Tracing Psychometric Inference in Large Language Models

Chih-Hao Hsu [1] **Feng-Chun B. Chou** [1] **Pin-Hao Andy Chen** [1]

Abstract

Do large language models represent psychological constructs internally, or do they only generate outputs that resemble psychometric structure? We address this question using a cross-persona paradigm across 14 models (Llama 3 and Qwen 2.5, 0.5B–14B, base and instruct). Given responses on one psychological scale from real individuals ($N = 272$), models predict responses on six other scales. At the behavioral level, LLMs reproduce human cross-scale correlation structure and systematically amplify it, with model-generated correlations exceeding human estimates even after correcting for measurement attenuation. We then examine whether this structure is reflected internally. Contrastive direction analysis reveals an organized geometry in activation space aligned with psychometric relationships. This structure emerges in large instruct models but is not observed in base models. Across models, geometry strength predicts behavioral amplification ($r = 0.68$, $p = 0.008$; partial $r = 0.69$ controlling for log size), independent of model size. To relate internal structure to output behavior, a matched activation-probing paradigm shows that representational amplification is less variable than behavioral amplification (1.38–1.77 vs 0.53–1.53). A synthetic control with known ground-truth structure suggests that a representation-level component remains after subtracting a ridge-probe baseline (≈ 1.3), with adjusted slopes still predicting behavior at $r = 0.88$. We term the systematic representation-to-behavior gap *readout attenuation*. Together, these findings suggest that LLMs encode structured representations aligned with psychological constructs, while differences in output primarily reflect how these representations are read out.

[1]National Taiwan University. Correspondence to: Pin-Hao Andy Chen <andyphchen@ntu.edu.tw>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

A growing body of work shows that large language models can simulate human psychological profiles at the group level, producing response patterns that resemble survey data in structure and covariance (Liu et al., 2026; Argyle et al., 2023; Serapio-García et al., 2023). This raises a fundamental interpretability question beyond behavioral evaluation: when an LLM reproduces psychometric structure in its output, does it reflect internal representations of psychological constructs, or merely generate similar patterns from surface-level statistical associations? A model that emulates without internal representation is a convincing imitator, not a system that "has" the structure in any meaningful sense. The same ambiguity arises in LLM interpretability regarding whether next-token predictors develop world models of domains they reason about (Li et al., 2023; Nanda et al., 2023), and makes the psychometric case particularly tractable: psychology supplies a rich, quantitative, population-validated reference in the 16×16 human subscale correlation matrix.

We design a cross-persona paradigm to probe this question: a subject's responses on one psychological scale are placed in the LLM prompt, and the model is asked to complete items from a different, unseen scale. This paradigm has two properties that make it well-suited for probing internal representation. First, by anchoring the model to a specific human subject's response pattern rather than eliciting generic self-report, it forces the model to perform cross-construct inference on a per-subject basis, generating a subject-by-construct response matrix directly comparable to the human ground truth. Second, by rotating which scale serves as the persona, every pair of the subscales is evaluated in both directions, yielding a bidirectional 16×16 correlation matrix that characterizes the full between-construct structure rather than a one-way projection from a single anchor scale (e.g., Big Five used as fixed input in Liu et al. 2026). Behaviorally, across 14 models (Llama 3 and Qwen 2.5, 0.5B–14B, base and instruct) and $N = 272$ participants, LLMs reproduce and systematically amplify human cross-scale correlations, with amplification persisting after correcting for measurement attenuation (§3). We trace this amplification to model internals using contrastive direction extraction, activation probing, and steering, asking whether (a) psychometric structure is encoded in mid-layer representations, (b) amplification arises at specific model depths, and (c)

representational strength predicts behavior across models.

We distinguish three levels of evidence throughout the paper. First, activation probing asks whether cross-scale psychometric information is linearly decodable from model states. Second, contrastive geometry asks whether construct directions form an organized similarity structure aligned with human psychometrics. Third, steering asks whether these directions are causally accessible during generation. Our strongest evidence concerns the first two levels; steering provides a complementary causal-accessibility check rather than a complete account of how the model uses these directions.

These analyses yield three contributions:

1. **Structured psychometric geometry emerges only with sufficient scale and instruction tuning.** Contrastive directions for 16 subscales recover a similarity structure aligned with human psychometric correlations in large instruct models, but not in base models of any size (Mantel r up to 0.66 for Qwen 14B Instruct; ≤ 0.14 for all base). Notably, base and instruct models with comparable behavioral performance can differ sharply in whether such representational structure is present.
2. **Representational amplification is detectable before behavioral readout.** Representational amplification is less variable than behavioral amplification (1.38–1.77 vs 0.53–1.53) and remains associated with behavior after subtracting a synthetic ridge-probe baseline (adjusted slope predicts behavior at $r = 0.88$, $p = 0.002$). We term the resulting representation–behavior gap *readout attenuation*, which decreases with model scale and instruction tuning.
3. **Representational geometry quantitatively predicts behavioral amplification across models** ($r = 0.68$, $p = 0.008$; partial $r = 0.69$ controlling for log size), with a stronger relationship observed among instruct models ($r = 0.90$, Fisher-z $p = 0.064$ given $n = 7$ per group).

2. Related Work

LLM psychometric simulation. A growing body of work examines the ability of LLMs to simulate human psychological profiles and individual agents (Park et al., 2024a; Aher et al., 2023; Dillion et al., 2023; Jiang et al., 2024; Choi & Li, 2024). Closest to our work, Liu et al. (2026) prompted LLMs with Big Five responses to predict nine other scales, reporting strong alignment and systematic amplification, which they attribute to noise filtering (we revisit this account in §7). Related work examines LLM personality profiles (Serapio-García et al., 2023; Sorokovikova et al., 2024), psychometric benchmarks (Huang et al., 2024; Samuel et al., 2024), response biases (Tjuatja et al., 2024), and the validity of applying psychological instruments to LLMs (Pellert et al., 2024; Dong et al., 2025). We extend this line of work by moving from behavioral characterization to a representation-level account.

Representation engineering and probing. Linear probing and contrastive methods identify interpretable directions in LLM activation space corresponding to concepts such as truthfulness (Marks & Tegmark, 2024; Burns et al., 2023; Li et al., 2024), spatial and temporal structure (Gurnee & Tegmark, 2024), linear features more generally (Park et al., 2024b; Elhage et al., 2022), world models in trained sequence models (Li et al., 2023; Nanda et al., 2023), and persona traits (Durmus et al., 2024; Zou et al., 2023). Activation Addition and Contrastive Activation Addition demonstrate that adding vectors along these directions causally steers model behavior (Turner et al., 2024; Zou et al., 2023). We adapt these methods to psychometric constructs, using contrastive direction extraction and activation steering to probe how psychological traits are represented and processed.

3. Setup: Cross-Persona Paradigm and Behavioral Baseline

3.1. Human psychometric data

We use psychometric data from $N = 272$ participants who completed seven validated Chinese-language scales encompassing 114 items and 16 subscales (Table 1). These scales span diverse psychological domains including empathy (IRI), affect (PANAS), well-being (POM, SWLS), personality (BFI-10), self-construal (SCS), and loneliness (UCLA). Consistent with prior cross-scale analyses of LLM psychometric behavior (Liu et al., 2026), we operate throughout at the *subscale* level, using the 16×16 human subscale correlation matrix (120 off-diagonal pairs) as ground truth; scale-level analysis (7×7) yields too few pairs while item-level (114×114) suffers from within-subscale redundancy. Bootstrap resampling (2000 iterations) confirms matrix stability ($r = 0.93$, 95% CI: [0.89, 0.96]; App B).

3.2. Cross-persona paradigm

Each of the seven scales serves as the *persona scale* in turn. A subject's item-level responses on the persona scale are placed in the system prompt ("You are completing psychological questionnaires as a person with the following responses on the [scale name]: Q1. [text] $\rightarrow$ [response] ..."), and the model is asked to respond to items from all other scales. Each target item is administered as a separate call, with the response extracted as the argmax over valid response option log-probabilities. This yields 7 rotations per model, each producing cross-scale responses for all 272 subjects. We verify the resulting 16×16 correlation ma-

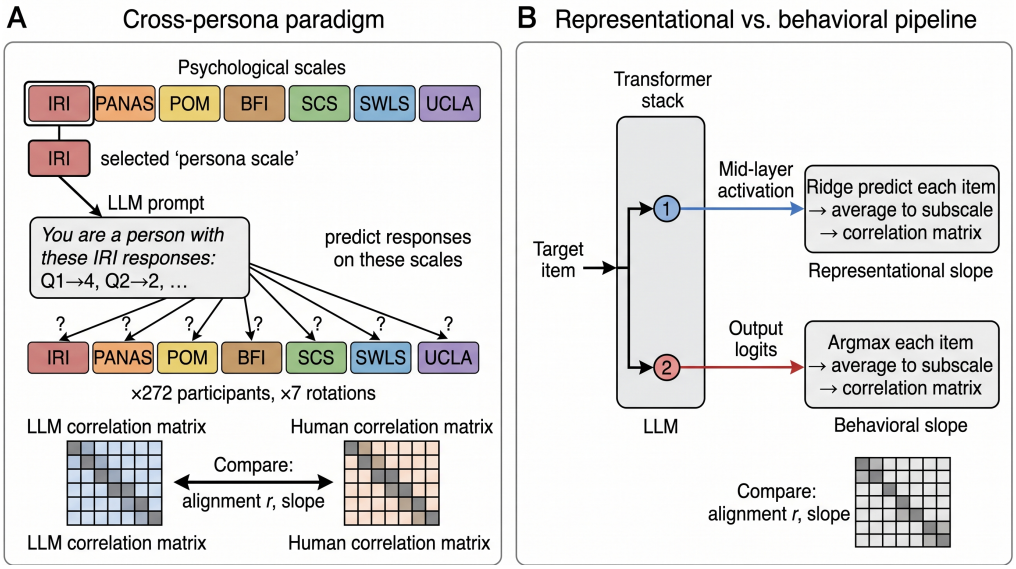

Figure 1. Overview of experimental design. **Left**: Cross-persona paradigm: a subject's responses on one persona scale are provided in the prompt, and the model predicts responses on other scales across all 7 scale rotations. **Right**: Behavioral and representational pipelines extract parallel 16×16 cross-scale correlation matrices from argmax outputs and mid-layer ridge-predicted responses, respectively, and compare them to the human ground truth.

trix against trivial in-prompt copying: within-scale subscale pairs (sharing the persona scale, where direct context is available) are not systematically closer to human correlations than cross-scale pairs (App C).

We evaluate 14 models: Qwen 2.5 (0.5B, 3B, 7B, 14B) and Llama 3 (1B, 3B, 8B), each in base and instruct variants. All models receive identical raw-text prompts (no chat template) to isolate the effect of instruction tuning weights from prompt formatting.

3.3. Behavioral results

For each model, we construct an implicit subscale-level cross-scale correlation matrix by correlating persona-conditioned subscale scores (mean of argmax integer item responses) with human subscale scores across subjects, and compare it to the human ground truth using Pearson correlation (alignment r) and OLS slope.

