Geometry of Human Perceptual Domains Emerges Transiently in LLM Representations

Anonymous Authors[1]

Abstract

While large language models (LLMs) are trained purely on textual data, prior work has shown that their internal representations can exhibit rich geometric structure in embedding space. Building on this line of work, we investigate whether such structure is similar to human perceptual organisation across different domains (e.g., color, pitch, emotion, and taste). Specifically, we study the layer-wise emergence of intrinsic geometrical structure corresponding to perceptual modalities within the residual streams of multiple open-weight transformer architectures. Our results reveal three key findings. First, we observe the emergence of layer-wise geometric structure across multiple perceptual domains, despite the absence of any direct perceptual supervision during training. Second, these perceptual domains exhibit distinct emergence profiles, with both geometric structure and its alignment with human baselines following domain- and model-specific trajectories across depth. Third, this emergence follows a consistent representational trajectory: geometry is weak or diffuse in early layers, becomes progressively organised in intermediate layers, and is attenuated in later layers, suggesting that perceptual geometry arises transiently as part of the model's internal transformation pipeline. This provides new insight into how and where human-like perceptual geometry arises in LLMs, offering a principled pathway for mechanistic analysis of internal representations.

1. Introduction

LLMs have demonstrated remarkable capabilities in capturing relational geometric structures from purely textual data. Recent work has shown that LLM representations exhibit structured geometry across a range of concepts. For instance, cyclical domains such as days of the week, months, or hues are often organized along circular manifolds, reflecting their inherent periodic structure (Engels et al., 2025; Modell et al., 2025). (Gurnee & Tegmark, 2024; Gurnee et al., 2026) show that LLM representations corresponding to continuums like years of history or character counts are organized in compact 1D manifolds. A unifying explanation for these geometric patterns is provided by (Karkada et al., 2026), who argue that such structures emerge naturally from translation symmetries in language co-occurrence statistics. Importantly, this phenomenon extends beyond simple conceptual domains to more complex scientific concepts. For example, representations of the periodic table have been shown to organize into a structured 3D spiral, where intermediate layers encode continuous relational properties and deeper layers sharpen categorical distinctions (Lei & Cooper, 2025).

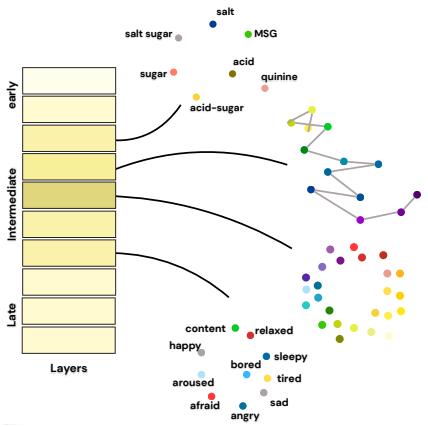

Figure 1. Perceptual geometric structures emerge within LLM representations across sensory and affective domains (taste, pitch, color, and emotion). Simpler structures (taste) peak earlier, followed progressively by pitch, color, and emotion. Across domains, geometry is weak in early layers, organizes in intermediate layers, and attenuates in later layers.

Despite lacking direct sensory grounding, these models often exhibit behaviors that suggest an implicit understanding of perceptual and affective concepts such as color, pitch, and

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

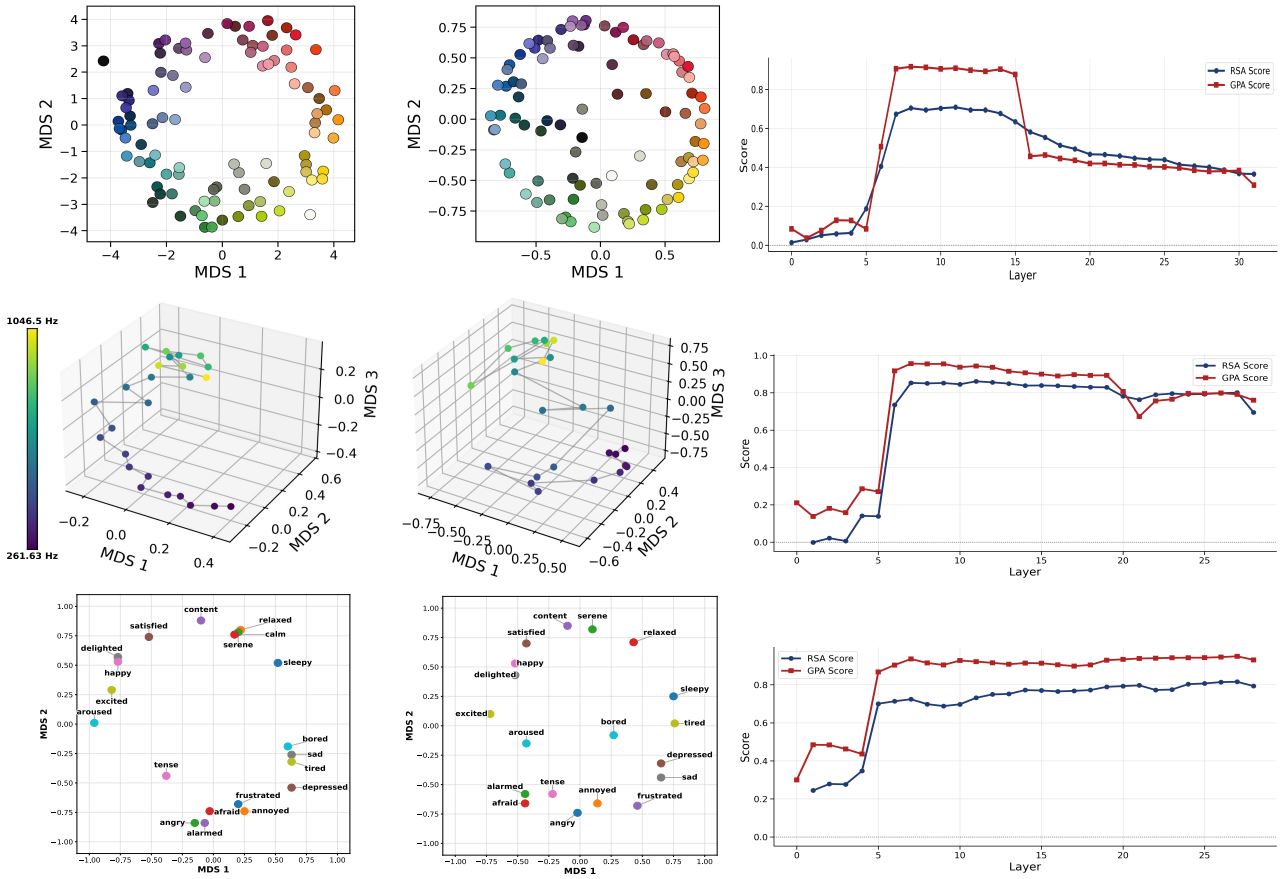

Figure 2. Overview of perceptual geometry emergence across modalities. Each row corresponds to a perceptual domain—color (top, LLaMA-3-8B), pitch (middle, Qwen-3-4B), and emotion (bottom, Gemma-7B), with the clearest representative model shown for each. Columns show the human perceptual baseline (left), the peak-alignment geometric representation from the model (middle), and the corresponding layer-wise alignment profile (right).

emotion. Prior work suggests that language itself may be closely linked to the organization of subjective experience, with internal representations in neural systems reflecting structured relationships between qualia and linguistic concepts (Taniguchi et al., 2025). This raises a fundamental question: *to what extent do internal representations in LLMs reflect human perceptual organization?*

Mechanistic interpretability provides a framework for addressing such questions by analyzing how information is represented and transformed across intermediate layers of neural networks (Singh et al., 2024; Dar et al., 2023; Alain & Bengio, 2018) and has shown that transformer representations evolve across layers, with early layers encoding lexical features, and middle layers capturing abstract semantic relationships (Skean et al., 2024; 2025).

