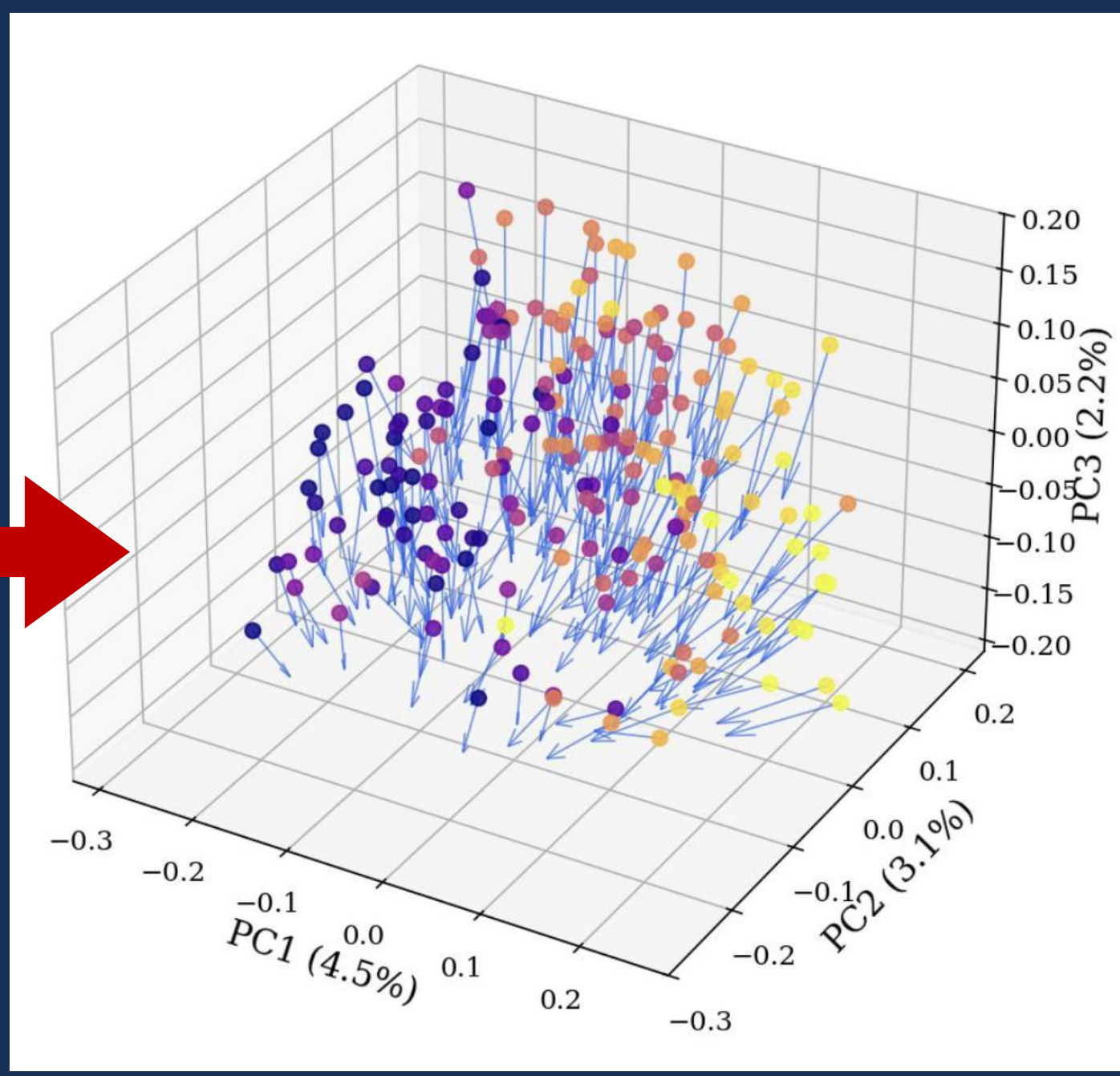