Two patterns emerge (Figure 2). First, alignment increases with model size, with larger models more closely reproducing human psychometric structure. Second, the slope transitions from attenuation (< 1) in smaller models to amplification (> 1) in larger models, with instruct models showing stronger amplification at scale. Correcting for human measurement attenuation using Cronbach's α (Nunnally, 1975) reduces but does not eliminate amplification in large models: corrected slopes for $\geq$ 7B models remain in the range 1.18–1.29 (see App H for subscale-level analyses and sensitivity checks over reliability thresholds). In contrast, Liu et al. (2026) report slopes near 1.0 after correction. The persistence of amplification after correction motivates our representation-level investigation.

4. Psychometric Geometry in Activation Space

4.1. Contrastive direction extraction

Adapting the persona vectors approach (Chen et al., 2025), for each of the 16 subscales we extract a contrastive direction $\mathbf{d}_k = \bar{\mathbf{h}}_k^{\text{high}} - \bar{\mathbf{h}}_k^{\text{low}}$ between mean residual-stream activations under *high* (target items at maximum, others randomized) and *low* (target items at minimum) conditions, averaged over 10 random backgrounds and 20 open-ended Chinese probe questions per condition (App I). We compare pairwise cosine similarity between the 16 directions to the human subscale correlation matrix via the Mantel test (Mantel, 1967) (10,000 permutations).

4.2. Psychometric geometry emerges with scale and instruction tuning

Our results reveal a striking dissociation (Figure 3). Small models (0.5B–3B) show no psychometric geometry (Mantel $r \approx 0$) regardless of instruction tuning. In contrast, large instruct models show clear structure: Llama 8B Instruct achieves Mantel $r = 0.37$, Qwen 7B Instruct $r = 0.42$, and Qwen 14B Instruct $r = 0.66$. Crucially, *all base models show near-zero geometry regardless of size*: Qwen 14B Base $r = 0.13$, Llama 8B Base $r = -0.06$. Instruction tuning, not just scale, is necessary for organized psychometric representations. A semantic similarity control confirms this

Table 1. Scale inventory. PT = Perspective Taking, FS = Fantasy, EC = Empathic Concern, PD = Personal Distress, PA = Positive Affect, NA = Negative Affect, E = Extraversion, A = Agreeableness, C = Conscientiousness, N = Neuroticism, O = Openness.

Scale	Abbrev.	Items	Range	Subscales
Interpersonal Reactivity Index	IRI	28	1–5	PT, FS, EC, PD
Positive & Negative Affect Schedule	PANAS	20	1–5	PA, NA
Peace of Mind	POM	7	1–6	POM
Big Five Inventory-10	BFI	10	1–7	E, A, C, N, O
Self-Construal Scale	SCS	24	1–7	Indep, Interdep
Satisfaction with Life Scale	SWLS	5	1–7	LifeSat
UCLA Loneliness Scale	UCLA	20	1–4	Loneliness

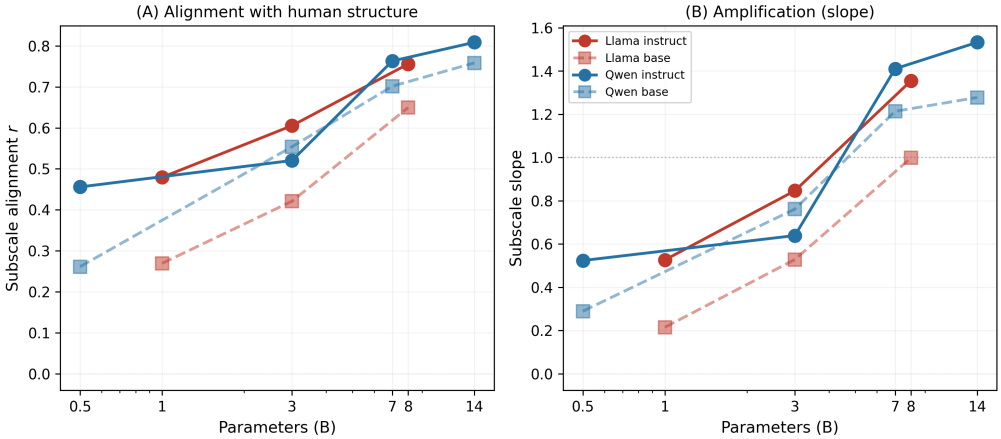

Figure 2. Behavioral scaling across 14 models at the subscale level. **Left**: alignment r between LLM and human cross-scale correlation matrices. **Right**: slope of LLM correlations regressed on human correlations (raw: solid/dashed; attenuation-corrected: dotted). Slopes > 1 indicate amplification. Both alignment and amplification increase with model size. Corrected slopes remain > 1 for large models, indicating amplification is not fully explained by human measurement noise.

geometry is not driven by surface-level item content. Across 14 models, controlling for item-text semantic similarity reduces the mean alignment between contrastive direction geometry and human structure from $r = 0.115$ to partial $r = 0.113$ (a 2% drop), with maximum absolute drop below 0.03 across models (App I).

4.3. Convergent evidence from steering specificity

To test whether extracted directions are causally active rather than read-only correlates, we fit regression-based directions (one per subscale) from 272 subjects' activations and add scaled direction vectors to the residual stream during inference. For each (source, target) subscale pair, we record the change in target-item expected-value response, yielding a steering-response matrix (9 source subscales $\times$ 16 target subscales; see App J). In one Llama 8B Instruct run, the double-centered Mantel test (which removes generic row/column shifts to isolate construct-specific structure) against the human psychometric matrix gives $r = 0.28$, $p = 0.009$ (App J). Across per-model steering runs, double-centered specificity is positive in multiple instruct models, with the strongest effects observed in Llama 3B and Qwen 3B Instruct; however, magnitudes do not scale monotonically with model size and are sensitive to layer and extraction choices. We therefore interpret steering as causal-accessibility evidence rather than as a quantitative scaling result.

5. Locating Amplification in Representations

5.1. Activation probing: paradigm and pipeline

The analyses above establish that psychometric structure exists in the activation space. We now ask whether amplification can be traced to the representational level. We design an activation probing paradigm that mirrors the behavioral experiment: the system prompt contains human responses from one scale, and we extract the decision-point activation (the last token position before the model produces its answer) when presented with a target item from a different scale. Crucially, the target scale's answers are *not* in the prompt, making this a clean test of cross-scale information encoding. For each target scale, we select the first 2 items per subscale to cover all 16 subscales (22–30 target items per rotation), and extract activations for 272 subjects $\times$ 7 rotations across all models.

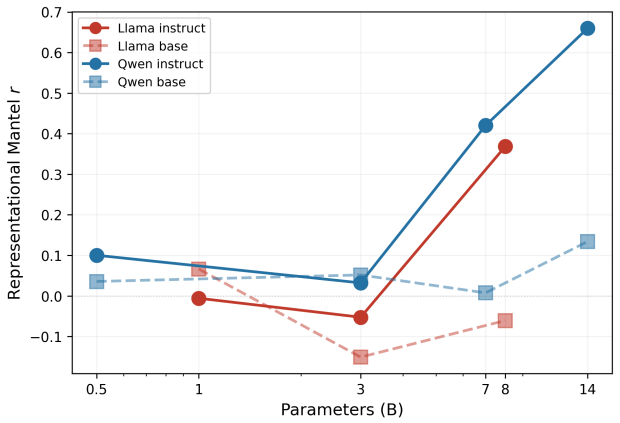

Figure 3. Representational psychometric geometry across 14 models. Mantel r between contrastive direction cosine similarity and the human subscale correlation matrix, as a function of model size. Psychometric geometry is absent in small models and in all base models, emerging only in large instruct models (Qwen 7B+, Llama 8B+).

Cross-scale information is linearly decodable from activations. Using 5-fold cross-validated ridge regression to predict each target subscale score from activations, most cross-scale subscale predictions are statistically significant ($p < 0.05$). For Llama 8B Instruct, 71 of 96 subscale-level cross-scale predictions reach significance with mean CV $r = 0.19$, confirming mid-layer activations encode substantial information about psychological constructs whose items are absent from the prompt.

Matched-pipeline comparison. To enable fair comparison between representational and behavioral amplification, we construct parallel pipelines that differ only in the item-level response extraction step:

- **Behavioral pipeline**: For each target item, extract the argmax over response option log-probabilities (discrete integer). Average item responses within each subscale to obtain subscale scores. Compute the 16×16 correlation matrix of subscale scores across subjects.
- **Representational pipeline**: For each target item, use cross-validated ridge regression to predict the item's continuous response from the decision-point activation. Average predicted item responses within each subscale (applying the same reverse coding). Compute the 16×16 correlation matrix of predicted subscale scores across subjects.

The two pipelines differ only at the item-level extraction step (continuous ridge prediction vs. discrete argmax); both aggregate to subscale scores via averaging, preserving the same variance structure.

5.2. Amplification at the representational level

Using the item-level representational pipeline (§5.1), we compute the predicted correlation matrix for each of nine models spanning two families (0.5B–14B), both instruction-tuned and base variants, including base–instruct pairs at 8B (Llama) and 14B (Qwen) (Table 2).

The central finding is that representational amplification varies substantially less than behavioral amplification (slope 1.38–1.77 vs 0.53–1.53). The gap between the two, which we term *readout attenuation*, decreases with model scale for instruct models: from $+0.91$ (Llama 1B) to $+0.24$ (Qwen 14B). A permutation test on Llama 8B Instruct (100 iterations, shuffling subject labels) rules out a regression artifact: 0 of 100 permuted slopes exceed the observed (permuted mean 0.38 ± 0.17, $p < 0.01$). What scales with model size is thus not amplification itself, already present in mid-layer representations of small models, but the fidelity with which readout expresses this structure in behavioral output.

The two base models are informative: representational slopes are comparable to their instruct counterparts (Llama 8B: 1.63 vs. 1.70; Qwen 14B: 1.70 vs. 1.77), but behavioral slopes are lower (1.00 vs. 1.35; 1.28 vs. 1.53), yielding larger readout attenuation gaps. Extending to all seven base–instruct pairs, instruction tuning reliably improves behavioral alignment (mean Δbeh. slope $+0.22$, paired $p = 0.011$), while the representational geometry gap (ΔMantel r) is positive only for models $\geq$ 7B (mean $+0.46$ for large vs. $+0.02$ for small pairs, consistent with the geometry emergence threshold in §4). Instruction tuning thus primarily improves readout alignment; emergence of organized representational geometry additionally requires a minimum scale threshold.

Discretization and readout decomposition. Comparing continuous ridge, rounded-integer, and argmax pipelines (mean slopes 1.60, 0.71, 1.01), behavioral slope typically *exceeds* the rounded slope, so late transformer layers partially recover what discretization destroys. This late-layer recovery is near-zero in small models but $+0.56$ in Qwen 14B Instruct (rounded $0.97 \rightarrow$ beh 1.53); what scales with model size is thus how effectively late layers route mid-layer amplified structure through the discretized bottleneck, not the bare cost of discretization (App B).