Emerging studies have begun to explore the relationship between language models and human perceptual structures. Evidence shows that LLMs can reproduce human sensory judgments across modalities and recover known perceptual structures such as the color wheel and pitch spiral through model outputs (Marjieh et al., 2023). Similarly, (Abdou et al., 2021) demonstrates alignment between Language models and color spaces like CIELAB, highlighting that perceptual structure can, to some extent, be inferred from language. However, these approaches primarily operate either at the level of model outputs or task specific probing and do not directly examine how perceptual structure emerges and evolves across layers within the model.

We present three main contributions:

- **Layer-wise emergence of perceptual geometry:** We demonstrate that LLMs develop structured geometric representations corresponding to multiple perceptual domains (color, pitch, taste, and emotion) within their residual streams, and quantitatively compare these representations with human perceptual baselines.
- **Modality-specific emergence profiles:** Different perceptual domains exhibit distinct emergence profiles, with both geometric structure and its alignment with human perceptual baselines following domain- and

model-specific trajectories across model depth.

- **Transient perceptual encoding:** Perceptual structure follows a consistent depth-wise trajectory across models: it is weak or diffuse in early layers, becomes progressively organised in intermediate layers, and attenuates in later layers, indicating that perceptual geometry is encoded transiently rather than uniformly across depth.

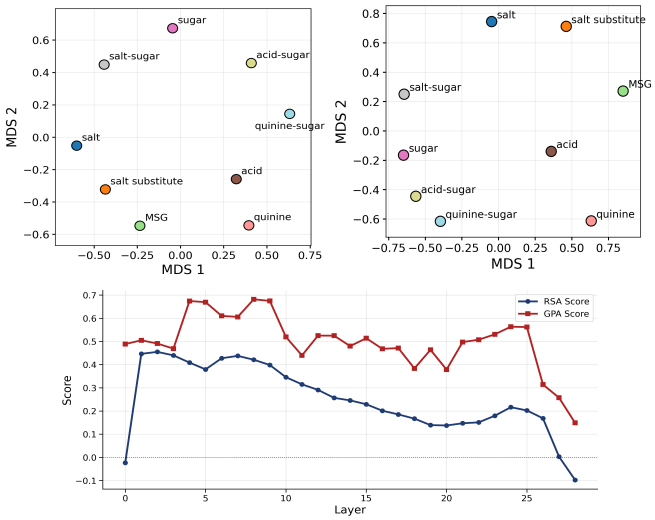

Figure 3. Comparison of the human perceptual map (top left) and peak-layer LLM representation for taste in Gemma-7B (top right). Bottom: layer-wise alignment scores (RSA and GPA).

2. Methodology

2.1. Overview

We investigate whether perceptual geometric structures exist within the embedding space of LLMs which aligns with the human similarity judgements. Our approach is fully *intrinsic*, requiring no probing or additional training. Given a set of stimuli (e.g., colors, emotions, pitch values), we extract layer-wise representations from pretrained LLMs and construct geometric maps, which are then quantitatively compared with human perceptual baselines.

Formally, let a set of stimuli be denoted as $\{x_i\}_{i=1}^N$, where each stimulus is associated with a minimal structured prompt P_i (see Appendix A.1 and A.2 for all prompt templates and their rationale). For example, for color stimuli, a prompt may take the form: *"The description of color given as* `#9B081A`*"*. Each prompt is passed through a model m to obtain last token hidden state activation at every transformer layer l, $\mathbf{h}_i^{(l,m)} \in \mathbb{R}^d$.

Then, for each layer l, we construct a geometric representation of the stimulus set through the following steps:

Dissimilarity Computation: Given embeddings $\{\mathbf{h}_i\}_{i=1}^N$, we compute pairwise cosine dissimilarities:

$$d_{ij} = 1 - \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{\|\mathbf{h}_i\|\|\mathbf{h}_j\|}. \tag{1}$$

This gives a symmetric dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$. Cosine distance isolates directional structure and is invariant to scale differences across layers.

Multidimensional Scaling (MDS): To obtain a low-dimensional geometric embedding, we MDS (Kruskal, 1964). Given $\mathbf{D}$, MDS finds coordinates $\{\mathbf{y}_i \in \mathbb{R}^2\}$ by minimizing the following stress function:

$$\min_{\{\mathbf{y}_i\}} \sum_{i<j} \left(d_{ij} - \|\mathbf{y}_i - \mathbf{y}_j\|\right)^2. \tag{2}$$

This produces a 2D representation that preserves pairwise dissimilarities, forming a qualia map for each layer. These model representations were compared against human similarity judgments (see Appendix A.3 for dataset details). The resulting human dissimilarity matrices from the datasets were projected into a low-dimensional space using MDS to construct the baseline perceptual geometry.

For quantifying the alignment between LLM and human representations we use the following metrics:

Representational Similarity Analysis (RSA): RSA measures the correlation between two dissimilarity matrices. Let $\mathbf{D}^{\text{LLM}}$ and $\mathbf{D}^{\text{human}}$ denote the model and human dissimilarity matrices. We compute RSA as the Spearman rank correlation between the vectorized upper-triangular entries of $\mathbf{D}^{\text{LLM}}$ and $\mathbf{D}^{\text{human}}$.

Generalized Procrustes Analysis (GPA): GPA measures geometric alignment between two geometric maps by finding the optimal orthogonal transformation:

$$\min_{\mathbf{R}} \|\mathbf{Y}^{\text{LLM}}\mathbf{R} - \mathbf{Y}^{\text{human}}\|_F^2. \tag{3}$$

The final score is computed as $\text{GPA} = 1 - \frac{\text{residual}}{2}$. Higher values indicate stronger geometric alignment.

2.2. Layer-wise Profiling

We apply the above pipeline across all layers $l \in \{0, \ldots, L\}$ for each model m, producing a sequence of geometric maps $\{\mathbf{Y}^{(l,m)}\}$. Each layer is independently compared with the human baseline using RSA and GPA, yielding a layer-wise alignment profile. This allows us to trace how perceptual structure evolves across depth rather than treating representation quality as a single endpoint. By evaluating each layer independently, we can identify where perceptual organisation first becomes measurable, how it strengthens through intermediate processing, and whether it is preserved or transformed in later layers.

3. Results and Discussion

We analyse the emergence of perceptual structure across four modalities: color, emotion, pitch, and taste. For each modality, we examine both qualitative geometry (via MDS maps) and quantitative alignment (via RSA and GPA) across layers. Figure 2 summarizes the central result for three representative modalities—color, emotion, and pitch by showing the human perceptual geometry, the peak-alignment model geometry, and the corresponding layer-wise alignment profile. Full layer-wise geometric trajectories, including early and late-layer maps as well as complementary Isomap visualizations, are provided in Appendix B.1.

For color, the peak-layer representation in Figure 2 exhibits a clear circular manifold closely resembling the human perceptual color wheel, indicating that the model recovers a smooth chromatic geometry from purely linguistic supervision. This stage corresponds to peak RSA and GPA values, where the model reorganizes raw lexical statistics into a structured geometry without any explicit supervision. The corresponding layer-wise profile shows that this structure emerges gradually, peaks in the early-middle layers, and weakens thereafter. This depth-wise pattern is broadly consistent across architectures (Figure 21): LLaMA-3-8B, LLaMA-3.2-3B, and Gemma-7B each exhibit a clear rise–peak–fall profile, where color alignment peaks at an intermediate layer followed by late-layer attenuation. Early and last layer maps (Appendix B.1) show that this intermediate organization is preceded by fragmented local structure and followed by geometric degradation in later layers as the model increasingly specialize for task specific representations. Qwen-3-4B follows the same overall trajectory but with a shifted late-layer profile: after the intermediate peak, alignment briefly rebounds in the deeper layers before degrading again toward the final layers. Taken together, these results suggest that perceptual color structure is transient, emerging most coherently in intermediate representations before weakening in later layers.