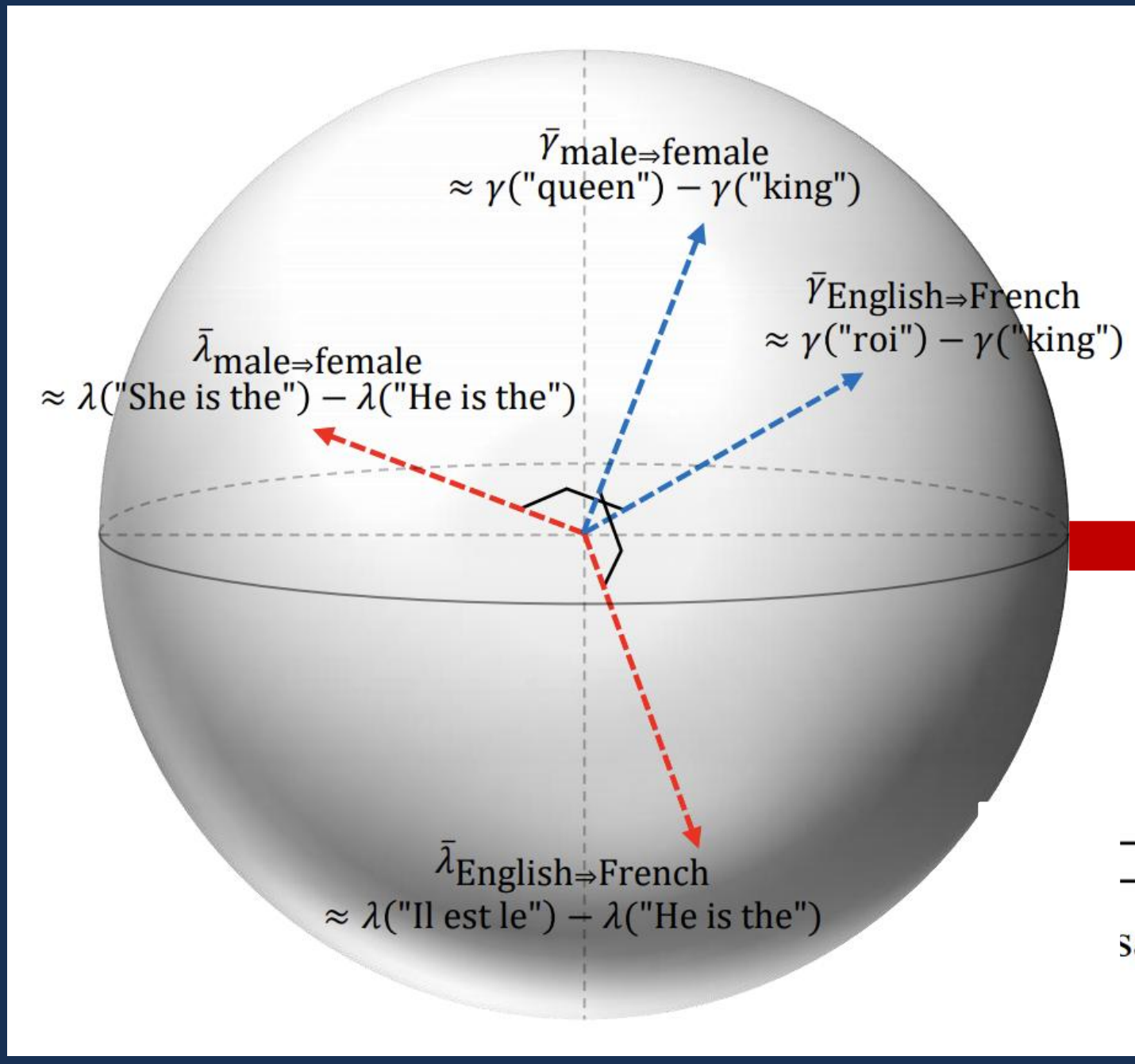