5.3. Synthetic pipeline control

Is the 1.38–1.77 representational range a genuine property of learned representations, or an artifact of the ridge probing pipeline? We address this question using a synthetic control with known ground-truth correlation structure. Specifically, we generate $N = 272$ latent subscale scores from the human 16×16 correlation matrix, project them through random

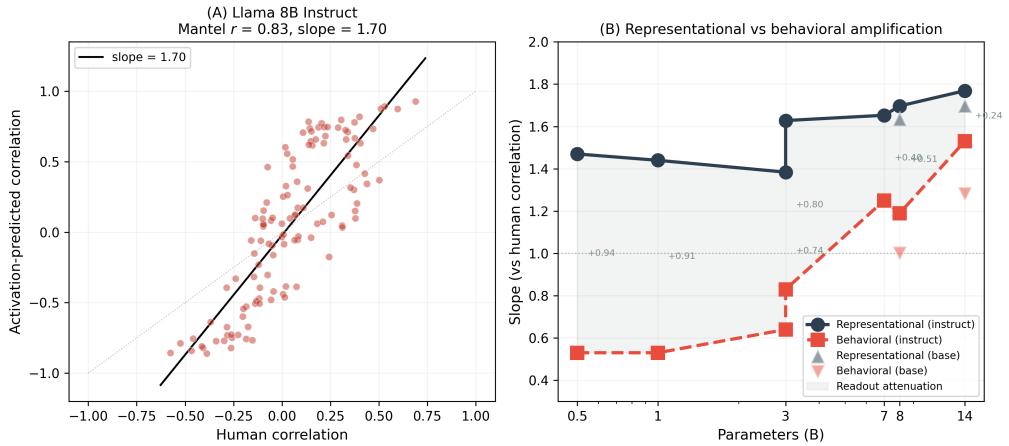

Figure 4. Locating amplification across 9 models. **Left**: Activation-predicted subscale correlations vs. human correlations for Llama 8B Instruct using the item-level pipeline (Mantel $r = 0.83$, slope $= 1.70$). **Right**: Representational slope (solid) vs. behavioral slope (dashed) across 9 models. Representational amplification varies less than behavioral (1.38–1.77 vs 0.53–1.53, spread 0.39 vs 1.00), while both increase with model scale at markedly different rates.

Table 2. Representational vs. behavioral amplification across 9 models. Representational slope is computed using the item-level pipeline (§5.1). Behavioral slope is from argmax integer responses averaged to subscale scores. Mantel r measures alignment between the item-level predicted correlation matrix and human correlation matrix. Best layer is selected post-hoc from a layer-wise sweep balancing Mantel r and slope on the plateau region; App B shows this selection is robust (mean ± 1-layer slope variation 0.025).

Model	Best layer	Repr. slope	Beh. slope	Gap	Mantel r
Qwen 0.5B Instruct	8	1.47	0.53	+0.94	0.73
Llama 1B Instruct	8	1.44	0.53	+0.91	0.81
Qwen 3B Instruct	14	1.38	0.64	+0.74	0.81
Llama 3B Instruct	14	1.63	0.83	+0.80	0.84
Qwen 7B Instruct	18	1.65	1.41	+0.24	0.82
Llama 8B Instruct	16	1.70	1.35	+0.34	0.83
Llama 8B Base	14	1.63	1.00	+0.63	0.82
Qwen 14B Instruct	32	1.77	1.53	+0.24	0.84
Qwen 14B Base	28	1.70	1.28	+0.42	0.82

Gaussian matrices $\mathbf{W} \in \mathbb{R}^{d \times 16}$ matched to each model's hidden dimension, and add isotropic noise at SNR levels $\in \{0.1, 0.3, 1.0, 3.0, 10.0\}$ across 20 random seeds. We then apply the same item-level ridge pipeline (App E).

Pipeline inflation is SNR-dependent, not a fixed offset: recovered slope is ≈ 1.05 at SNR $= 10$ but ≈ 1.5 at SNR $= 0.1$ (Fig. 5A). Mantel r between recovered and ground-truth matrices remains > 0.97 throughout, so the pipeline preserves correlation *structure* faithfully while inflating magnitude. Using each model's contrastive-direction reliability ρ as an approximate SNR proxy (SNR $\approx \rho/(1-\rho)$), we interpolate the synthetic grid for per-model baselines (range 1.11–1.48). Because this proxy is not an item-level split-half reliability estimate, we treat the adjusted slopes as approximate and emphasize their ordering and scale dependence rather than exact magnitudes. Adjusted slopes (observed $-$ baseline) range from -0.09 (Qwen 3B Inst) to $+0.66$ (Qwen 14B Inst) and still predict behavioral slope at $r = 0.88$ ($p = 0.002$; Fig. 5C), compared to $r = 0.91$ for raw, and correlate with log size at $r = 0.68$. If representational slopes were pure pipeline artifact, adjusted values should cluster at zero without scale dependence; instead they show clear ordering (large instruct models $+0.38$ to $+0.66$ vs small models $+0.09$ to $+0.11$).

6. From Representations to Behavior

If representational structure contributes to amplification, models with stronger representational structure should exhibit greater behavioral amplification. Across all 14 models, representational Mantel r correlates with behavioral subscale slope at $r = 0.68$ ($p = 0.008$; LOO range 0.58–0.74; partial $r = 0.69$ controlling for log size; Fig. 6). Within-group correlations are suggestive though underpowered at $n = 7$: $r = 0.90$ ($p = 0.006$) among instruct models, $r = 0.17$ ($p = 0.71$) among base models (Fisher-z $p = 0.064$).

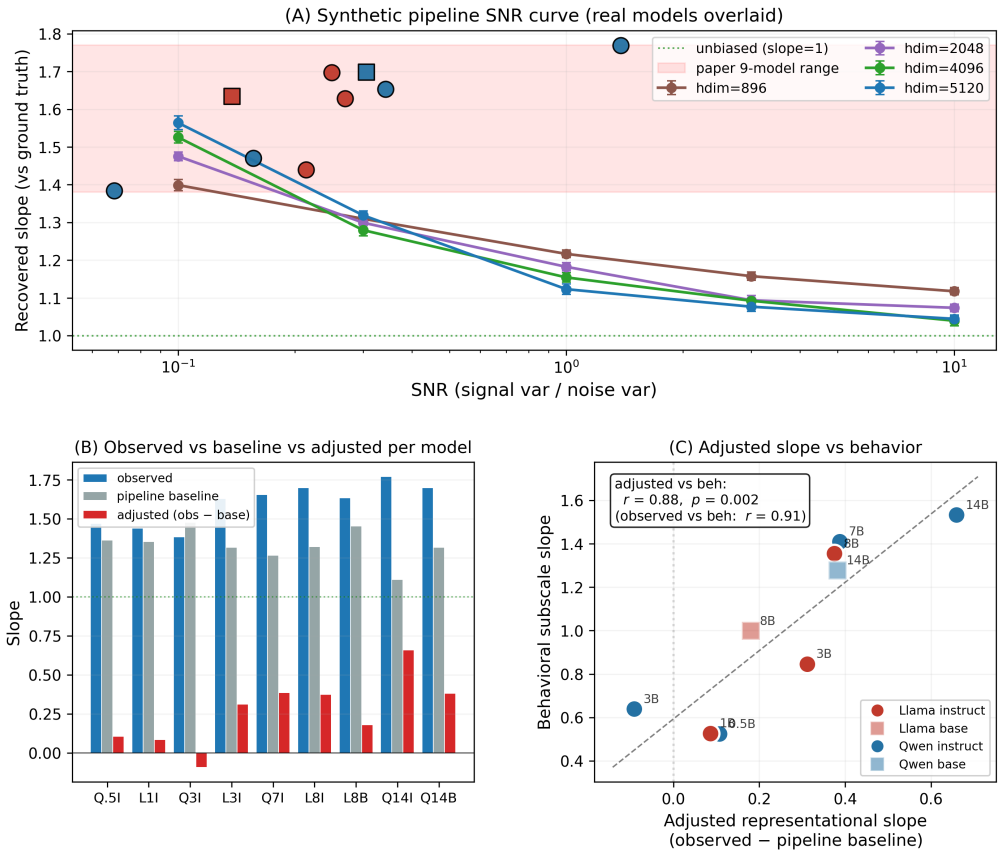

Figure 5. Synthetic pipeline control. **(A)**: Recovered slope of the item-level ridge pipeline on synthetic data as a function of SNR, for four hidden dimensions. Slope approaches 1.0 at high SNR and rises to ≈ 1.5 at low SNR. Real models overlaid using SNR estimated from an approximate contrastive-direction reliability proxy

; most fall within the paper 9-model range (1.38–1.77, red band). **(B)**: Per-model observed slope (blue), SNR-matched synthetic baseline (grey), adjusted slope (red). **(C)**: Adjusted slope still predicts behavioral slope ($r = 0.88$, $p = 0.002$; raw $r = 0.91$).

A mechanistic dissociation between base and instruct models. Several base models exhibit behavioral amplification despite near-zero representational geometry (Qwen 14B Base: slope 1.28, Mantel $r = 0.13$; Llama 8B Base: slope 1.00, Mantel $r = -0.06$). This suggests base and instruct models achieve cross-scale inference through qualitatively different mechanisms, with instruction tuning leaving a geometric signature in mid-layer representations that the base pathway does not.

7. Discussion

Amplification and representation-level structure. Our central finding is that representational amplification varies less than behavioral amplification across models (1.38–1.77 vs 0.53–1.53). Large instruct models preserve most of their representational amplification during readout, whereas smaller models exhibit substantial attenuation. A synthetic pipeline control (§5.3) further shows that, after subtracting the ridge-probe baseline, the adjusted slope remains scale-dependent (-0.09 to $+0.66$) and continues to predict behavioral amplification ($r = 0.88$). These results suggest that at least part of the amplification is already present in learned representations and is not solely a decoding artifact, with readout alignment emerging as the dimension most sensitive to model scale and instruction tuning.

Relation to prior interpretations. Liu et al. (2026) interpret structural amplification as a form of noise filtering, where unilateral attenuation correction should bring slopes toward 1.0. Our results complicate this account. Corrected behavioral slopes remain above 1.0 for models $\geq$ 7B (1.18–1.29), whereas Liu et al. report clustering near 1.0. One potential source of divergence is methodological: our rotating-persona design spans the full bidirectional correlation structure across constructs, whereas their single-source design projects from Big Five inputs onto other scales. The residual amplification we observe may therefore arise from

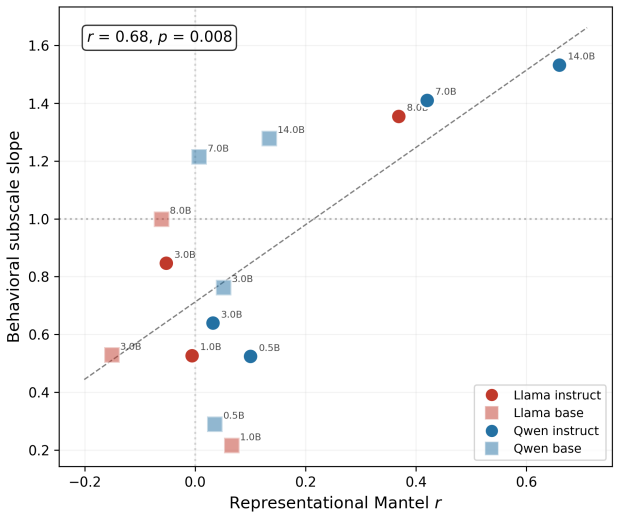

Figure 6. Representational structure predicts behavioral amplification across 14 models ($r = 0.68$, $p = 0.008$; LOO: [0.58, 0.74]). Circles = instruct, squares = base; red = Llama, blue = Qwen; labels indicate model size.

correlations not directly constrained by a Big Five input. In addition, activation-level slopes exceed 1.0 even in smaller models whose behavioral outputs are attenuated below 1.0, indicating that behavioral attenuation alone cannot account for the absence of amplification at smaller scales.