For emotion, the peak-layer representation recovers a well-organized affective manifold aligned with the human valence–arousal structure (Russell, 1980). In contrast to color, the associated layer-wise profile shows that this alignment is not only strong at its peak but remains comparatively stable across later layers, indicating more persistent retention of emotional geometry through depth. Our observation of a low-dimensional affective space aligns with concurrent findings by (Choi & Weber, 2026; Sun et al., 2026); however, our analysis additionally localizes where this structure becomes most coherent and how its stability differs from other perceptual domains. Detailed layer-wise maps (Figures 11 and 12) further illustrate this progressive organization and relative preservation.

For pitch, the peak-layer geometry reveals a smooth arc-like organization consistent with the continuous and ordinal nature of human pitch perception. The layer-wise profile shows a more localized emergence pattern in which relational structure becomes coherent in intermediate layers before progressively deforming at greater depth. The full sequence in Figures 16 and 17 shows that early layers exhibit only weak partial ordering, intermediate layers undergo a clear structural transition in which frequencies become arranged along a arc-like manifold, and later layers progressively deform this relational organization. The absence of discrete clusters and the emergence of a smooth trajectory suggest that pitch is represented in a relational, continuous manner rather than as a set of categorical groupings.

Taste representations also recover a clearly organized geometry that closely matches the human perceptual structure (Figure 3). The peak-layer map recovers a qualitatively well-formed taste manifold, with the relative ordering of primary tastes and mixtures broadly matching the human perceptual arrangement (Figures 13, 14). This is reflected in strong geometric alignment (High GPA scores) at peak layers, indicating that the model captures the overall global structure of taste relations. However, taste differs from the other perceptual domains in its lower RSA scores and noisier layer-wise profiles (Figures 23) , indicating that this structure is less stable and less precise at the level of fine-grained pairwise relations. The layer-wise maps and metric profiles show that taste geometry emerges early, but degrades rapidly thereafter, with a markedly noisier and less stable trajectory across subsequent layers than observed in the other perceptual domains.

4. Limitations

Our analysis is primarily descriptive: while we identify where perceptual geometry emerges and how it evolves across depth, we do not explain the mechanisms that give rise to these manifolds. This includes architecture-specific deviations—such as the late layer rebound observed in Qwen's color trajectory, which we can localize descriptively but do not yet explain mechanistically. Understanding why such structure forms remains an important direction for future mechanistic work. Second, our study is limited to four perceptual domains. Although the observed trends are broadly consistent, broader evaluation across domains, architectures, and prompting conditions is needed to establish generality. We further restrict evaluation to RSA and GPA; while these capture complementary aspects of representational geometry, other metrics may reveal additional aspects of organization. Finally, geometric alignment with human baselines should not be interpreted as evidence of human-like perception, but only as evidence of structurally similar representational organization.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. Can language models encode perceptual structure without grounding? a case study in color, 2021. URL https://arxiv.org/abs/2109.06129.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Bradley, M. M. and Lang, P. J. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report. URL https://pdodds.w3.uvm.edu/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf.

Choi, B. J. and Weber, M. Latent structure of affective representations in large language models, 2026. URL https://arxiv.org/abs/2604.07382.

Dar, G., Geva, M., Gupta, A., and Berant, J. Analyzing transformers in embedding space, 2023. URL https://arxiv.org/abs/2209.02535.

Engels, J., Michaud, E. J., Liao, I., Gurnee, W., and Tegmark, M. Not all language model features are one-dimensionally linear, 2025. URL https://arxiv.org/abs/2405.14860.

Gurnee, W. and Tegmark, M. Language models represent space and time, 2024. URL https://arxiv.org/abs/2310.02207.

Gurnee, W., Ameisen, E., Kauvar, I., Tarng, J., Pearce, A., Olah, C., and Batson, J. When models manipulate manifolds: The geometry of a counting task, 2026. URL https://arxiv.org/abs/2601.04480.

Karkada, D., Korchinski, D. J., Nava, A., Wyart, M., and Bahri, Y. Symmetry in language statistics shapes the geometry of model representations, 2026. URL https://arxiv.org/abs/2602.15029.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, Mar 1964. ISSN 1860-0980. doi: 10.1007/BF02289565. URL https://doi.org/10.1007/BF02289565.

Lei, G. and Cooper, S. J. Layerwise recall and the geometry of interwoven knowledge in llms, 2025. URL https://arxiv.org/abs/2502.10871.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., and Griffiths, T. L. Large language models predict human sensory judgments across six modalities, 2023. URL https://arxiv.org/abs/2302.01308.

Modell, A., Rubin-Delanchy, P., and Whiteley, N. The origins of representation manifolds in large language models, 2025. URL https://arxiv.org/abs/2505.18235.

Russell, J. A circumplex model of affect. 39:1161–1178, 12 1980. URL https://www.researchgate.net/publication/235361517_A_circumplex_model_of_affect.

Singh, C., Inala, J. P., Galley, M., Caruana, R., and Gao, J. Rethinking interpretability in the era of large language models, 2024. URL https://arxiv.org/abs/2402.01761.

Skean, O., Arefin, M. R., LeCun, Y., and Shwartz-Ziv, R. Does representation matter? exploring intermediate layers in large language models, 2024. URL https://arxiv.org/abs/2412.09563.

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. Layer by layer: Uncovering hidden representations in language models, 2025. URL https://arxiv.org/abs/2502.02013.

Sun, L., Yan, L., Lu, X., Lee, A., Zhang, J., and Shao, J. Valence-arousal subspace in llms: Circular emotion geometry and multi-behavioral control, 2026. URL https://arxiv.org/abs/2604.03147.

Taniguchi, T., Oizumi, M., Saji, N., Horii, T., and Tsuchiya, N. Constructive approach to bidirectional influence between qualia structure and language emergence, 2025. URL https://arxiv.org/abs/2409.09413.

A. Appendix

A.1. Prompts Used

We use a set of minimal, structured prompts to extract representations corresponding to each domain. The list of prompts used across modalities is provided below:

Color: "The description of the color given as `#HEXCODE`"
Emotion: "Describe the person who is feeling `emotion`"
Taste: "The description of taste on tongue given by `taste`"
Pitch: "The description of sound of musical tone `VALUE Hz`"

We intentionally employ minimal and structurally simple prompts to avoid injecting additional semantic bias or task-specific reasoning into the representations. This ensures that the extracted embeddings primarily reflect the model's intrinsic encoding of the stimulus rather than artifacts introduced by external or additional linguistic context.

A.2. Prompt Completion Examples

To complement the minimal prompt templates described above, we provide representative prompt-completion examples generated by the LLMs for each modality. Each example contains the queried prompt for a given stimulus followed by the top descriptor phrases generated by Qwen3-8B, LLaMA-3-8B, and Gemma-7B. These examples verify that the prompts consistently elicit coherent and modality-relevant descriptions despite their intentionally minimal structure. They also ground the embedding analysis by showing the natural language completions from which latent representations are extracted, demonstrating that the resulting embeddings arise from semantically meaningful and stimulus-consistent textual descriptions rather than arbitrary vector artifacts.

