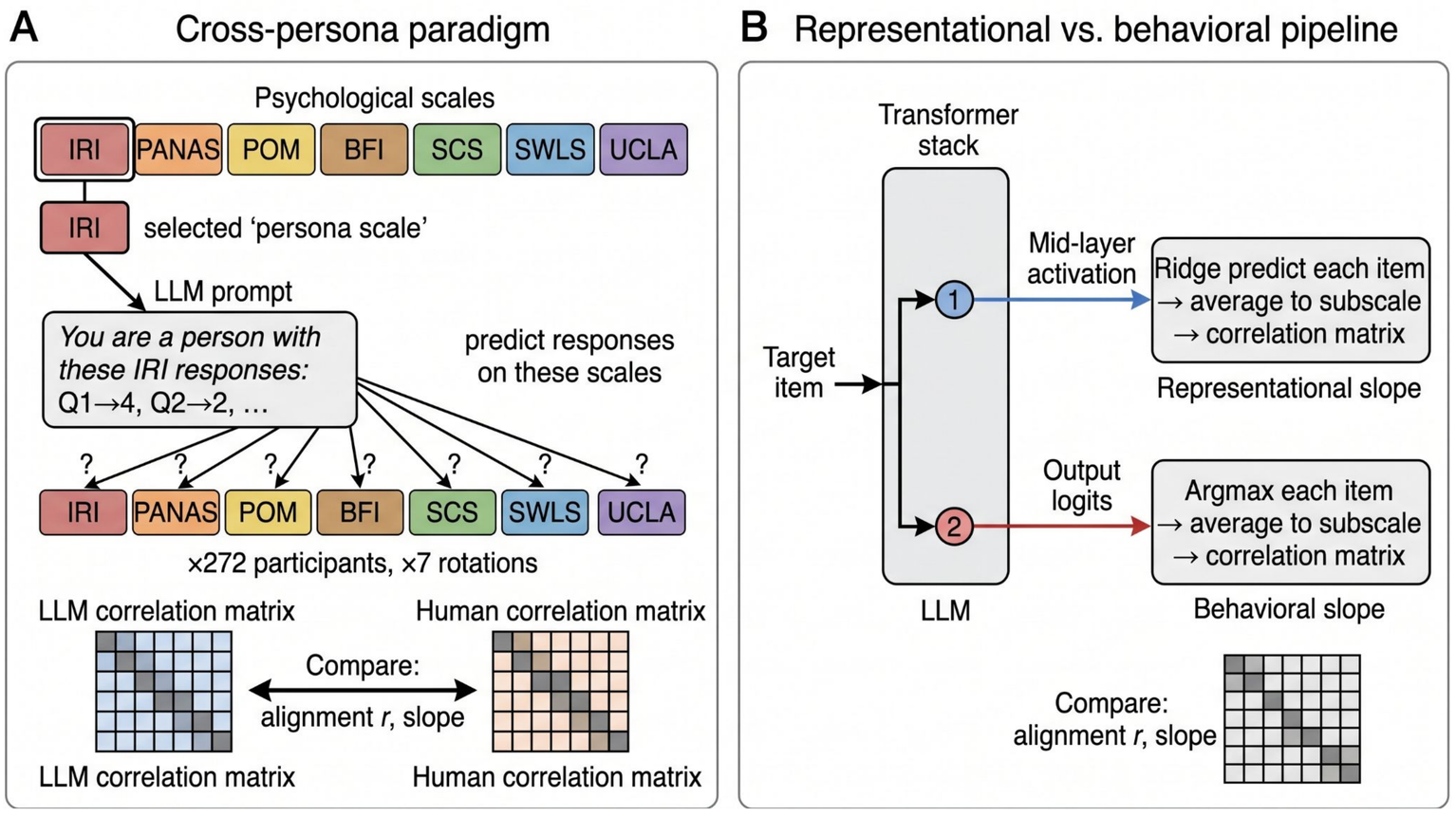
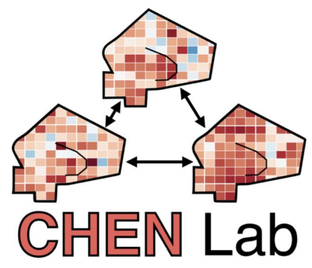
# **Psychometric** information is **linearly decodable** from activations, but human-aligned **activation geometry** emerges mainly in **large instruct** models.