Slope amplification and baseline shifts. Our slope-based analyses do not exhaust the ways LLM correlations differ from human correlations. Appendix H shows a systematically positive OLS intercept: subscale pairs that are near zero in the human matrix tend to receive mildly positive predicted correlations. This baseline shift does not show the same clean scaling trend as slope amplification, and may reflect response-style effects, priors introduced by the persona prompt, or residual lexical/semantic co-occurrence. We therefore treat slope amplification and intercept inflation as separable distortions requiring separate controls.

Emergence, instruction tuning, and implications. Instruction tuning appears to have two separable effects. First, it is associated with the emergence of organized directional geometry in activation space (Mantel r increasing from 0.13 to 0.66 in Qwen 14B base vs. instruct), which is not observed in base models. Second, it reduces readout attenuation: base and instruct models exhibit similar representational slopes but differ in behavioral expression. Extending this comparison across seven base–instruct pairs clarifies the pattern: instruction tuning reliably improves behavioral alignment at all scales (paired $p = 0.011$), whereas improvements in representational geometry are observed only above $\approx$ 7B. Smaller instruct models therefore show behavioral gains without corresponding geometric reorganization. This pattern aligns with the dissociation observed in base models (§6). Base models appear to encode linearly decodable psychometric information (representational slope $\approx$ 1.6–1.7) without organized directional geometry. Amplified cross-scale structure thus requires only linear decodability, whereas instruction tuning additionally introduces directional organization that supports more faithful readout. For LLM-based psychological simulation, this implies that behavioral outputs may underestimate the degree of amplification present in internal representations. More broadly, psychometric structure provides a quantitative and population-grounded testbed for examining how structured representations emerge in LLMs.

Why instruction tuning aligns with personality structure. This mechanistic account raises a further question: why does instruction tuning align representations with *human* personality structure, despite not explicitly targeting such constructs? One possibility is incidental trait alignment. Signals related to human preference and role consistency, which are abundant in instruction-tuning data, may organize latent personality-relevant representations already present in the base model into directions that happen to align with human trait structure. Under this view, structured trait representations may emerge without explicit supervision, as a byproduct of optimization toward other objectives.

Limitations. Our human data are drawn from a single cultural sample (Taiwanese college students), whereas LLMs are trained predominantly on English text. This mismatch raises the possibility that some observed amplification reflects cross-linguistic differences in psychometric structure rather than purely representational effects. An exploratory post-hoc analysis (App D) suggests that culturally specific subscales (e.g., POM, SCS Indep/Interdep) exhibit stronger amplification, although this effect is driven largely by a small number of subscale pairs and requires replication. Importantly, amplification is not confined to culturally specific constructs: within culturally stable subscales alone, $\geq$ 3B instruct models still exhibit slopes above 1.0. Moreover, our core findings rely on within-model differences and cross-model rank ordering, which are less sensitive to such mismatches. Our steering analyses support a qualitative interpretation that contrastive directions are functionally accessible, but do not establish a quantitative relationship between direction strength and behavioral effects. Contrastive direction reliability is also limited in smaller models, constraining interpretation of null geometry results. Differences across pipelines (ridge vs. argmax) are addressed through discretization controls (§5.2) and synthetic pipeline controls (§5.3), which bound potential artifacts. Finally, the SNR proxy is imperfect, and adjusted slope magnitudes should be interpreted as approximate. Nonetheless, the qualitative pattern—namely, that observed amplification reflects both

a pipeline baseline and a scale-dependent representational component—remains robust.

8. Conclusion

We trace psychometric inference in LLMs from behavioral outputs to internal representations. Structured psychometric geometry in activation space emerges with model scale and instruction tuning and predicts behavioral amplification across models (within-instruct $r = 0.90$; partial $r = 0.69$ controlling for log size), while steering provides suggestive evidence that these directions are causally accessible. Representational and behavioral amplification both increase with model size but at different rates: representational slopes span 1.38–1.77, whereas behavioral slopes span 0.53–1.53. A synthetic pipeline control suggests that this range includes a representation-level component (-0.09 to $+0.66$) beyond a ridge-probing baseline of approximately 1.3, with adjusted slopes continuing to predict behavior ($r = 0.88$). Together, these findings provide a representation-level account of how LLMs capture and amplify human psychological structure, with readout fidelity emerging as the dimension most sensitive to model scale and instruction tuning.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pp. 337–371, 2023.

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

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. 2023.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Choi, H. K. and Li, Y. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. *arXiv preprint arXiv:2405.02501*, 2024.

Cortina, J. M. What is coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology*, 78(1):98–104, 1993.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

Dong, W., Zhao, Y., Sun, Z., Liu, Y., Peng, Z., Zheng, J., Zhang, Z., Zhang, Z., Wu, J., Wang, R., et al. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *arXiv preprint arXiv:2505.00049*, 2025.

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2024.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *International Conference on Learning Representations*, 2024.

Huang, W., Song, R., Chen, Y., Wang, Z., and Qi, Y. Quantifying ai psychology: A psychometrics benchmark for large language models. In *International Conference on Learning Representations*, 2024.

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., and Kabbara, J. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3605–3627, 2024.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

Liu, Y.-F., Lu, Y.-L., He, D., and Zhang, H. From five dimensions to many: Large language models as precise and interpretable psychological profilers. In *International Conference on Learning Representations*, 2026.

Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2): 209–220, 1967.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Nunnally, J. C. Psychometric theory—25 years ago and now. *Educational Researcher*, 4(10):7–21, 1975.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024a.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2024b.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., and Strohmaier, M. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024.

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

Samuel, V., Zou, H. P., Zhou, Y., Chaudhari, S., Kalyan, A., Rajpurohit, T., Deshpande, A., Narasimhan, K., and Murahari, V. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024.

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., and Matarić, M. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

Sorokovikova, A., Fedorova, N., Rezagholi, S., and Yamshchikov, I. P. Llms simulate big five personality traits: Further evidence. *arXiv preprint arXiv:2402.01765*, 2024.

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., and Neubig, G. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2024.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Prompt Examples

Single-scale system prompt (behavioral experiment and cross-persona probing).

```
You are completing psychological questionnaires as a person
with the following responses on the [Scale Name]:
[instruction]
Response scale:
  1 = [label] ... 5 = [label]

Q1. [item text] -> [response]
Q2. [item text] -> [response]
...

Respond to all questions as this person would, maintaining
consistency with this psychological profile. Answer with just
the number corresponding to your choice.
```

Contrastive extraction system prompt (mega-prompt). Same format but includes all 7 scales (114 items). The target subscale's items are set to maximum (high) or minimum (low); all other items are randomized.

Target item prompt.

```
[instruction]

Statement: [item text]

Response options:
  1 = [label] ... 5 = [label]

Please respond with a single number from 1 to 5.
```

B. Robustness Analyses

Bootstrap stability and split-half baseline for alignment metrics. Bootstrap resampling (2000 iterations, subject-level) of the human 16×16 subscale correlation matrix yields mean pairwise CI width of 0.25 (median 0.25, max 0.29). The overall structure is highly stable ($r = 0.93$ between original and bootstrap samples; 95% CI: $[0.89, 0.96]$).

To establish a reference scale for our LLM-vs-human alignment metrics at the subscale level (the unit of analysis throughout this paper, §3), we additionally compute a split-half human baseline: 2000 random partitions of the 272 subjects into two halves, each yielding an independent 16×16 subscale correlation matrix, with the alignment statistics (r, OLS slope, OLS intercept) computed between the two halves. This distribution characterises how closely two independent human samples reproduce each other under the same metric, and provides a natural reference against which to compare LLM-vs-human alignment (Figure 7). Three observations follow. (i) On alignment r, our 8B/14B base LLM matrices fall in the left tail of the human split-half distribution. LLM cross-scale predictions therefore approach but do not yet reach the level of agreement that two independent halves of the human sample show with one another, framing the LLM alignment numbers reported in the main text as substantial but sub-ceiling. (ii) On OLS slope, the same models exceed the bulk of the human distribution. The amplification claim of §3 is thus not relative to a degenerate or trivially noisy baseline: LLM slopes are large precisely in the dimension where the human reference is itself centred near 1.0 by construction. (iii) On intercept, both LLM models lie beyond the right tail of the human distribution. We discuss this intercept offset, which is independent of the slope amplification, in App H (Intercept inflation).

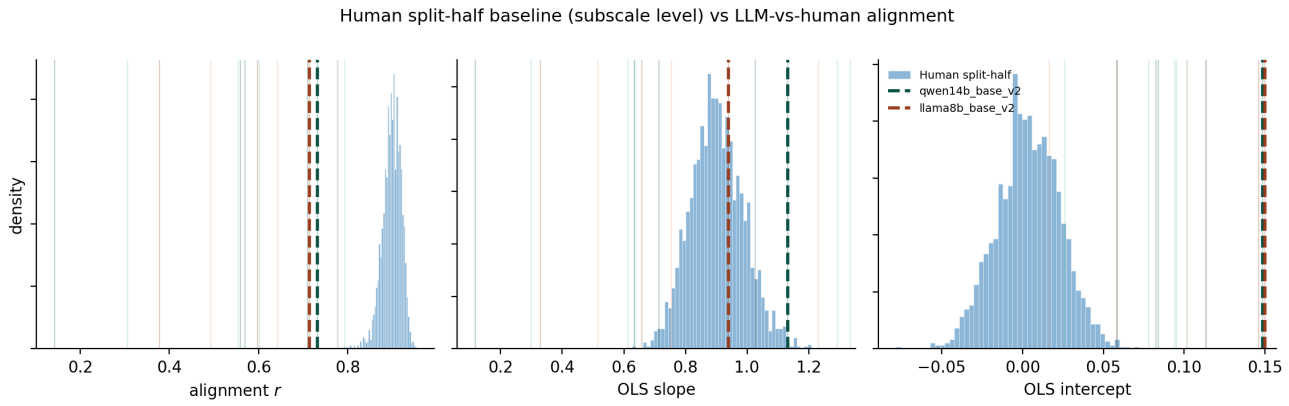

Figure 7. Human split-half baseline at the subscale level (2000 random splits of $N = 272$, blue histograms) versus two LLM base models (Qwen 14B Base, green; Llama 8B Base, red), for alignment r (left), OLS slope (middle), and OLS intercept (right). LLMs fall *below* the human distribution on r, *above* it on slope, and *above* it on intercept. The three panels correspond to three independent components of the LLM-vs-human relationship, and the LLM offsets are signed differently across them.