Prompt: Description of color given by #9B081A

"dark red", "maroon", "60% red, 3% green, 10% blue", "hue 0°", "sat 94.7%", lightness 12.7%, shade of red, "warm, deep tone"

Prompt: Describe the person who is feeling Afraid

"muscles are tense, breathing is shallow and quick, heart is pounding", "mind is racing... on edge, expecting something bad to happen", "tense and alert"

Prompt: Description of color given by #73B7C3

"cool light blue", "slight green undertone", "lightness 84.3%", "saturation 31.4%", "calm, serene"

Prompt: Describe the person who is feeling Satisfied

"content, calm, and at ease... sense of accomplishment and fulfillment", "relaxed posture, a sense of inner peace", "breathing is steady and even"

Prompt: Description of color given by #AAC901

"yellow-green", "bright, high luminance", "67.8% red, 81.2% green, 0.4% blue", "hue ~75°"

Prompt: Describe the person who is feeling Tense

"physical tightness in their muscles... on edge, restless", "thoughts might be racing... mind might be filled with negative thoughts", "feeling of impending doom or dread"

Figure 4. Representative prompt completions for color and emotion. Each panel shows the queried prompt together with the top descriptor phrases produced by Qwen3-8B, LLaMA-3-8B, and Gemma-7B.

Prompt: The description of sound of musical tone 261.63 Hz

"a pitch height of middle C", "smooth, clear, and full sound", "lack of overtones or harmonics", "feelings of calmness, serenity", "bright, clear sound"

Prompt: The description of taste on tongue given by Quinine-Sugar

bitter and astringent, bitter and sweet, bitter compound, balance of bitter and sweet, unique and complex flavor experience

Prompt: The description of sound of musical tone 554.37 Hz

"high C sharp (C#) above middle C", "well-defined frequency without harmonic overtones", "bright and uplifting", "evoke feelings of joy and happiness", "pleasant and easily recognizable"

Prompt: The description of taste on tongue given by MSG

savory, savory enhancer that increases the savory taste of food, Umami

Prompt: The description of sound of musical tone 783.99 Hz

"relatively high pitch", "very sharp and clear quality due to its lack of harmonics or overtones", "above the typical range of human speech", "still within the upper range of human hearing", "excitement, alertness, and energy"

Prompt: The description of taste on tongue given by Quinine

bitter alkaloid... causes a bitter sensation, sharp, unpleasant sensation, well-known bitter compound, famously used in tonic water

Figure 5. Representative prompt completions for pitch and taste. Each panel shows the queried prompt together with the top descriptor phrases produced by Qwen3-8B, LLaMA-3-8B, and Gemma-7B.

A.3. Human Baseline Dataset Collection

The datasets used in this work were collected from established sources of human perceptual similarity judgments. For color, pitch, and taste, we use datasets compiled in (Marjieh et al., 2023), which aggregate similarity judgments and confusion matrices from multiple psychological studies and provide normalized dissimilarity matrices. For emotion, we use the ANEW benchmark database (Bradley & Lang), which provides human ratings along valence, arousal, and dominance (VAD) dimensions. These coordinates are converted into pairwise dissimilarities using cosine distance in the VAD space to construct the baseline dissimilarity matrix. For each modality, the resulting dissimilarity matrices are projected into low-dimensional space using MDS to obtain the corresponding human baseline geometry.

B. Additional Results

B.1. Layer-wise Geometric Emergence (MDS and Isomap)

While MDS provides a faithful reconstruction of pairwise dissimilarities, it is important to verify that observed structures are not artifacts of the embedding method. Therefore, we additionally employ Isomap, a nonlinear dimensionality reduction technique that preserves geodesic distances on the underlying manifold. Consistency between MDS and Isomap visualizations provides stronger evidence that the observed geometric structures reflect intrinsic properties of the representation space rather than projection artifacts.

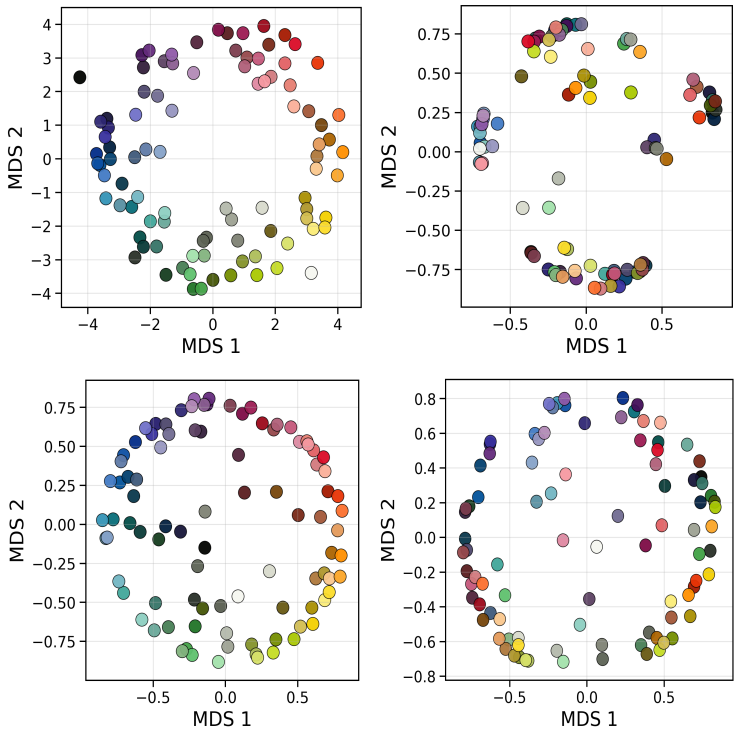

Figure 6. Layer-wise emergence of 2D color geometry in LLaMA-3-8B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

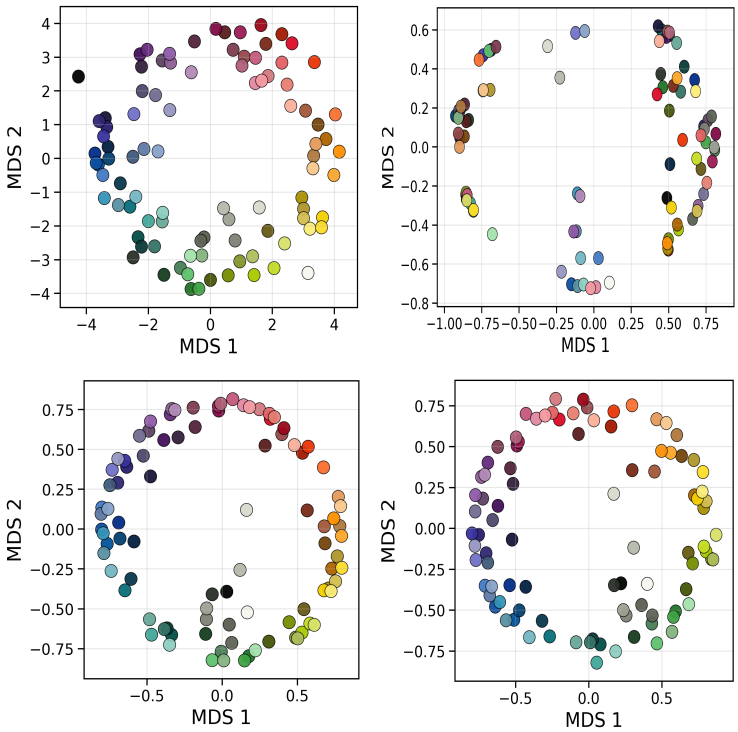

Figure 7. Layer-wise emergence of 2D color geometry in Qwen-3-4B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