# Tracing Psychometric Inference in Large Language Models

Chih-Hao Hsu, Feng-Chun Ben Chou, Pin-Hao Andy Chen

## 1 TL;DR

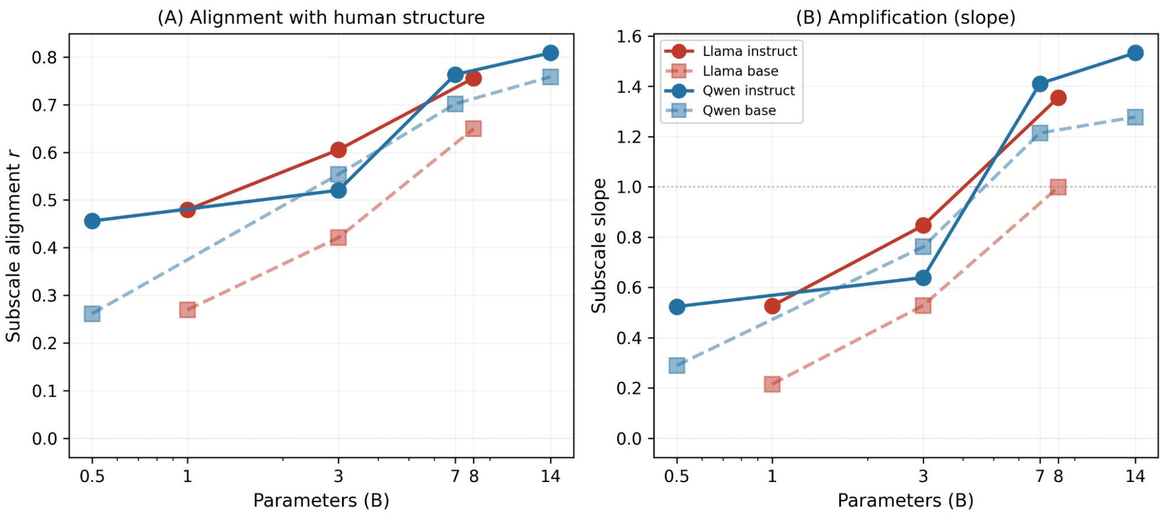
- In a rotating cross-persona task, LLMs reproduce and amplify human cross-scale psychometric correlations.
- Psychometric information is linearly decodable from mid-layer activations, while organized human-aligned geometry emerges mainly in large instruct models.
- Probe-predicted responses show less variable amplification than final model answers, revealing a representation-to-behavior gap.

## 2 Motivating Question

- LLMs can generate survey responses that resemble human psychometric data, but behavior alone cannot tell us whether psychological constructs are internally represented, linearly decodable, or causally used.

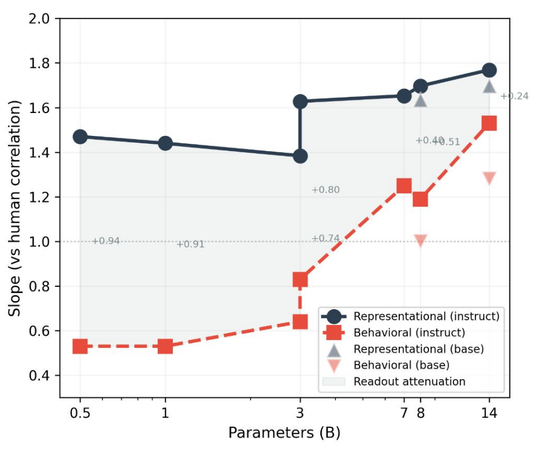
## 3 Cross-Persona Paradigm

- Responses from 272 human participants across 7 validated psychological scales, covering 114 items and 16 subscales.
- Each scale is used as the persona scale in turn. The model views a subject's responses on one scale and predicts responses on the remaining scales, producing a 16 x 16 cross-scale correlation matrix.
- Across 14 Llama 3 and Qwen 2.5 models, larger models better align with the human correlation matrix and increasingly amplify human correlations.