Leave-one-out analysis on structure–behavior correlation. The $r = 0.68$ correlation between representational Mantel r and behavioral slope (Figure 6) is robust to removal of individual models. LOO range: [0.58, 0.74]; LOO mean: 0.68. The most influential point is Qwen 14B Instruct (removal reduces r to 0.58).

Pipeline sensitivity analysis (Llama 8B Instruct). Three representational pipeline variants are compared:

Pipeline	Slope	Mantel r
Ridge $\rightarrow$ subscale score (continuous)	1.44	0.82
Ridge $\rightarrow$ item (continuous) $\rightarrow$ subscale mean **[main]**	1.70	0.83
Ridge $\rightarrow$ item $\rightarrow$ round to integer $\rightarrow$ subscale mean	0.83	0.78
Behavioral (argmax integer $\rightarrow$ subscale mean)	1.35	—

The item-level continuous pipeline (second row) is used as the main representational pipeline throughout the paper, as it provides an apples-to-apples comparison with the behavioral pipeline (both aggregate item-level responses to subscale scores via averaging).

Discretization across all 9 models. The continuous $\rightarrow$ rounded-integer slope reduction reported above on Llama 8B Instruct ($1.70 \rightarrow 0.83$) is consistent across the full 9-model panel. Continuous slopes range 1.38–1.77 (mean 1.60); rounded slopes range 0.47–0.97 (mean 0.71); within-model differences range $+0.80$ to $+0.97$. Paired t-test on the 9-model paired differences yields $p < 10^{-4}$ (Wilcoxon $p = 0.004$). Mantel r, being rank-based and scale-invariant, is largely preserved after discretization and in two models (Qwen 0.5B Instruct, Llama 1B Instruct) marginally increases, consistent with continuous predictions having inflated tail variance at small human correlations.

Ridge regularization sensitivity. On Llama 8B Instruct (layer 16), we vary the ridge regularization parameter λ across seven orders of magnitude. The item-level pipeline slope remains above 1.5 for $\lambda \in [1, 100]$ (the range selected by cross-validation), peaking at 1.81 ($\lambda = 100$) and declining to 0.98 at $\lambda = 0.01$ (under-regularized) and 0.44 at $\lambda = 10{,}000$ (over-regularized). The main result is not sensitive to the specific regularization strength within the cross-validated range (Figure 8).

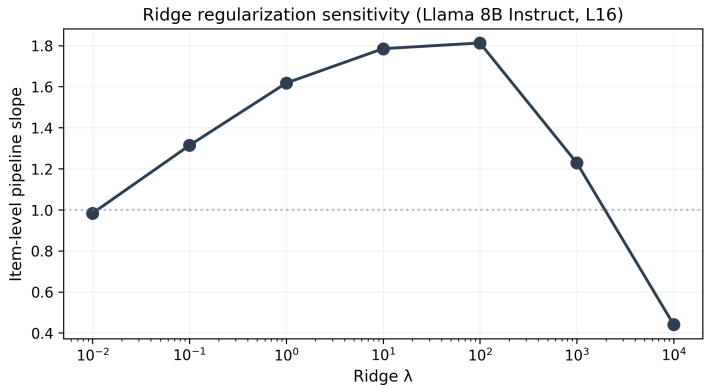

Figure 8. Ridge regularization sensitivity on Llama 8B Instruct (layer 16). Item-level pipeline slope as a function of ridge λ (log scale), exhibiting the expected inverted-U: underfitting at $\lambda \to 0$ (OLS-like, activation noise overfit) and overfitting-to-zero at $\lambda \to \infty$ (weights shrunk to zero). The CV-selected range ($\lambda \in [1, 100]$) falls on the plateau, where slope remains above 1.5. The main result is therefore robust to the specific regularization choice within the range that cross-validation actually selects.

Random features baseline. We shuffle the subject–activation mapping within each rotation (breaking the correspondence between persona identity and activation while preserving the marginal distribution of activations) and recompute the item-level pipeline. Over 10 shuffled iterations on Llama 8B Instruct, the mean slope is 0.25 ± 0.02, compared to 1.70 for real activations, a $6.7\times$ difference. This confirms that the representational slope reflects genuine subject-specific psychometric information in activations, ruling out pipeline-only explanations in which the slope is independent of subject-specific activation structure.

Layer selection and robustness. For each model, the "best layer" reported in Table 2 was selected post-hoc from a per-model layer-wise sweep on the plateau region of high Mantel r and slope. To assess whether this selection materially affects reported slopes, we examined slope variation across all 14 models at layers immediately adjacent to the selected peak. The mean ± 1-layer slope spread across models is 0.025 (max 0.118, observed for Qwen 7B Instruct); the ± 2-layer spread is 0.043 (max 0.137). Both are an order of magnitude smaller than the cross-model representational spread (0.39). Post-hoc layer selection therefore contributes negligibly to cross-model slope differences.

Layer-wise slope trajectory. The item-level pipeline slope varies across layers for all nine models, generally increasing from early layers (≈ 1.3–1.5) to a peak in mid-to-late layers (≈ 1.65–1.77) before declining slightly in the final extracted layers (Figure 9). The peak layer (selected for Table 2) corresponds to approximately 50–70% of model depth across architectures. This trajectory is consistent across model sizes and families, with base and instruct models of the same size showing similar layer-wise profiles. Notably, the slope remains elevated (> 1.5) through the final extracted layer in all models, indicating that most readout attenuation occurs during the unembedding and argmax steps rather than through gradual decay across late transformer layers.

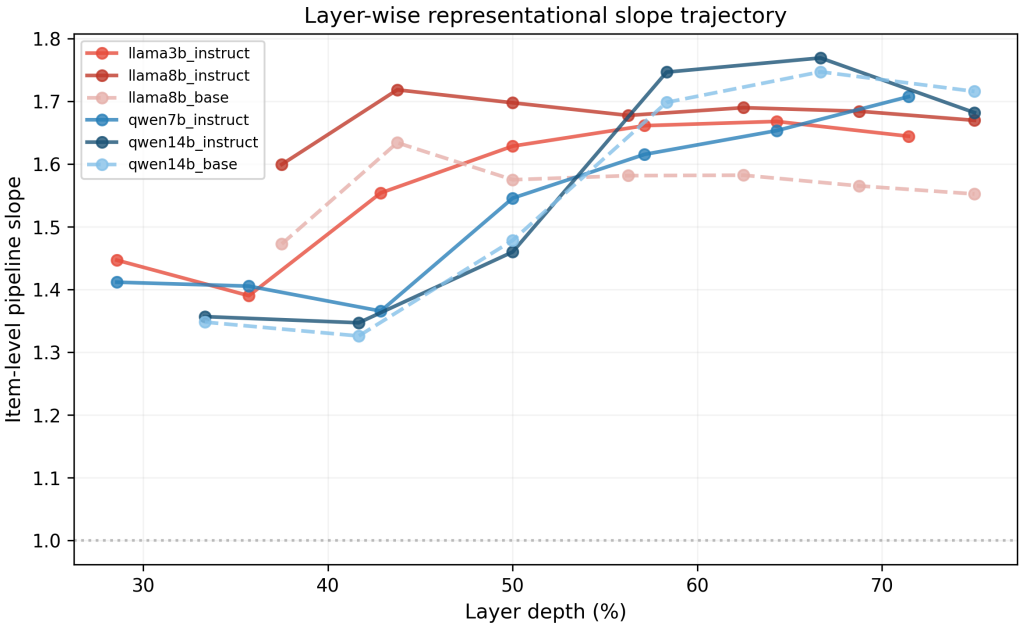

Figure 9. Item-level pipeline slope as a function of relative layer depth (layer index divided by total layers) across the 9 cross-persona models. Circles mark each model's "best layer" reported in Table 2. Trajectories peak at approximately 50–70% depth and decline gradually toward the final layers; cross-model spread at any fixed relative depth is smaller than the cross-family variation in best-layer indices, supporting the claim that post-hoc layer selection contributes negligibly to cross-model slope differences.

Within-group and partial correlations. The $r = 0.68$ structure–behavior correlation (§6) was decomposed to test scale and instruct-status confounds. Within the 7 instruct models, the correlation is $r = 0.90$ ($p = 0.006$); within the 7 base models, $r = 0.17$ ($p = 0.71$); Fisher-z comparison $p = 0.064$. The partial correlation controlling for log model size is $r_{\text{partial}} = 0.69$ (versus zero-order $r = 0.68$), confirming that the relationship is not mediated by scale. Additional control for family yielded $r_{\text{partial}} = 0.55$, retaining a substantial effect size.

C. Trivial-Context Control

In the cross-persona paradigm, a subject's item-level responses on the persona scale are placed in the prompt. For target subscales from the *same* scale as the persona ("within-scale" subscale pairs), direct supporting context is therefore available to the model; for subscales from a different scale ("cross-scale" pairs), no direct context exists. A natural concern is that alignment on within-scale pairs could reflect trivial prompt copying rather than structured inference, inflating the overall alignment metrics.

To test this, we partition the cross-persona predicted correlation matrices of all 9 cross-persona models into within-scale and cross-scale pairs, and compute the mean absolute deviation $|r_{\text{pred}} - r_{\text{human}}|$ for each group. Pooling across models, within-scale pairs yield mean $|r_{\text{pred}} - r_{\text{human}}| = 0.293$ (162 pairs) vs. cross-scale $= 0.284$ (918 pairs); the ratio is 1.030, indistinguishable from no effect. Paired per-model testing yields t-test $p = 0.24$, Wilcoxon $p = 0.25$. We conclude that within-scale prompt context does not systematically improve alignment over cross-scale pairs, ruling out trivial copying of in-prompt responses as a driver of the observed structure.

D. Cultural Specificity Exploratory Analysis

As an exploratory check on whether the observed amplification is uniform across culturally stable versus culturally specific constructs, we classify subscales based on prior cross-cultural literature: *stable* = Big Five E/A/C/N, IRI EC/PT, PANAS NA, SWLS (8 subscales); *specific* = POM, SCS Indep/Interdep (3 subscales); remaining subscales (IRI Fantasy and Personal Distress, PANAS PA, BFI Openness, UCLA Loneliness) are marked ambiguous and excluded from the contrast because cross-cultural evidence on their replicability in East Asian samples is mixed. For each of 9 models with cross-persona data, we compute OLS slope and Mantel r on the subset of subscale pairs within each group.

Paired comparisons indicate that culturally specific subscales exhibit *greater* slope amplification than stable subscales (mean Δslope $= -0.42$, paired t-test $p < 10^{-4}$, Wilcoxon $p = 0.004$; the Mantel r difference is similarly signed, $\Delta = -0.25$, paired t-test $p < 10^{-5}$). The direction survives restricting to cross-scale pairs only ($\Delta = -0.73$, $p < 10^{-4}$) and using the magnitude-invariant Mantel r instead of slope.

However, leave-one-pair-out analysis on the three within-specific pairs reveals that this contrast is driven substantially by a single pair: POM $\times$ SCS Independent self (mean $|\Delta$slope$|$ across 9 models $= 1.71$). The other two within-specific pairs are SCS Indep $\times$ SCS Interdep (mean $|\Delta$slope$| = 0.45$) and POM $\times$ SCS Interdependent self (mean $|\Delta$slope$| = 0.12$). The small subscale count in the specific group (3 subscales $\rightarrow$ 3 within-group pairs) limits statistical generalization.