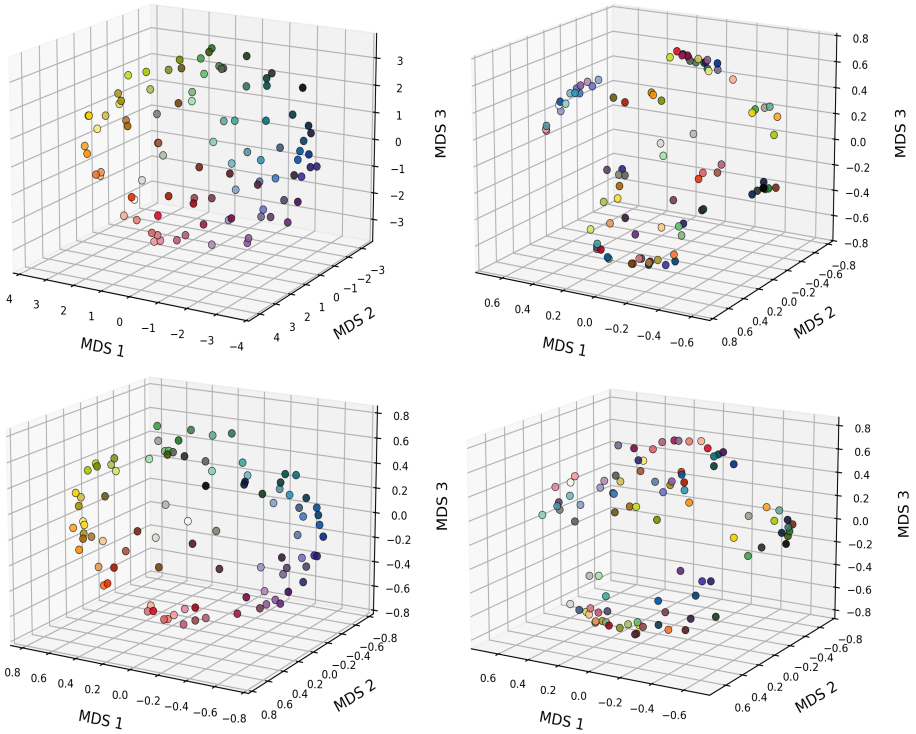

Figure 8. Layer-wise emergence of 3D color geometry in LLaMA-3-8B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

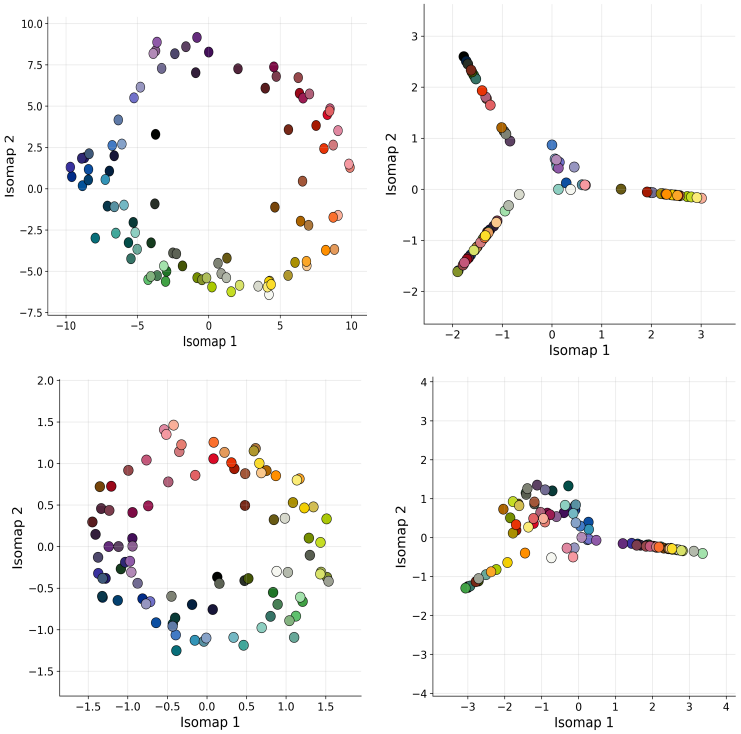

Figure 9. Layer-wise emergence of 2D color geometry in LLaMA-3-8B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer isomap representation.

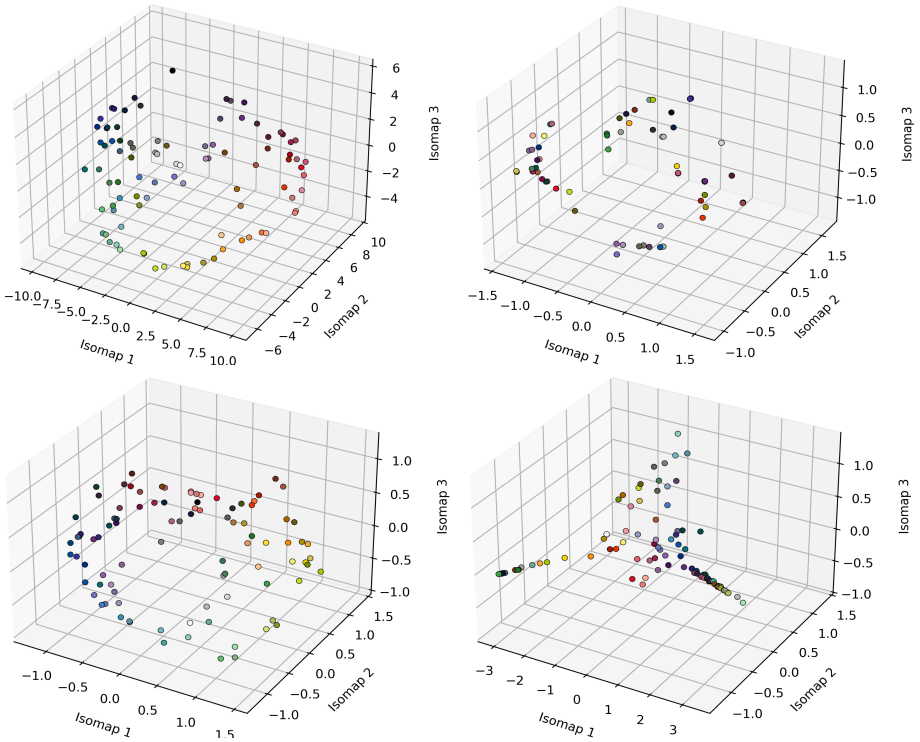

Figure 10. Layer-wise emergence of 3D color geometry in LLaMA-3-8B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer isomap representation.

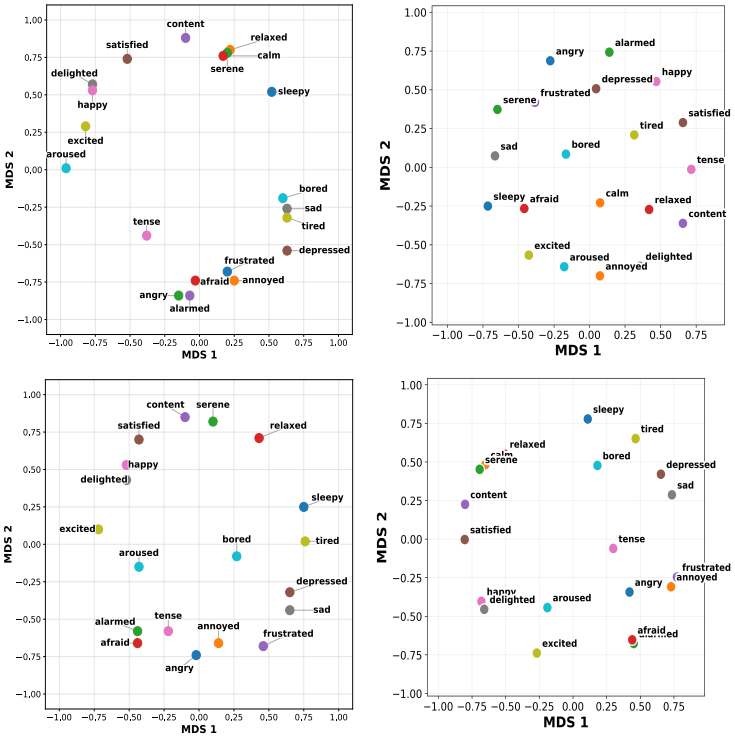

Figure 11. Layer-wise emergence of emotion geometry in Gemma-7B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