The transformation of concepts between contexts is not uniform but semantically meaningful and shared across models

# Language Models Represent and Transform Concepts with Shared Geometry

Zhimin Hu, Lanhao Niu, Sashank Varma

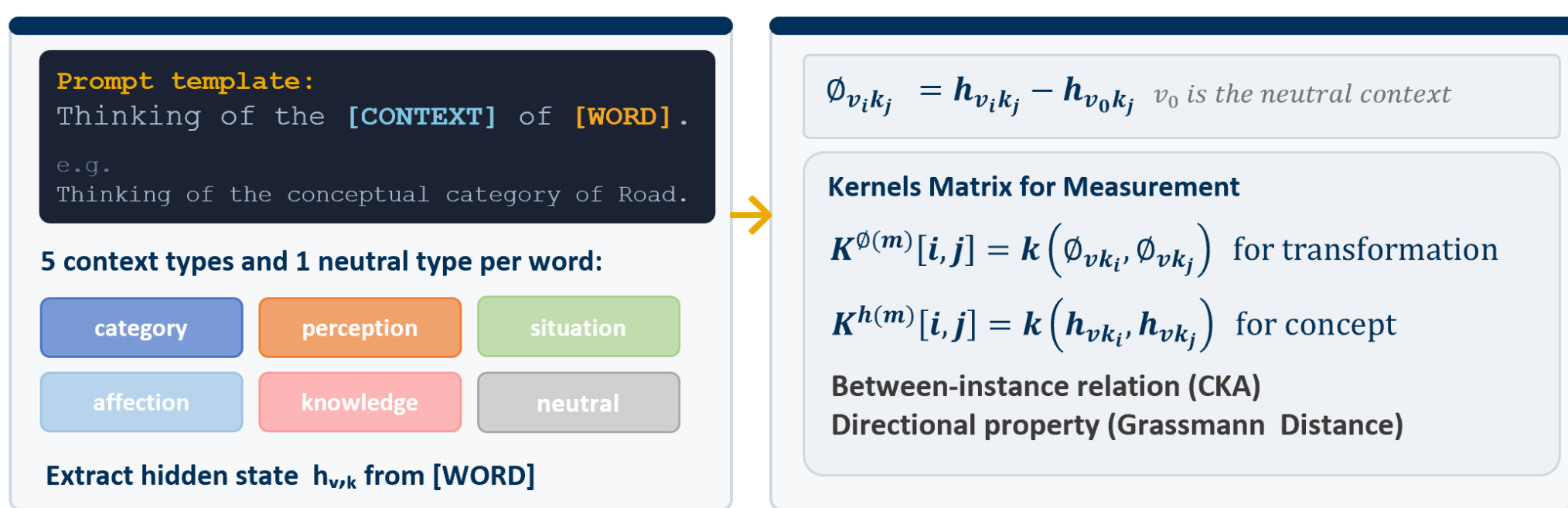
## 1 TL;DR

- Context moves each concept differently, not uniformly - in its own direction and magnitude
- This variation is semantically organized, not noise. They vary with lexical concreteness and density
- Models share a common geometry - not only in how concepts are represented, but in how context transforms them - across six model families