We interpret this finding cautiously: the observation that LLMs over-amplify correlations among culturally specific constructs is consistent with stereotyped reliance on English-language training content when encountering culturally distinctive Chinese scales, but the single-pair sensitivity and small n preclude strong claims. Cross-cultural replication with English-language scales and expanded construct sets is needed to confirm this as a general property.

E. Synthetic Pipeline Control Details

Item-level synthetic pipeline. For items-per-subscale $\in \{3, 5, 7\}$ and target Cronbach's $\alpha \in \{0.60, 0.75, 0.85\}$, the 20-seed recovered slope (vs. sample latent correlation) ranges from 1.17 to 1.43 (Table below). The mean slope across all $9 \times 20 = 180$ runs is 1.289 (std 0.096). Pipeline inflation decreases as items-per-subscale and α increase, consistent with higher measurement precision yielding lower ridge regularization and thus less information sharing across items.

items/subscale	$\alpha = 0.60$	$\alpha = 0.75$	$\alpha = 0.85$
3	1.430 ± 0.007	1.394 ± 0.005	1.196 ± 0.006
5	1.343 ± 0.006	1.259 ± 0.003	1.231 ± 0.002
7	1.380 ± 0.007	1.193 ± 0.003	1.172 ± 0.002

High-dimensional SNR sweep. For the activation-level synthetic pipeline, we generated latents from the human correlation matrix, projected through a random Gaussian matrix $\mathbf{W} \in \mathbb{R}^{d \times 16}$ with $d \in \{896, 2048, 4096, 5120\}$, and added isotropic Gaussian noise at SNR $\in \{0.1, 0.3, 1.0, 3.0, 10.0\}$. Over 400 runs ($4 \times 5 \times 20$ seeds), recovered slope depends on SNR with minimal dependence on d within our range:

SNR	$d = 896$	$d = 2048$	$d = 4096$	$d = 5120$
0.1	1.39	1.48	1.54	1.56
0.3	1.32	1.30	1.31	1.31
1.0	1.22	1.17	1.14	1.14
3.0	1.16	1.10	1.08	1.07
10.0	1.11	1.07	1.04	1.03

Mantel r between recovered and human matrices is consistently ≥ 0.97 for all cells, confirming that the pipeline recovers correlation structure faithfully regardless of magnitude inflation.

Per-model adjustment. Pipeline baselines for each of the 9 cross-persona models, obtained by interpolating the synthetic grid at each model's hidden dimension and SNR proxy ($\rho/(1-\rho)$), range from 1.11 (Qwen 14B Instruct) to 1.48 (Qwen 3B Instruct). Adjusted slopes (observed $-$ baseline) range from -0.09 (Qwen 3B Instruct) to $+0.66$ (Qwen 14B Instruct); mean adjusted slope is $+0.27$. Correlations of adjusted slope with behavioral slope ($r = 0.88$, $p = 0.002$) and log model size ($r = 0.68$, $p = 0.045$) remain substantial.

SNR proxy caveat. The SNR proxy used for per-model adjustment is derived from the contrastive-directions pipeline's empirical reliability (`mean_reliability` from §4), not from a reliability estimate of the item-level pipeline itself. Proper correction would require split-half reliability of item-level subscale predictions, which requires re-running the full item-level pipeline across all rotations with predictions saved. The reported baseline and adjusted slopes should therefore be interpreted as approximate; given the relatively smooth shape of the high-dimensional SNR sweep in the relevant range (SNR $\in [0.1, 1.4]$), reasonable shifts in the SNR proxy are expected to produce small shifts in the adjusted values but preserve model ordering and the overall scale-dependent pattern.

Table 3. Complete subscale composition and reverse-coded items per scale.

Scale (range)	Subscale	Items (1-indexed)	Reverse-coded (within scale)
IRI (1–5)	Perspective Taking	3, 8, 11, 15, 21, 25, 28	3, 4, 7, 12, 13, 14, 15, 18, 19
	Fantasy	1, 5, 7, 12, 16, 23, 26	
	Empathic Concern	2, 4, 9, 14, 18, 20, 22	
	Personal Distress	6, 10, 13, 17, 19, 24, 27	
PANAS (1–5)	Positive Affect	1, 3, 5, 9, 10, 12, 14, 16, 17, 19	(none)
	Negative Affect	2, 4, 6, 7, 8, 11, 13, 15, 18, 20	
POM (1–6)	Peace of Mind	1–7 (all)	5, 7
BFI-10 (1–7)	Extraversion	1, 6	2, 6, 8, 9, 10
	Agreeableness	2, 7	
	Conscientiousness	3, 8	
	Neuroticism	4, 9	
	Openness	5, 10	
SCS (1–7)	Independent self	1, 6, 7, 9, 11, 12, 15, 17, 20, 21, 22, 23	(none)
	Interdependent self	2, 3, 4, 5, 8, 10, 13, 14, 16, 18, 19, 24	
SWLS (1–7)	Life Satisfaction	1–5 (all)	(none)
UCLA (1–4)	Loneliness	1–20 (all)	1, 5, 6, 9, 10, 15, 16, 19, 20

F. Scale Details

Table 3 lists all 7 scales, their subscale composition, item numbers (1-indexed, matching the original Chinese-language instrument item orderings), and reverse-coded items. Response ranges vary by scale. All 114 items are administered in the Chinese translation used with our $N = 272$ participants. Reverse-coded items are un-reversed (via $v' = \text{max_val} + 1 - v$) prior to any scoring or analysis; subscale scores are the mean of the un-reversed item responses within the subscale.

G. Cross-Persona Details

Activation extraction. For the single-scale activation probing paradigm (§5), we hook the post-MLP residual-stream output of each transformer layer (i.e., the output of `model.model.layers[i]`, equivalent to the input to layer $i{+}1$) and record the activation at the *last token position* of the prompt under each (subject, rotation, target-item) combination. All prompts use raw-text concatenation (persona system prompt ‖ target-item question ‖ "`Answer: `")[1] rather than the model's chat template, matching the behavioral pipeline and ensuring comparability between base and instruct variants. A candidate layer set is recorded per model (scaled with depth: layers 12–24 for Llama 8B; analogous mid-to-late ranges for other sizes); the "best layer" in Table 2 is selected post-hoc from a per-model layer-wise sweep on the plateau region of high Mantel r and slope, with ties broken in favour of higher Mantel r. Appendix B reports layer-wise sensitivity.

Directional asymmetry. The LLM cross-scale correlation matrix is not symmetric: cell (i, j) and (j, i) come from different persona rotations (Scale i as persona vs. Scale j as persona). We compute the asymmetric component of both the activation-predicted and behavioral correlation matrices and correlate them. On Llama 8B Instruct, the activation asymmetry pattern correlates with the behavioral asymmetry pattern at $r = 0.81$ ($p < 0.0001$); on Qwen 14B Base, $r = 0.74$ ($p = 0.0001$). This indicates that the directional asymmetries in cross-scale inference (e.g., personality predicting affect better than affect predicting personality) are already encoded in mid-layer representations and faithfully expressed in behavioral output.

H. Additional Behavioral Results

Cronbach's alpha for attenuation correction. The attenuation-corrected slopes shown in Figure 2 (dotted lines) are computed by disattenuating the human correlation matrix only (unilateral correction; the LLM matrix is left untouched, as

[1]The trailing space after the colon is intentional. The BPE tokenizers used by Llama and Qwen treat the response digit as its own token only if separated from the colon by whitespace; without the trailing space, the colon and digit can fuse into a single token, which would invalidate the option-likelihood scoring described in §3.

Table 4. Subscale-level disattenuated slope for $\geq$ 7B models as a function of the minimum-reliability cut. As more low-α subscales are excluded, the corrected slope rises monotonically and converges on the scale-level estimate.

Cut	$n_{\text{subscales}}$	n_{pairs}	$\geq$ 7B range	$\geq$ 7B mean
All 16 subscales	16	112	$[0.62, 0.88]$	0.76
$\alpha \geq 0.60$	12	64	$[0.78, 1.06]$	0.89
$\alpha \geq 0.70$	9	36	$[0.80, 1.04]$	0.91
$\alpha \geq 0.80$	5	10	$[1.19, 1.36]$	1.28
Scale-level (Fig. 2)	7	21	$[1.18, 1.29]$	1.23

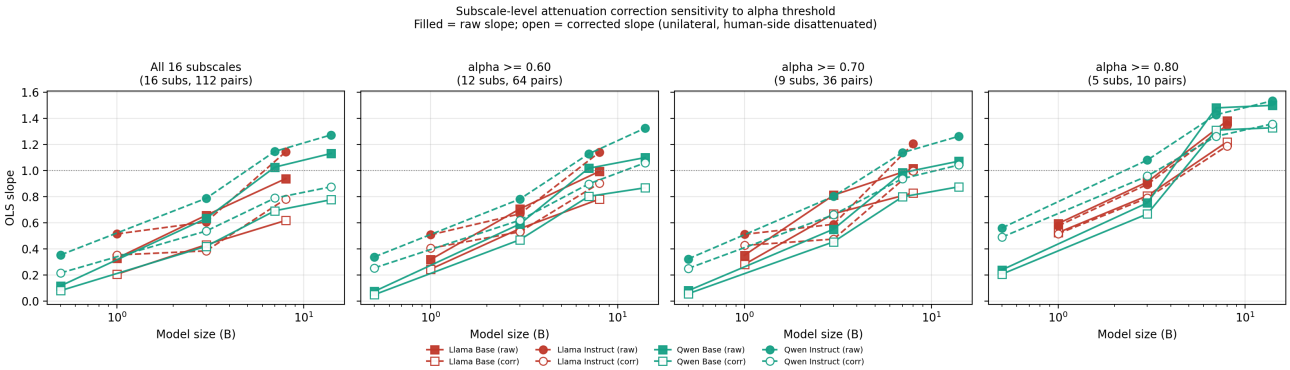

Figure 10. Subscale-level attenuation correction sensitivity to the minimum-α cut. Filled markers: raw slope; open markers: corrected slope (unilateral, human-side disattenuated). At the most permissive cut (all 16 subscales) the four BFI-10 short subscales pull the corrected slope downward; restricting to $\alpha \geq 0.80$ removes them and recovers the scale-level 1.18–1.29 range.

the LLM has no measurement-error reliability analogue at the subject level). Cronbach's α is estimated directly from the $N = 272$ Taiwanese participant responses per subscale using the standard formula $\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_i \sigma_i^2}{\sigma_T^2}\right)$, where k is the number of items, σ_i^2 is the variance of item i, and σ_T^2 is the variance of the total subscale score. Disattenuated correlations are $r_{\text{true}}(x, y) = r_{\text{obs}}(x, y)/\sqrt{\alpha_x \alpha_y}$; values that cross ± 1 after disattenuation are clipped. The dotted lines in Figure 2 use scale-level reliability estimates: one Cronbach's α is computed for each of the seven scales from all items in that scale (e.g., 28 items for IRI, 20 for PANAS, 7 for POM). The human scale-level correlation matrix is then disattenuated unilaterally using these seven scale reliabilities.