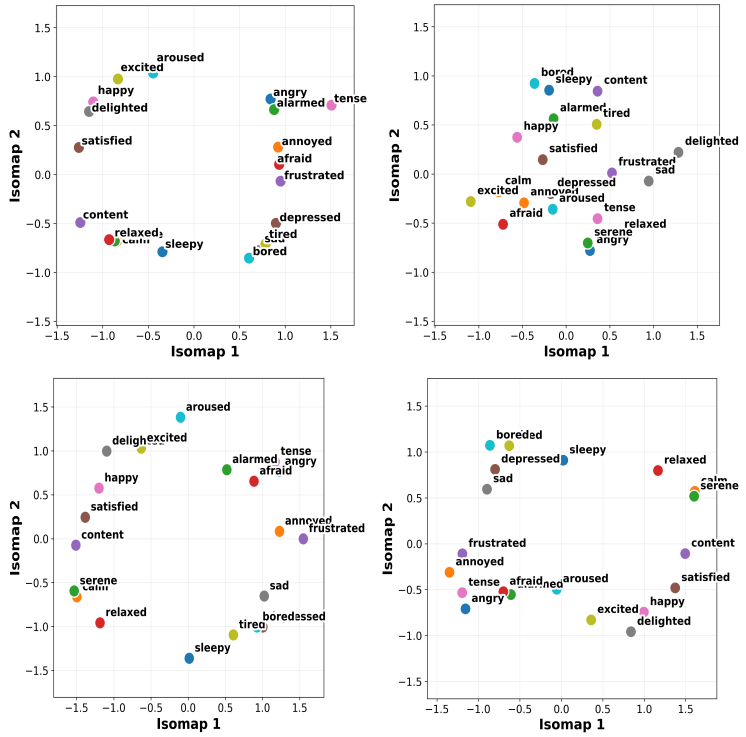

Figure 12. Layer-wise emergence of emotion geometry in Gemma-7B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer isomap representation.

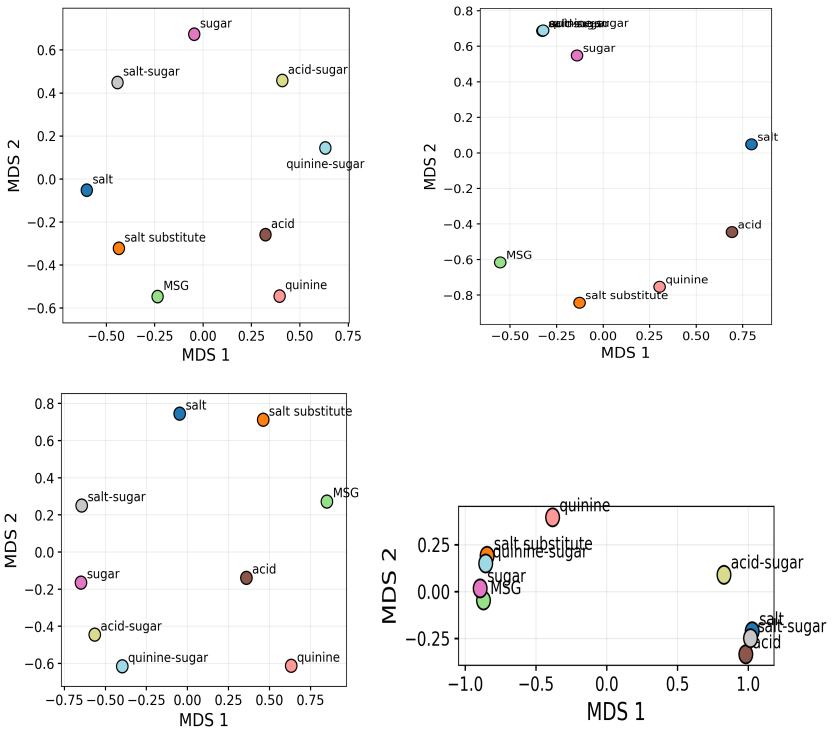

Figure 13. Layer-wise emergence of Taste geometry in Gemma-7B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

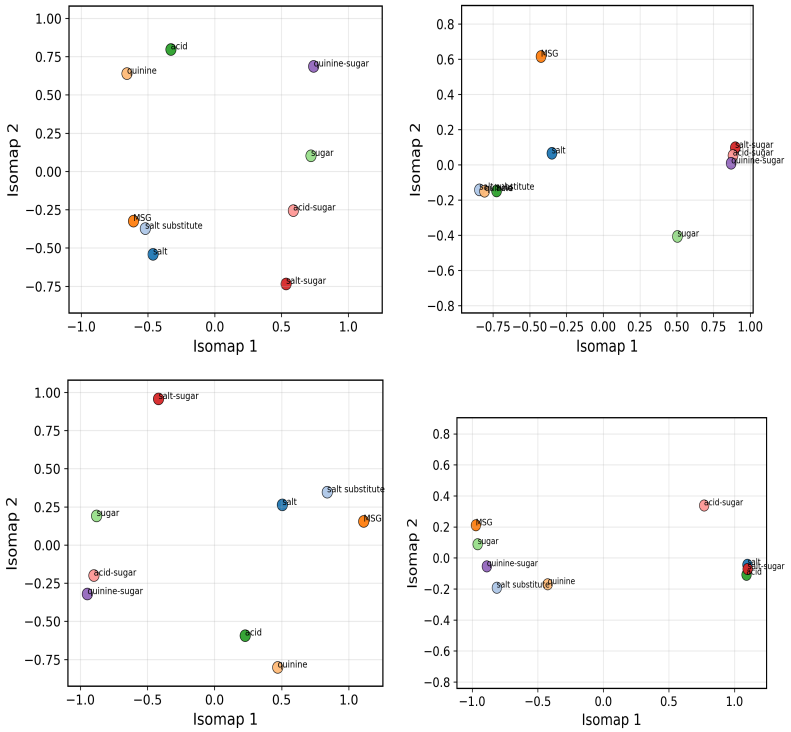

Figure 14. Layer-wise emergence of Taste geometry in Gemma-7B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer isomap representation.

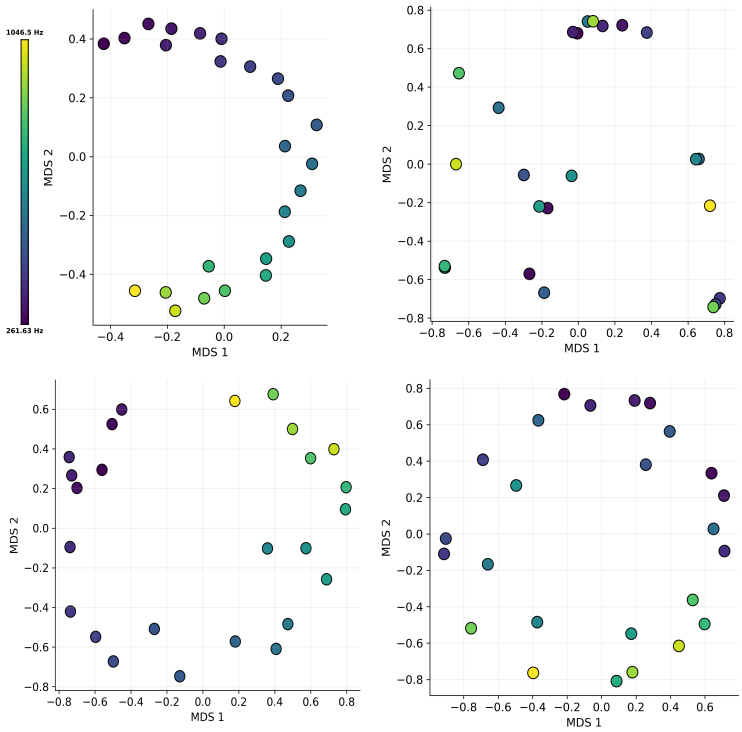

Figure 15. Layer-wise emergence of 2D Pitch geometry in Qwen-3-4B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