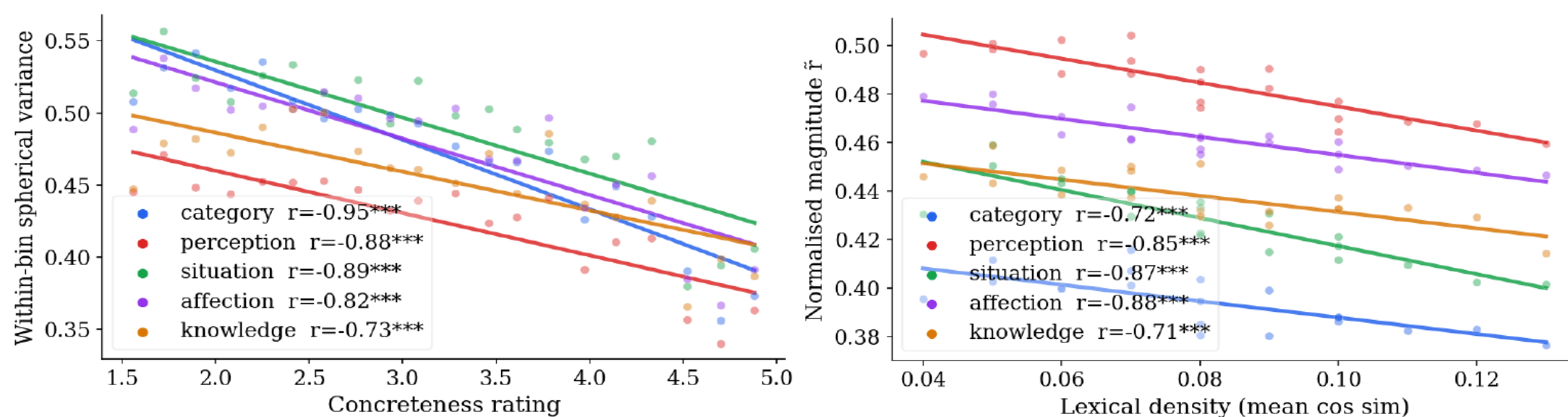
## 2 Background / Motivating Question

- Existing view treats a concept as a static vector
- Yet concepts appear in context, and context moves them.
- Does the way context transforms a concept have structure, or even shared across models?

