condition gives persona×item shares of 50.46% (Likert) and 53.39% (binary), with model identity at 0.31% and 0.41%. The held-family decoder remains 99.4–99.7% for one through five samples per cell. Mean absolute score error relative to five samples is 0.0189 at one sample and 0.0074 at three; split-half domain MAE is 0.028 for

Default and 0.018 for MBTI conditions. These checks quantify the full sampling and keying audit behind Figures 4 and 5.

E Model manifest

The archive contains Claude-Opus-4.6, Claude-Sonnet-4.6, DeepSeek-V3.2, DeepSeek-V4-Flash, DeepSeek-V4-Pro, Gemini-3-Flash-Preview, Gemini-3.1-Flash-Lite, Gemini-3.1-Pro-Preview, Gemini-3-Pro-Preview, GLM-4.6V, GLM-4.7, GLM-5.1, GPT-5.5, GPT-5.2, Kimi-K2.5, Kimi-K2.6, MiniMax-M2.7, Qwen3-235B-A22B, Qwen3.5-122B-A10B, and Qwen3.5-397B-A17B. Family labels and file hashes are in 'code/results/iclr/data_manifest.csv'. The replication suite adds Qwen3.8-27B and Mistral Small 3.2 24B (Section 3.7); provider versions and serving details are recorded in the accompanying manifest.

F Theory-facing diagnostics

Differential persona functioning. Measurement-invariance work treats item-level parameter changes as differential item functioning (DIF) (Vandenberg & Lance, 2000; Zumbo, 1999). For each MBTI axis c, we define the keyed item contrast

$$\Delta_{ci} = \mathbb{E}[y_i \mid c = +1] - \mathbb{E}[y_i \mid c = -1],$$
$$\mathrm{DIFVar}(c) = \frac{1}{I}\sum_i (\Delta_{ci} - \bar{\Delta}_c)^2, \qquad \mathrm{SCI}(c) = \frac{1}{I}\left|\sum_i \mathrm{sign}(\Delta_{ci})\right|. \tag{8}$$

The four axes have $(\mathrm{SCI}, \mathrm{DIFVar})$ equal to $(0.290, 0.1261)$ for EI, $(0.122, 0.0265)$ for SN, $(0.271, 0.0648)$ for TF, and $(0.104, 0.0832)$ for JP. The low sign-consistency values and wide contrast ranges are descriptive evidence against a single common-slope/item-invariant shift under the inventory's keyed direction. The S/N weakness is not a pooling artifact: recomputed within each model, S/N has the smallest DIFVar of the four axes in all 20 models and the smallest SCI in nine of 20 (mean within-model SCI rank 1.6 of 4; SCI below 0.05 in four models). They motivate a DIF or multidimensional measurement model; they do not identify a human latent trait or establish a causal semantic mechanism.

Fano information lower bound. Let P be uniform over the 16 MBTI labels, X the observed profile, and $e = \Pr[\hat{P}(X) \neq P]$. Fano's inequality gives

$$I(P; X) \geq \log_2 16 - h_2(e) - e \log_2 15, \tag{9}$$

with h_2 the binary entropy (Feder & Merhav, 1994); the reported values plug in the empirical held-family, equal-key, and model-held-out errors.

Label-specificity of Walsh concentration. A 2,000-draw permutation test that randomly reassigns the 16 persona labels within each model leaves the response profiles fixed. The null first-order mass has mean 0.266 and standard deviation 0.091, close to the analytic random-basis baseline 4/15, while the observed mass is 0.908 (no permutation exceeded it; plus-one Monte Carlo $p_{\mathrm{MC}} = 4.9975 \times 10^{-4}$ for 2,000 permutations). This check addresses the possibility that any arbitrary 16-way ordering would produce the reported Walsh concentration.

G Latent-trait null simulation

A single latent trait with item-specific response curves generates persona×item covariance without any item-specific persona modulation, so it is the natural null for the interaction share. We simulated a deliberately favorable but specific version: θ is each observed profile's first principal component, every item's characteristic curve is saturated (estimated freely on a ten equal-width θ-bin×item grid, so monotone and non-monotone ICCs are both available), and complete archives are redrawn with matched within-cell noise (B=200). This model reproduces 68% of the observed Likert interaction and 81% of the observed binary

interaction—32.9% versus 48.7% on Likert (simulated 95% envelope ≤ 33.2%) and 41.9% versus 52.0% on binary (≤ 42.3%)—and underpredicts the persona main effect (1.7% versus 4.1% on Likert), which the letter regressions of Section 3.6 show to be multidimensional.

Table 6: Multidimensional latent-trait nulls. Respondents are the 340 $(model, persona)$ cells; θ is a d-dimensional projection of the keyed item profile, and each simulated archive resamples every cell's five draws from the pooled draws of its leave-one-out k-nearest neighbors in θ-space (B=200). k=34 is a matched neighborhood size for the one-dimensional comparison; k=17 shows the effect of estimator saturation. "Absorbed" is the null's mean persona×item share as a percentage of the observed value (Likert / binary).