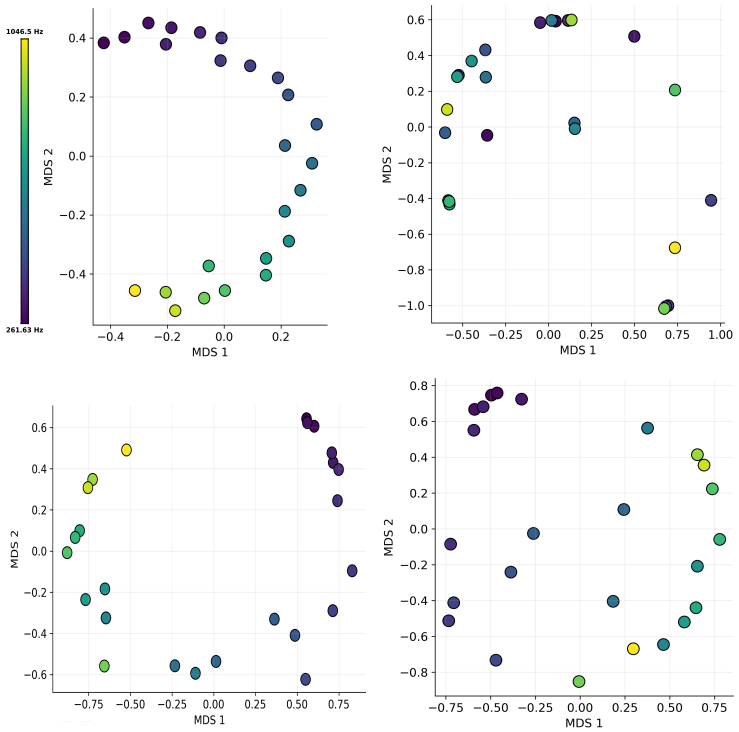

Figure 16. Layer-wise emergence of 2D Pitch geometry in Gemma-7B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

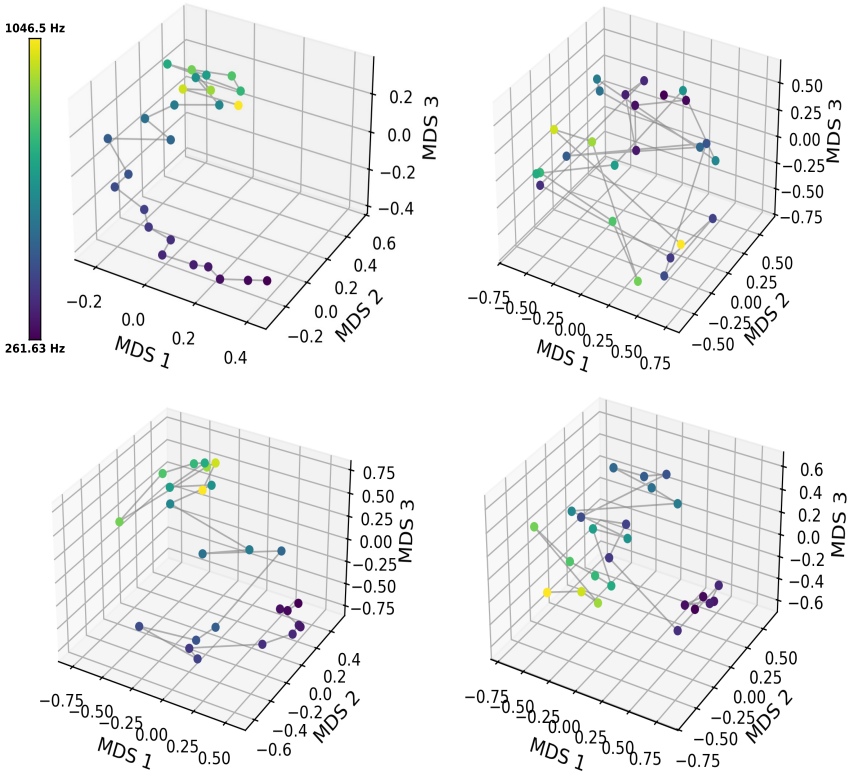

Figure 17. Layer-wise emergence of 3D Pitch geometry in Qwen-3-4B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

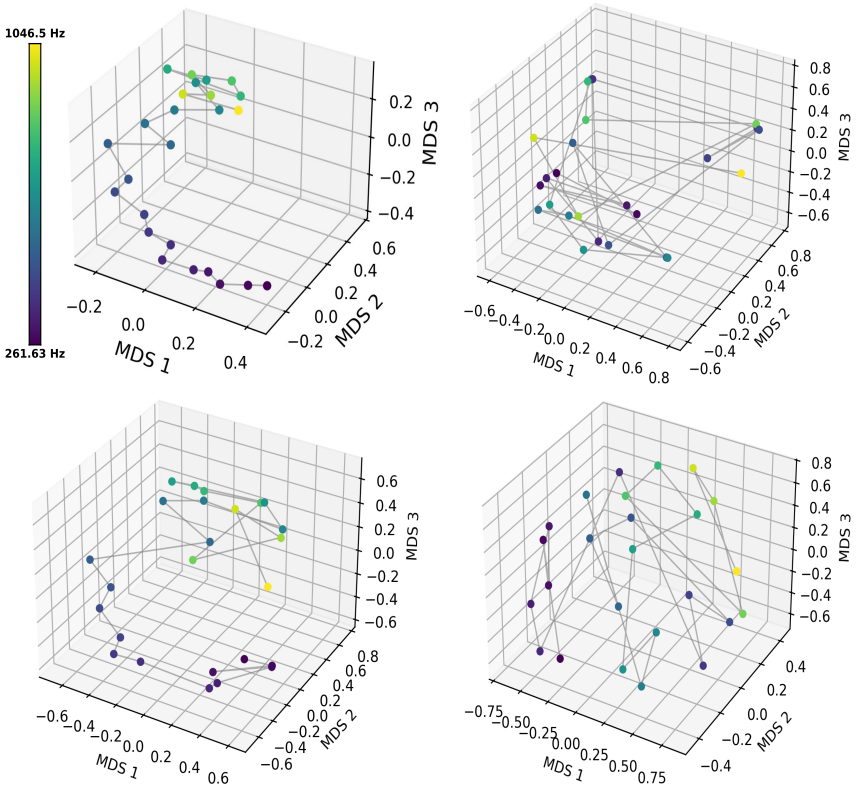

Figure 18. Layer-wise emergence of 3D Pitch geometry in Gemma-7B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer MDS representation.

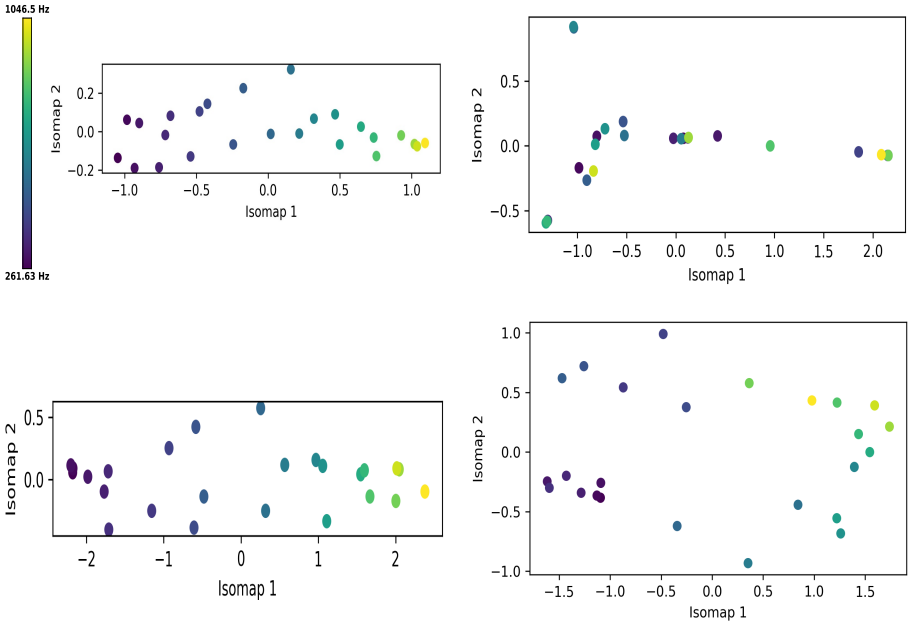

Figure 19. Layer-wise emergence of 2D Pitch geometry in Qwen-3-4B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer isomap representation.