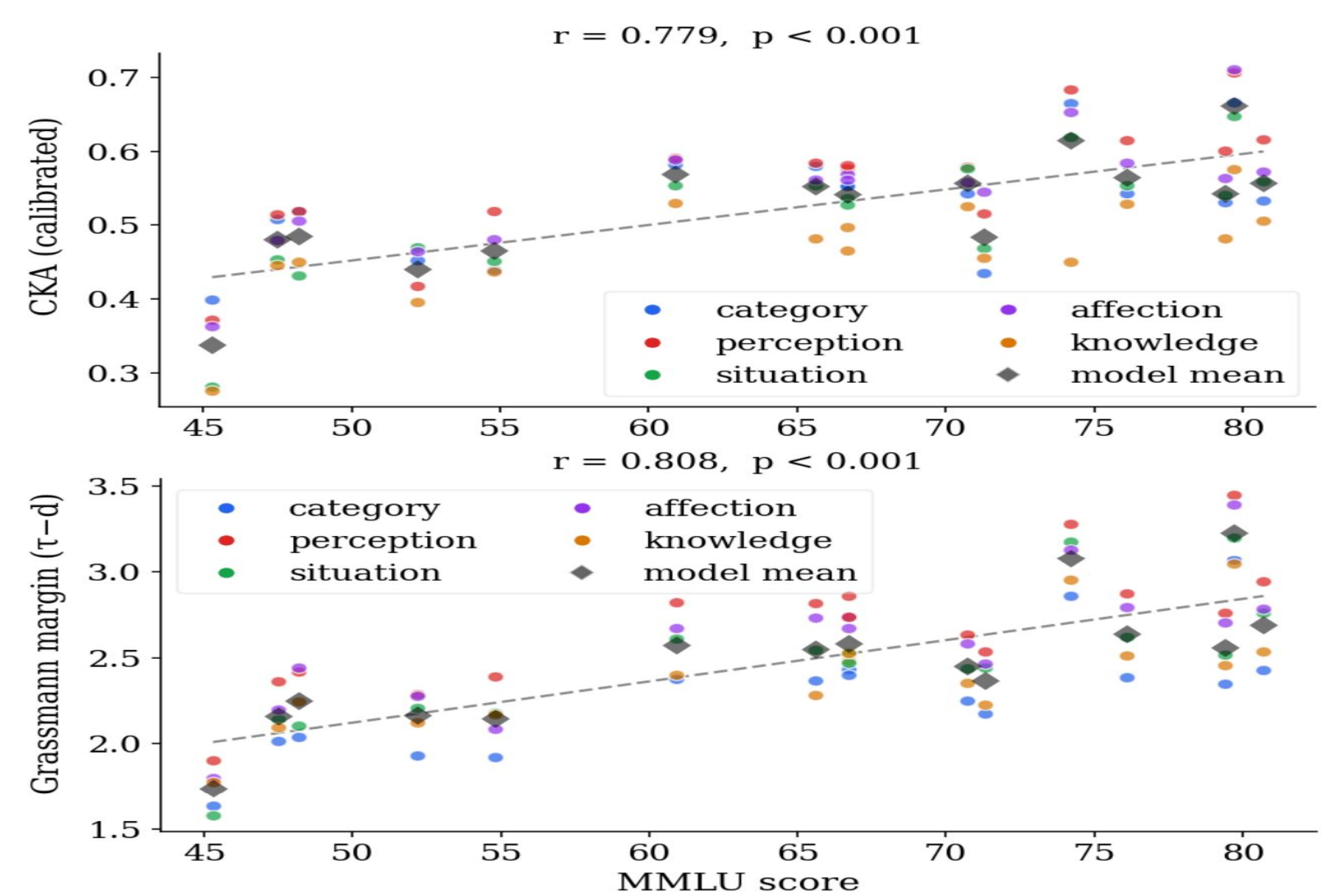
## 3 Measurement



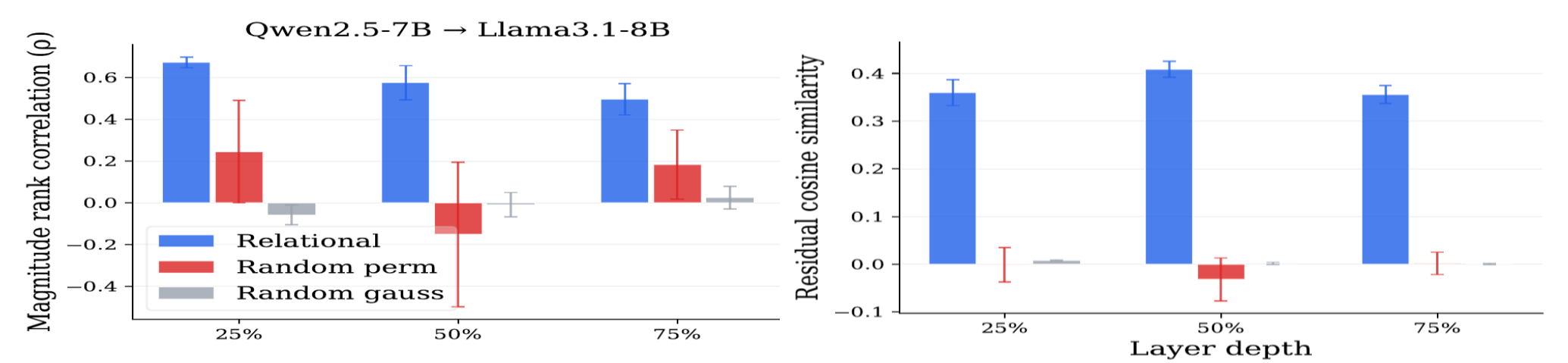
## 4 Share Geometry of Transformation



- More concrete concepts are displaced in more consistent directions across contexts.
- Concepts in denser neighborhoods (from Word2Vec) are displaced by smaller magnitudes.
- Shared displacement structure correlates more strongly with MMLU score than with parameter count.