Subscale-level sensitivity to reliability cut. Computing the correction at the subscale level (16 subscales) instead of the scale level exposes a sensitivity to short-instrument reliability. Subscale-level α values range from 0.30 (BFI Agreeableness, $k = 2$) to 0.92 (POM Peace of Mind, $k = 7$); the four BFI-10 subscales with two items each (Agreeableness, Conscientiousness, Neuroticism, Openness) yield $\alpha \in [0.30, 0.60]$, well below the conventional reliability bar. Cronbach's α is known to be unstable when computed on $k = 2$ items (Cortina, 1993); BFI-10 was designed to use test-retest rather than internal-consistency reliability for this reason (Rammstedt & John, 2007). Disattenuating with these underestimated α values inflates the human correlations more than warranted, suppressing the corrected LLM-vs-human slope. Sweeping a minimum-α threshold and recomputing corrected slopes for $\geq$ 7B models gives Table 4: as more low-α subscales are excluded, the corrected slope rises monotonically and converges on the scale-level estimate. Restricting to $\alpha \geq 0.80$ subscales (the cut at which Cronbach's estimates are themselves reliable) reproduces the scale-level range of 1.18–1.29 within rounding. The convergence of two independent restrictions (scale-level aggregation and high-α subscale subset) on the same range supports the §3 claim that amplification persists after disattenuation, while making explicit that the conclusion depends on having reliability estimates that are themselves trustworthy.

Per-scale alignment. Table 2 reports alignment averaged across all 16 subscales. Breaking this down per scale reveals that most scales show similar cross-scale alignment across models, with IRI (empathy) and BFI (personality) typically showing higher alignment than POM (peace of mind) on smaller models, consistent with the former constructs having richer English-language literature in LLM training data. Per-scale numbers are in the project repository (`per_scale_alignment.csv`); they do not materially change the main conclusions.

Intercept inflation as a separable distortion. The amplification analyses in §3 characterise LLM cross-scale correlations through their slope when regressed onto human correlations. OLS, however, has a second free parameter, the intercept, which captures a baseline offset: the predicted LLM correlation at $r_{\text{human}} = 0$. We report this subscale-level intercept across all 14 models in Figure 11. Two patterns are notable. First, the intercept does not exhibit a clean scaling trend: it does not decrease with model size for either family. Second, intercepts are systematically positive across models (median across the 14 models: 0.09), and the two base models for which we computed split-half baselines (App B) lie beyond the right tail of the human split-half intercept distribution. LLMs therefore appear to introduce a positive baseline shift on top of the slope-based amplification documented in the main text: even subscale pairs that are uncorrelated in the human data tend to be predicted with mildly positive correlations.

We do not interpret this as a separate causal mechanism on the basis of the present data. Possibilities include a generic acquiescence-style response style that is incompletely cancelled by reverse-coding, a mild positive prior over correlations introduced by the in-context persona format, or residual influence of token-level co-occurrence in the prompt. Distinguishing these would require a controlled probe (e.g., correlations between truly orthogonal constructs in the prompt, or randomised-persona controls) that we did not run. We report the intercept here so that readers interpreting the slope numbers as a complete summary of LLM-vs-human distortion are aware that a non-trivial intercept component coexists with the slope effect and is not absorbed by it.

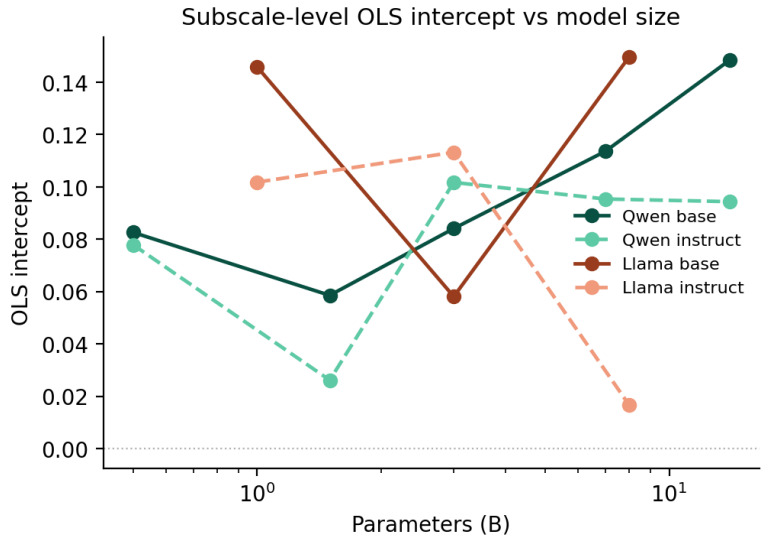

Figure 11. OLS intercept of LLM-vs-human subscale-level cross-scale correlation regression as a function of model size. Colour indicates family (red: Llama, green: Qwen) and base/instruct as marked in the legend. The intercept does not exhibit a monotone scaling trend, is predominantly positive across models, and indicates a baseline offset in LLM-predicted correlations that is separable from the slope-based amplification reported in §3.

Behavioral semantic similarity control. A potential concern for the representational-geometry results (App I) is that the apparent structure might be inherited from surface lexical similarity between item wordings. The same concern applies in principle to the behavioral pipeline; if LLMs match human structure, perhaps they are matching item-word similarity instead. We rerun the same control here: compute item-text embedding similarity, average within subscale to a 16×16 semantic similarity matrix S, and partial it out of the behavioral $r(B, H)$ alignment. We report two embedding sources for robustness: (a) each model's own input embedding layer (mean-pooled across item tokens), the same source used in App I; (b) the multilingual-e5-large sentence encoder, an external semantic encoder shared across all 14 models. Across 14 models, both sources yield mean partial correlations very close to the zero-order alignment (mean drop 0.002 for source (a), 0.008 for source (b); maximum absolute drop across the 14 models is 0.013 for (a) and 0.036 for (b)). Behavioral alignment is therefore not explained by item-text semantic similarity, regardless of which embedding is used as the control. Per-model values are in the code release.

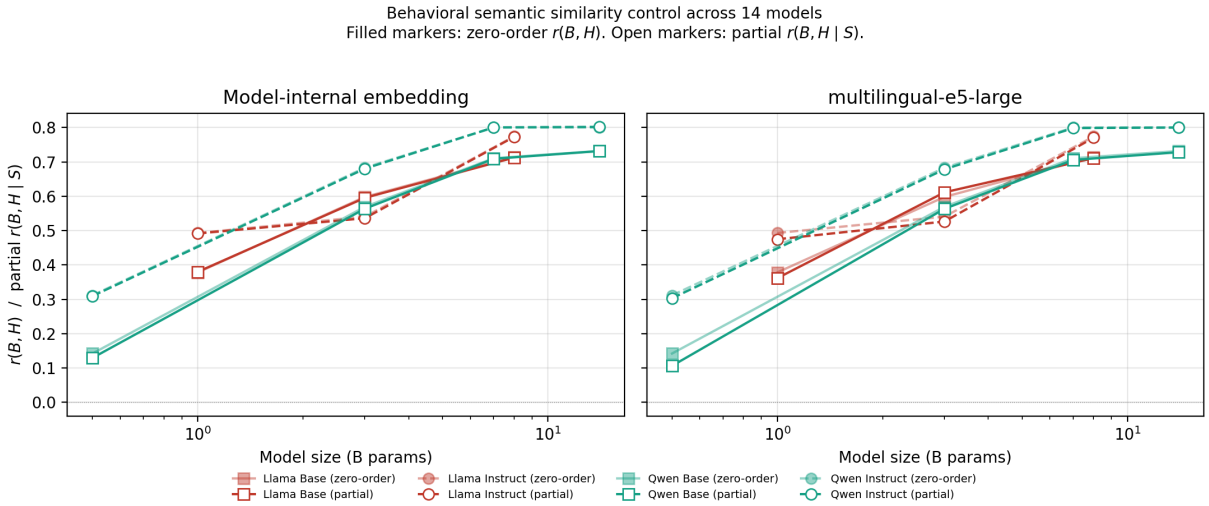

Figure 12. Behavioral semantic similarity control across 14 models. Filled markers: zero-order $r(B, H)$. Open markers: partial $r(B, H \mid S)$ controlling for item embedding similarity S. Left: each model's input embedding layer; right: multilingual-e5-large. Partial and zero-order values are visually indistinguishable for all models in both panels.

I. Representational Geometry Details

Qwen 14B Instruct geometry: visual comparison. Figure 13 makes the Mantel $r = 0.66$ reported in §4 concrete. The human subscale correlation matrix (left) and Qwen 14B Instruct's activation cosine similarity matrix (right) show matching block structure: within-scale subscale groupings (e.g., IRI PT/FS/EC/PD; Big Five E/A/C/N/O) cluster in both, and cross-scale relationships such as positive PA–LifeSat and negative NA–Loneliness appear in both. Qwen 14B Base, which shares pretraining but differs only in instruction tuning, shows Mantel $r = 0.13$ and no comparable block structure (matrix omitted; available in the code release).

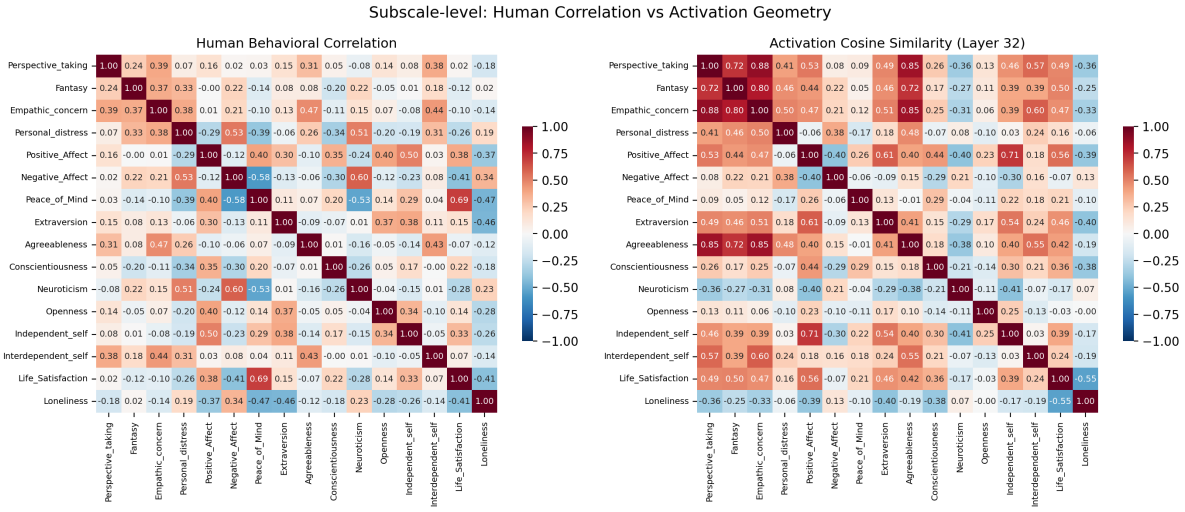

Figure 13. Human subscale correlation matrix (left) vs. Qwen 14B Instruct activation cosine similarity matrix (right, layer 32). Mantel $r = 0.66$. Visible block structure corresponds to within-scale subscale groupings (e.g., IRI PT/FS/EC/PD; Big Five E/A/C/N/O) and cross-scale relationships (e.g., positive PA–LifeSat, negative NA–Loneliness). The corresponding Qwen 14B Base matrix (Mantel $r = 0.13$) shows no comparable block structure.