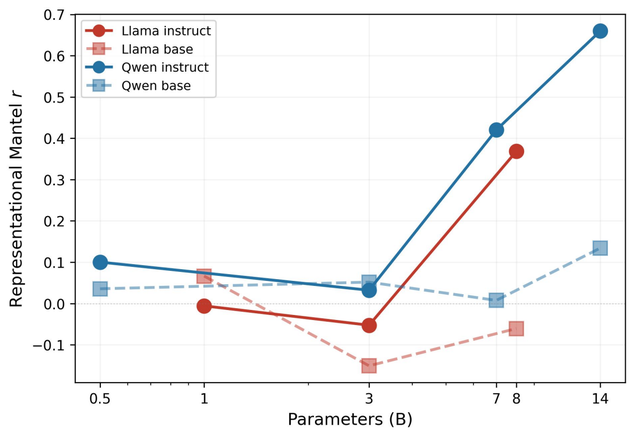



## 4 Probing, Readout, and Geometry

- Probing shows that cross-scale information is linearly decodable from mid-layer activations even when target-scale answers are absent from the prompt.

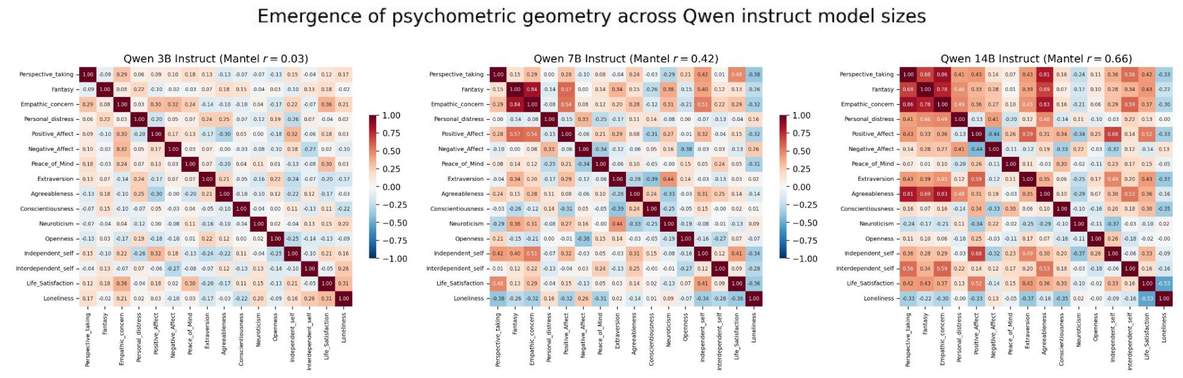



- Probe-based representational amplification is less variable than behavioral amplification.
- This representation-to-behavior gap decreases with scale and instruction tuning.
- Steering produces construct-specific effects in some instruct models.




- Extract one contrastive direction per subscale from high-vs-low psychometric conditions; compare pairwise cosine similarities to the human subscale correlation matrix.
- Organized geometry appears mainly in large instruct models

Emergence of psychometric geometry across Qwen instruct model sizes




## 5 Limitations and Discussion

- Human data come from a single cultural and linguistic sample.
- Steering provides suggestive causal-accessibility evidence, but not a quantitative scaling result
- The results support a representation-level account of psychometric structure so far.
- Future works (e.g., steering via contrastive directions) are needed for a causal mechanistic account.

PAPER

Mechanistic Interpretability Workshop, ICML 2026

howardhsuuu@gmail.com