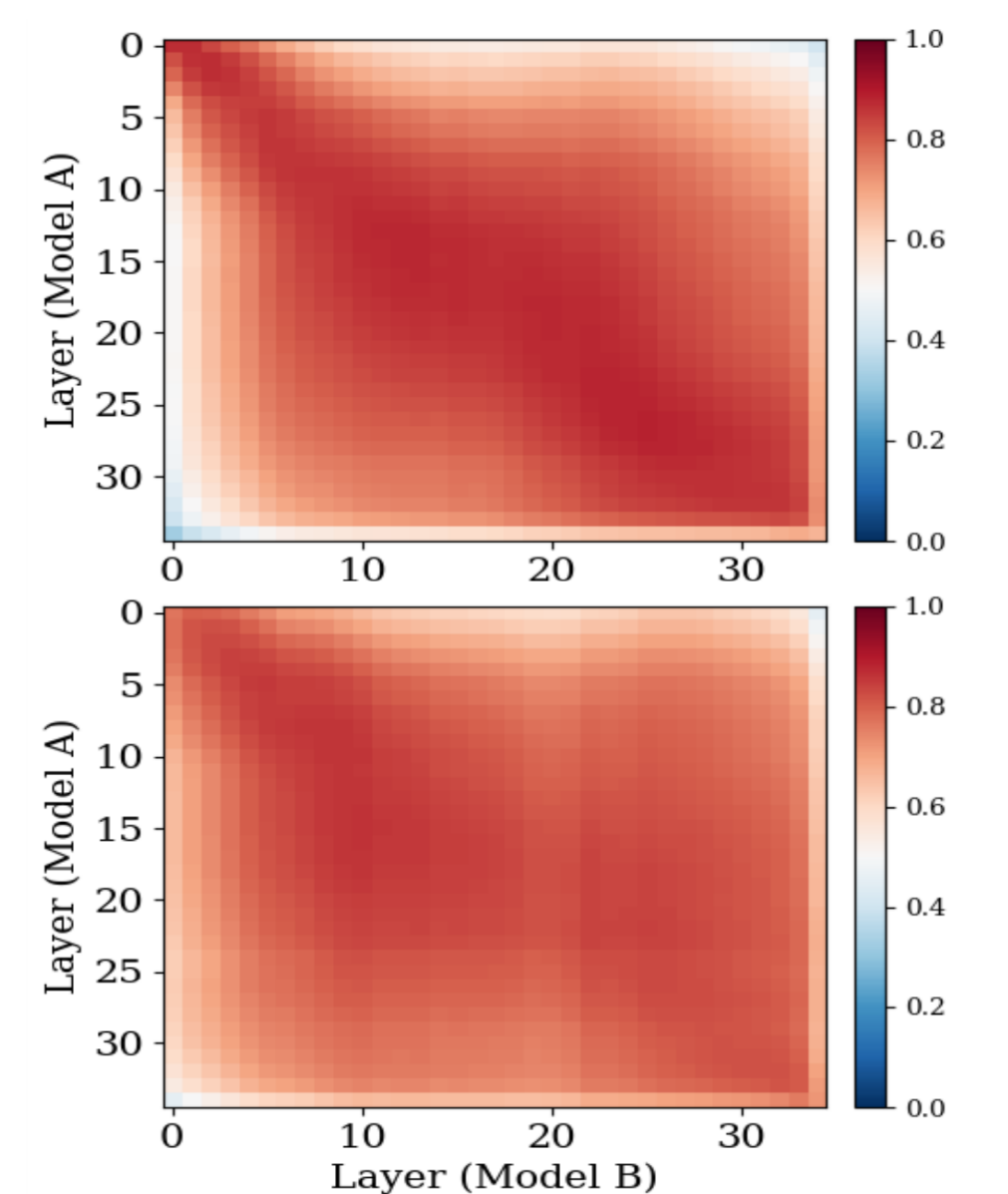


- Displacement structure from Model A predicts held-out displacements in Model B



## 4 Share Geometry of Concept

- The structure of various concepts under the same context (top) and the same concept under various contexts (down) is aligned (averaged across models)
- Alignments can not be reduced to surface statistical co-occurrence and scramble context degrade the alignments



## 5 Limitations and Discussion

- Scope of contexts and the size of vocabulary
- Steering and alignment methods manipulate something richer than a single direction—when we steer a representation, what exactly are we moving?
- Geometric structure support concepts to be simultaneously flexible and stable.



PAPER



CODE