Emergence of geometry with scale. Figure 14 visualises the scaling curve reported in Figure 3. Qwen 3B Instruct (Mantel $r = 0.03$) shows near-random pattern; Qwen 7B Instruct ($r = 0.42$) shows partial block structure; Qwen 14B Instruct ($r = 0.66$) shows clean alignment with the human correlation matrix. Geometry emergence is continuous rather than a sharp threshold, but the rate of emergence is rapid between 7B and 14B.

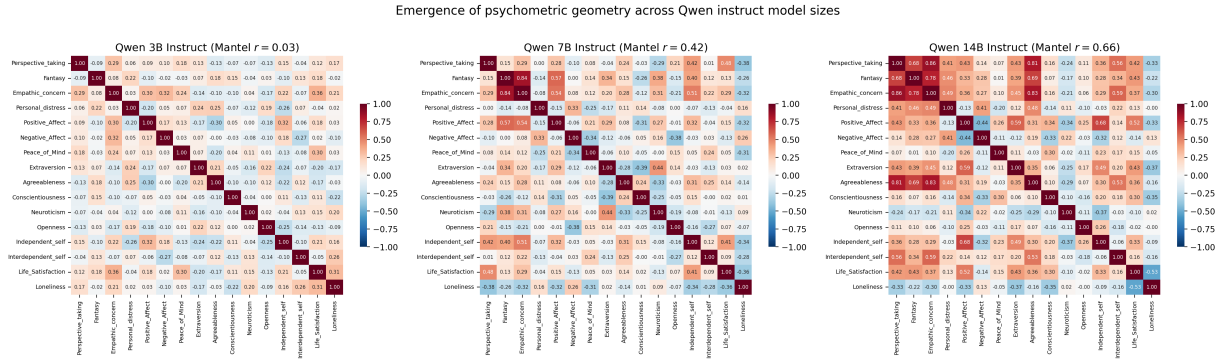

Figure 14. Emergence of psychometric geometry across Qwen instruct model sizes. Activation cosine similarity matrices for Qwen 3B Instruct (left, Mantel $r = 0.03$), Qwen 7B Instruct (middle, $r = 0.42$), and Qwen 14B Instruct (right, $r = 0.66$). Matrices use identical subscale ordering and colour scale. Block structure corresponding to human psychometric correlations becomes increasingly visible with scale.

Contrastive direction extraction procedure. For each of the 16 subscales, we construct paired *high* and *low* conditions. In the high condition, the target subscale's items (e.g., all 7 IRI Perspective Taking items) are set to the maximum value of the scale's response range (e.g., 5 for IRI); in the low condition, to the minimum. All other items across all 7 scales (to make a 114-item mega-prompt) are set to randomly sampled values from the scale range, independently per replication. The resulting system prompt is concatenated with a probe question (see below); the model generates 128 continuation tokens, and we record the mean residual-stream activation at each target layer over the generated token positions. We use $N = 10$ random background replications per (subscale, condition) pair, and $N = 20$ probe questions per condition, yielding $16 \times 2 \times 10 \times 20 = 6{,}400$ forward passes per model. The direction $\mathbf{d}_k$ for subscale k is the difference of means between the 200 high-condition activations and the 200 low-condition activations.

Probe questions. The 20 open-ended probe questions are in Traditional Chinese (matching the language of the Taiwanese subjects' responses) and deliberately neutral with respect to any psychometric content. They span 8 life domains (daily life, social/interpersonal, self-reflection, emotion, values, work/study, change/challenge, preferences), with questions such as "Describe how you typically spend a weekend," "How do you typically make important decisions," and "What do you consider the most important thing in maintaining a friendship." These are not psychometric items; their role is to elicit natural continuations whose activations differ systematically between the high and low conditions of the system prompt. The complete probe set is included in the project code release.

Layer sweep. We record activations at candidate layers spanning mid-to-late transformer depth, scaled to each model's number of layers: for Llama 8B ($L = 32$) we use layers {12, 14, 16, 18, 20, 22, 24}; for Qwen 14B ($L = 48$), {20, 24, 28, 32, 36}. Layer 16 ($\approx$50% depth) is our default for Llama 8B, matching both the CAA literature and our steering effects. For the contrastive direction analysis specifically, the "best layer" is chosen post-hoc to maximize the relevant Mantel r; Appendix B confirms negligible cross-layer variation.

Mantel test. Throughout, we use the Mantel test with 10,000 random permutations of one matrix's row/column indices to generate a null distribution for the off-diagonal correlation. Double-centered Mantel (used for steering specificity) first applies $\mathbf{M}_{\text{dc}} = \mathbf{M} - \mathbf{r}_{\text{row}} - \mathbf{c}_{\text{col}} + \mu$, removing generic row/column shifts before permutation.

Semantic similarity control. To rule out that direction geometry merely reflects lexical overlap between item wordings, we compute an item-embedding similarity matrix and partial it out of the alignment between contrastive direction geometry and human psychometric structure. We report two embedding sources to ensure the control is not specific to either choice: (a) each model's own input embedding layer (mean-pooled across item tokens), which captures the model's own representation of item lexical content; (b) the multilingual-e5-large sentence encoder, an external general-purpose semantic encoder shared across all 14 models. For each source we aggregate to a 16×16 subscale-level semantic similarity matrix S by mean-pooling item embeddings within each subscale. Across 14 models the mean zero-order $r(D, H)$ between contrastive direction geometry and the human psychometric matrix is 0.115; the mean partial $r(D, H \mid S)$ is 0.113 with embedding source (a) and 0.110 with source (b). The mean drop from zero-order to partial is below 0.01 for both sources, with maximum absolute drop across the 14 models below 0.03. On Llama 8B Instruct specifically, $r(D, H) = 0.368$ and the partial drops by 0.007

(source (a)) or 0.020 (source (b)). The geometry-structure link is therefore not driven by item-text semantic similarity. The same control is applied to the behavioral pipeline in App H (Figure 12); the conclusion is the same.

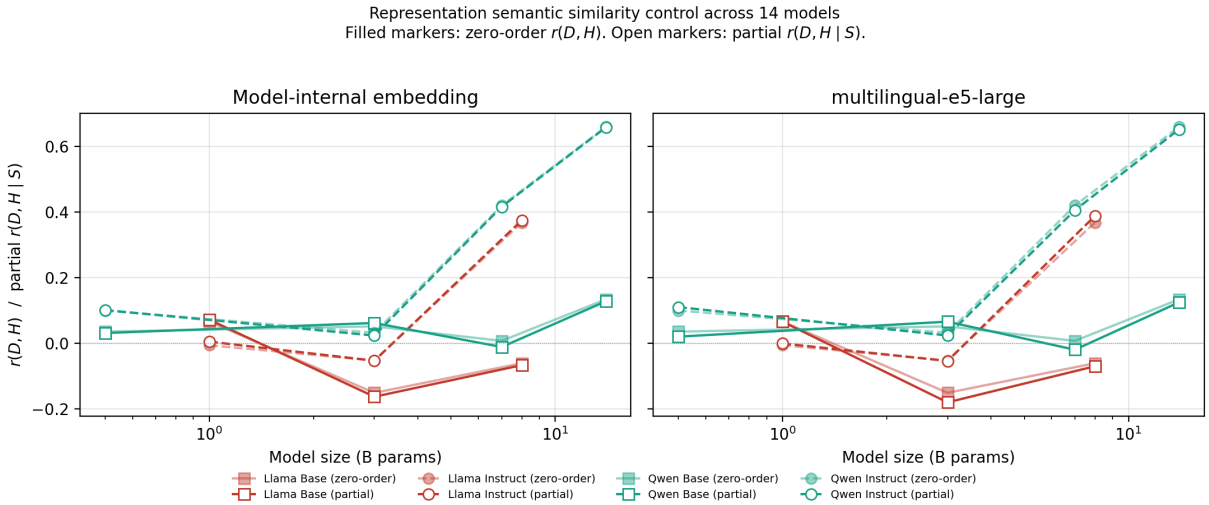

Figure 15. Representational semantic similarity control across 14 models. Filled markers: zero-order $r(D, H)$ between contrastive direction cosine similarity and the human subscale correlation matrix. Open markers: partial $r(D, H \mid S)$ controlling for item embedding similarity S. Left: each model's input embedding layer; right: multilingual-e5-large.

J. Steering Details

Regression direction extraction. Steering directions differ from the contrastive directions used for geometry analysis: we fit ridge regression per subscale, mapping $N = 272$ subjects' layer-L activations to their observed subscale scores. Ridge regularization is selected by 5-fold cross-validation over $\alpha \in \{0.1, 1, 10, 100, 1000\}$ (via `RidgeCV`). The direction for subscale k is the normalized weight vector $\mathbf{d}_k = \mathbf{w}_k / \|\mathbf{w}_k\|_2$. Cross-validated prediction r for each subscale provides a quality check; on Llama 8B Instruct, 14 of 16 subscales achieve CV $r > 0.2$ (the remaining two are IRI Fantasy and Personal Distress, which are also noisiest in our data).

Activation steering. We register a forward hook on the residual-stream output of layer L (`model.model.layers[L]`) that adds $\alpha \cdot \mathbf{d}_k$ to the residual at the *last token position* of the prompt (the position that immediately precedes generation), using the unit-normalised direction $\mathbf{d}_k$ defined above. We sweep $\alpha \in \{-2, -1, -0.5, 0, 0.5, 1, 2\}$ for each (source subscale, target item) pair and record the model's response log-probability distribution. Steering is applied at one token position in the model's native fp16 activation units; we do not rescale the direction by an empirical activation magnitude.

EV metric and steering-response matrix. For each target item with response options $\{1, 2, \ldots, K\}$, we compute the expected value $\text{EV} = \sum_{v=1}^{K} v \cdot p(v)$, which respects both the central tendency and the spread of the response distribution (unlike argmax). The steering effect is quantified per (source subscale s, target subscale t) as $\Delta\text{EV}_{s,t}$ = slope of EV vs. α, computed by OLS over the 7 α values. This yields a 9×16 steering-response matrix (a subset of 9 source subscales is used for steering to bound compute cost, spanning within- and between-scale pair types; all 16 target subscales are read out). Specificity is assessed by computing Mantel r (and its double-centered variant) between the steering-response matrix and the corresponding submatrix of the human psychometric correlation matrix.

Per-model results. In one Llama 8B Instruct steering run (layer 16), double-centered Mantel specificity is significant ($r = 0.28$, $p = 0.009$), as reported in the main text. Per-model steering runs show positive specificity in multiple instruct models, including strong double-centered effects in Llama 3B Instruct and Qwen 3B Instruct. The effects do not scale monotonically with model size, and a separate Llama 8B Instruct run is weaker than the layer-16 run reported above. This likely reflects two factors: higher-dimensional activation spaces make 272-subject direction extraction proportionally noisier, and fixed α magnitudes are not optimal across scales. Formal layer-wise and α-magnitude sweeps to disentangle these factors are left to future work.