θ basis	d	k	Likert / binary r_{pi}	Absorbed
(observed)	–	–	48.7% / 52.0%	100%
θ-deciles (1D, above)	1	bin	32.9% / 41.9%	68 / 81%
PC1	1	34	33.3% / 42.1%	68 / 81%
PC1–2	2	34	41.1% / 49.2%	84 / 95%
PC1–4	4	34	40.7% / 53.1%	84 / 102%
Letter loadings	4	34	42.1% / 52.7%	86 / 101%
Letter loadings	4	17	53.3% / 58.8%	109 / 113%
Random orthonormal	4	34	16.4% / 22.8%	34 / 44%

Table 6 extends the null to higher-dimensional traits. The kNN version of the 1D null (33.3%/42.1%) reproduces the equal-width-bin version (32.9%/41.9is not an artifact of changing the one-dimensional estimator. Absorption rises with d only when the added directions are aligned with the data's principal or letter structure: two dimensions absorb 84–95%, four letter-aligned dimensions 86–101%, while four random directions absorb less than one PC1 (34–44%). The multidimensional nulls also close the persona-main-effect gap (4D-letter: 3.8% versus 4.1% observed on Likert, where the 1D null gave 1.7%). Two qualifications bound the reading. First, saturation: with only 16 personas, a four-dimensional response surface has no residual degrees of freedom at the persona level—any $16 \times item$ mean pattern is representable—and the absorbed share indeed grows with estimator flexibility (86%→109% on Likert as k drops from 34 to 17), so the 86–101% figure is an upper envelope on what a 4D trait could express under this estimator, not evidence that one generated the data. Second, specificity: the random-basis row shows the absorption is not generic neighborhood smoothing. What this analysis rejects at matched smoothing is the specified one-dimensional trait account (null 95% envelope ≤33.5% Likert versus 48.7% observed). The item-level finding is therefore best stated as: the persona code is expressed through item non-uniformity organized by the same low-dimensional axes (effective rank 3.9, first-order Walsh dominance), not through a single common slope and not through item-idiosyncratic freedom.

H Replication and ablation protocol

The suite uses Qwen3.8-27B (served locally, non-thinking mode) and Mistral Small 3.2 24B Instruct 2506 (AWQ INT4, three replicas on RTX 3090 GPUs via vLLM; a chat template with the provider's default system prompt and date interpolation removed, rendered byte-identically to the provider template otherwise), both at temperature 0.7 with the same parser, normalization, and keyed transform as the main archive, five samples per cell for the replication and paraphrase modules and three for the template ablation. Totals per model: 58,444 calls (including a 100-call protocol pilot); parse failures 6 (0.01%) for Qwen3.8 and 26 (0.04%) for Mistral (24 in Default and 2 in non-default conditions); no network failures. The replication module repeats the full 17-condition grid (18,785 responses). The template ablation administers the masked and reworded prompts (21,216 responses); the paraphrase module administers meaning-preserving paraphrases of all 221 items under the original prompts, with direction and negation preserved and 20 items manually audited (17,680 responses). The memorization module elicits greedy continuations of each item's first half without a chat template in three conditions—unframed, framed as a verbatim questionnaire item, and the same probes on paraphrased items as a no-exposure control—on items with at least a two-word suffix (123 original, 161 para-

phrased). Because the Mistral serving stack's raw-completion endpoint returns generations with word-internal spaces deleted (a detokenizer fault confirmed on all three replicas); verbatim hits are counted with a whitespace-insensitive criterion—the whitespace-stripped true suffix must appear as a contiguous substring of the continuation—which reproduces the word-level counts on the Qwen3.8 archive exactly (0/123, 8/123, 2/161). Decoding uses archive centroids standardized by training statistics: the seven non-Alibaba families (17 models) for Qwen3.8 and all 20 models for Mistral (its family is absent from the archive). Scripts: 'code/qwen_tier1_runner.py', 'code/qwen_tier1_probe.py', 'code/qwen_tier1_analysis.py', and 'code/e1_supplement.py'; per-item records and the paraphrase manifest are archived alongside the main archive.

Persona-system and format arms (Section 3.7). Three arms administered on the same endpoint and protocol: (F) 32 Big-Five factorial personas (2^5, each axis named at both poles by a fixed sentence, masked-style wrapper, no inventory or letter tokens), all 221 items, five samples (35,360 responses); (G) 16 natural-language role personas matched in length and wrapper with no axis structure (17,680 responses); (H) the 120 IPIP-NEO items re-administered as forced yes/no questions under the 16 MBTI personas (9,600 responses). Parse failures: 0, 1, and 1. Analysis decomposes each arm with the same estimators as the tier-1 suite (two-way shares, Walsh energy, letter regressions, effective rank = participation ratio of squared singular values); held-out separability decodes domain profiles trained on samples 1–2 against samples 3–5. Cross-system construct maps correlate per-domain axis loadings between arm F and module A of the same model over the 17 shared domains. Script: 'code/e2_e3_analysis.py'; outputs in 'code/results/iclr/round4/'.

Table 7: Persona-system and format generalization on Qwen3.8-27B (Section 3.7). Effective rank is the participation ratio of the persona-effect spectrum; the MBTI and Big-Five rows are the same model's factorial grids.

Arm	r_{pi}	first-order energy	effective rank
MBTI letters, 16 personas	57.7%	0.895	4.40
Big-Five poles, 32 personas (F)	73.1%	0.982	5.12
Natural-language roles, 16 personas (G)	53.6%	–	3.18
IPIP-120 forced binary (H)	59.2%	–	–

Table 8: Cross-system construct map: Pearson correlations between Big-Five axis loadings (arm F) and MBTI letter loadings (same model, 17 shared domains). Signs follow the coding E/N/F/P = +1 and I/S/T/J = −1, so the negative C–J/P entry places high-C on the Judging pole; the A and N rows align with the Feeling pole in this archive. These correlations describe this response map and do not establish construct equivalence.

	E/I	S/N	T/F	J/P
Extraversion	0.956	−0.04	−0.10	0.10
Openness	−0.10	0.774	0.19	0.69
Conscientiousness	−0.08	−0.43	−0.06	−0.90
Agreeableness	−0.15	−0.03	0.65	−0.16
Neuroticism	−0.44	0.21	0.55	0.08