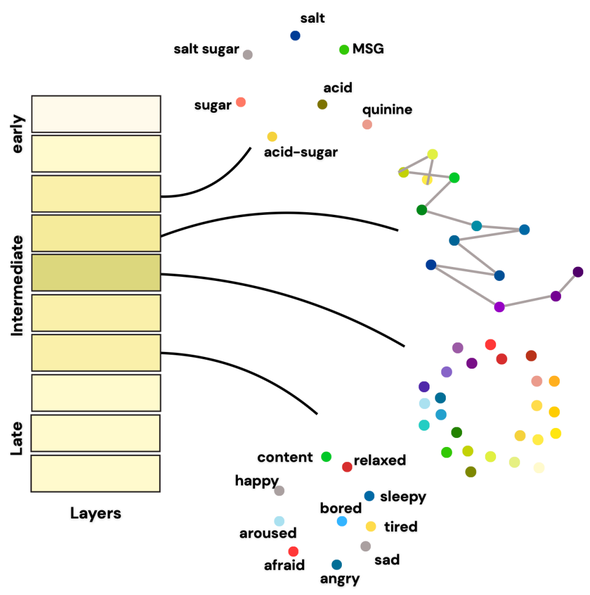
Geometry of Human Perceptual Domains Emerges Transiently in LLM Representations

Simardeep Singh
simardeep_s@mt.iitr.ac.in
Indian Institute of Technology- Roorkee, India

Paras Chopra
paras@lossfunk.com
Lossfunk, India

MOTIVATION

Large Language Models (LLMs) are trained purely on text, yet they often capture surprisingly rich relational structures in their internal representations. Previous work has shown that concepts such as colors, numbers, and time can form meaningful geometric manifolds inside these models. This raises an important question: do LLMs also organize perceptual concepts in ways that resemble human perception? In this work, we investigate how perceptual geometry develops across transformer layers for four domains color, pitch, taste, and emotion and identify where these human-like structures emerge within the model.



Methodology

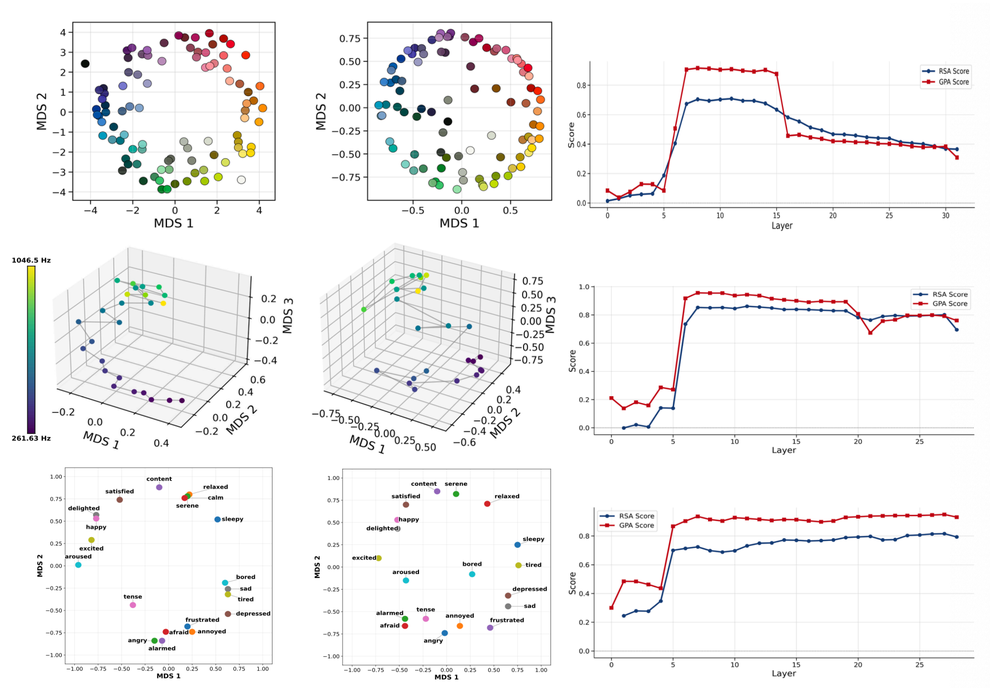
For each perceptual domain, we construct simple prompts and extract hidden representations from every transformer layer of multiple open-weight LLMs. Pairwise cosine distances between these representations are used to build geometric maps through Multidimensional Scaling (MDS). These layerwise geometries are then compared with corresponding human perceptual maps using Representational Similarity Analysis (RSA) and Generalized Procrustes Analysis (GPA), allowing us to quantify how closely model representations align with human perception across model depth.

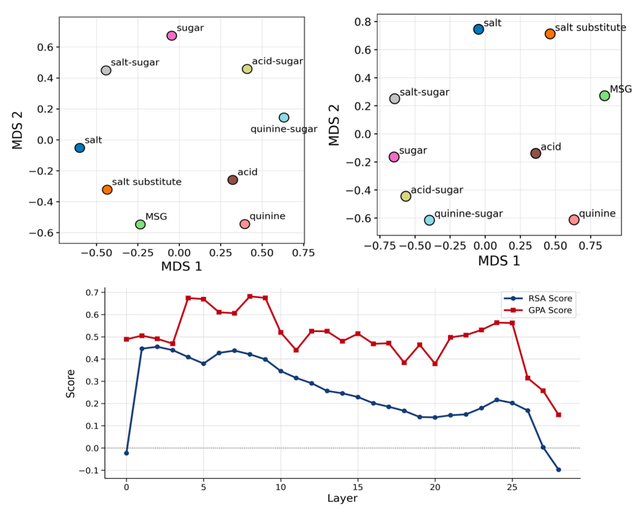
Key Results

Human-like perceptual geometry emerges naturally within LLM representations despite the absence of explicit perceptual supervision during training. Across all evaluated models, geometric organization is weak in early layers, becomes highly structured in intermediate layers, and gradually attenuates in deeper layers. This consistent rise–peak–fall trajectory suggests that perceptual geometry is transient rather than permanently encoded, indicating that intermediate layers play a central role in organizing perceptual information before later layers specialize for downstream language processing.

Domain-specific Findings

Although all perceptual domains exhibit transient emergence, each follows a distinct representational trajectory. Color forms a circular manifold closely resembling the human color wheel. Emotion produces biaxial valence-arousal structure which is the strongest and most persistent geometry across deeper layers. Pitch develops a smooth continuous manifold reflecting ordinal relationships between frequencies. Taste emerges earlier but degrades more rapidly, indicating that simpler perceptual structures are formed quickly but are less stable than more complex affective representations.





Limitations and Discussion

Our findings reveal that perceptual organization is an emergent property of language models rather than a consequence of explicit sensory training. Different perceptual domains emerge at different stages of computation, demonstrating that representational geometry is not uniformly distributed across the network. This study is descriptive and focuses on identifying where perceptual geometry emerges rather than explaining the underlying mechanisms responsible for its formation.
Understanding the neural circuits responsible for these transient geometric representations remains an important direction toward mechanistically explaining how LLMs organize perceptual knowledge. Additionally, geometric alignment with human baselines should not be interpreted as evidence of human-like perception, but only as evidence of structurally similar representational organization.