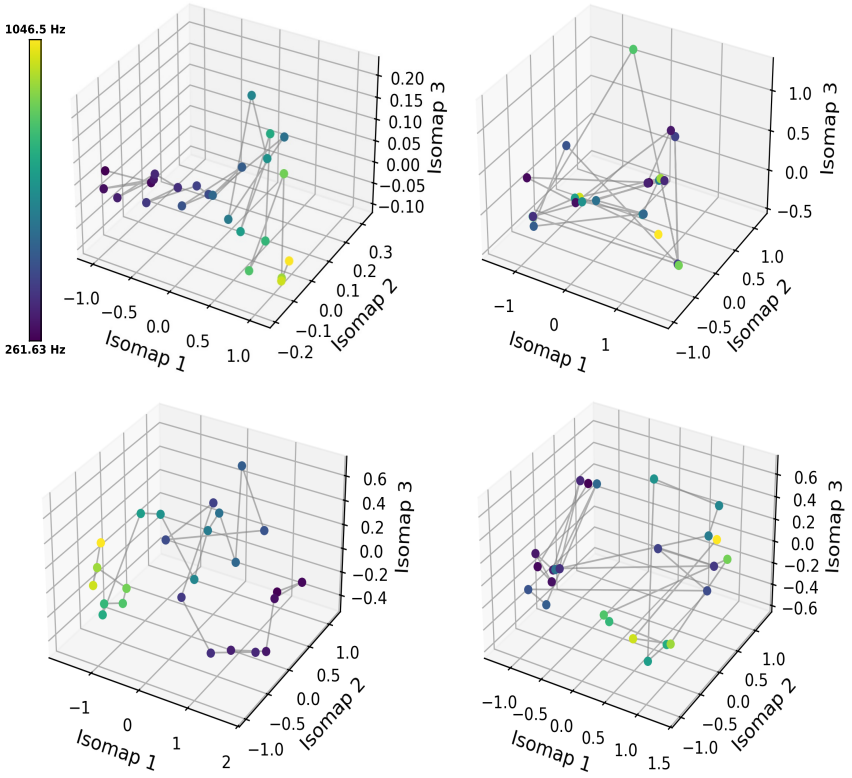

Figure 20. Layer-wise emergence of 3D Pitch geometry in Qwen-3-4B across model depth. Panels show (top-left) human perceptual geometry, (top-right) early-layer representation, (bottom-left) peak-alignment layer, and (bottom-right) final-layer isomap representation.

B.2. Layer-wise Metric Profiles

To complement the qualitative geometric visualizations, we report additional layer-wise alignment profiles for each stimulus domain.

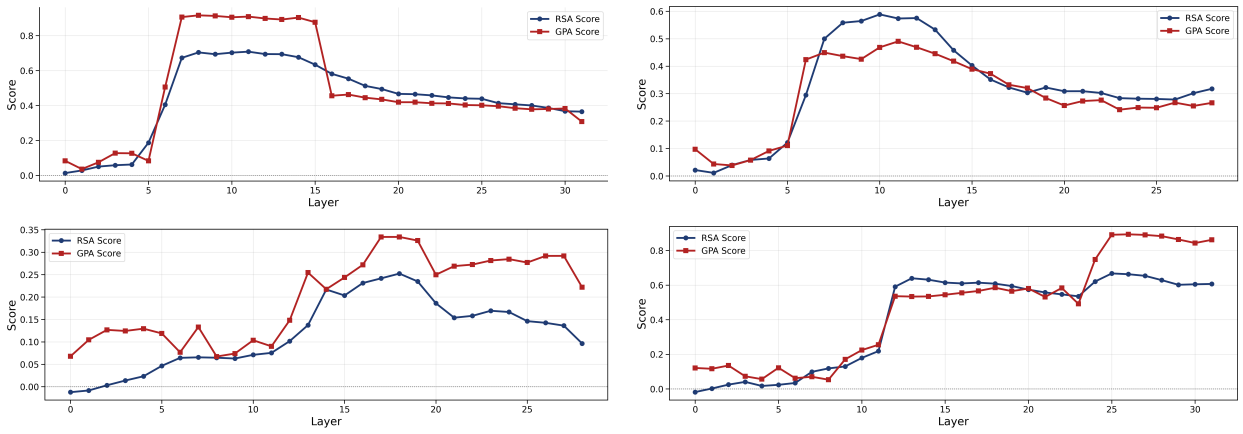

Figure 21. Layer-wise metric profiles for color across four models: (Llama-3-8B) model 1, (Llama-3.2-3B) model 2, (Gemma-7B) model 3, and (Qwen-3-4B) model 4.

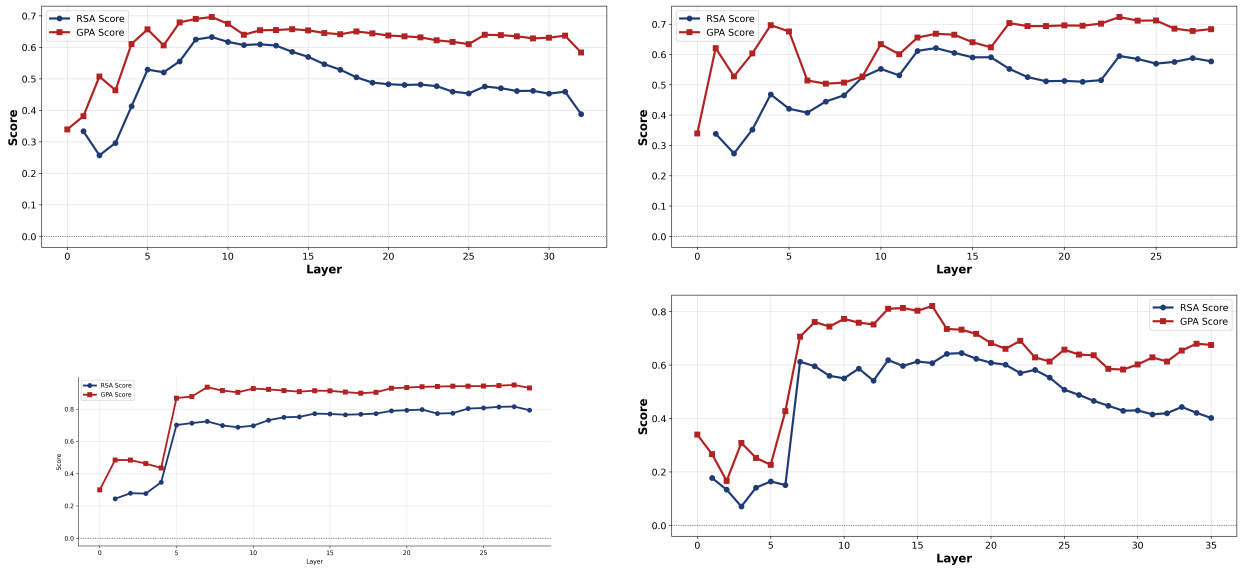

Figure 22. Layer-wise metric profiles for Emotion across four models: (Llama-3-8B) model 1, (Llama-3.2-3B) model 2, (Gemma-7B) model 3, and (Qwen-3-4B) model 4.

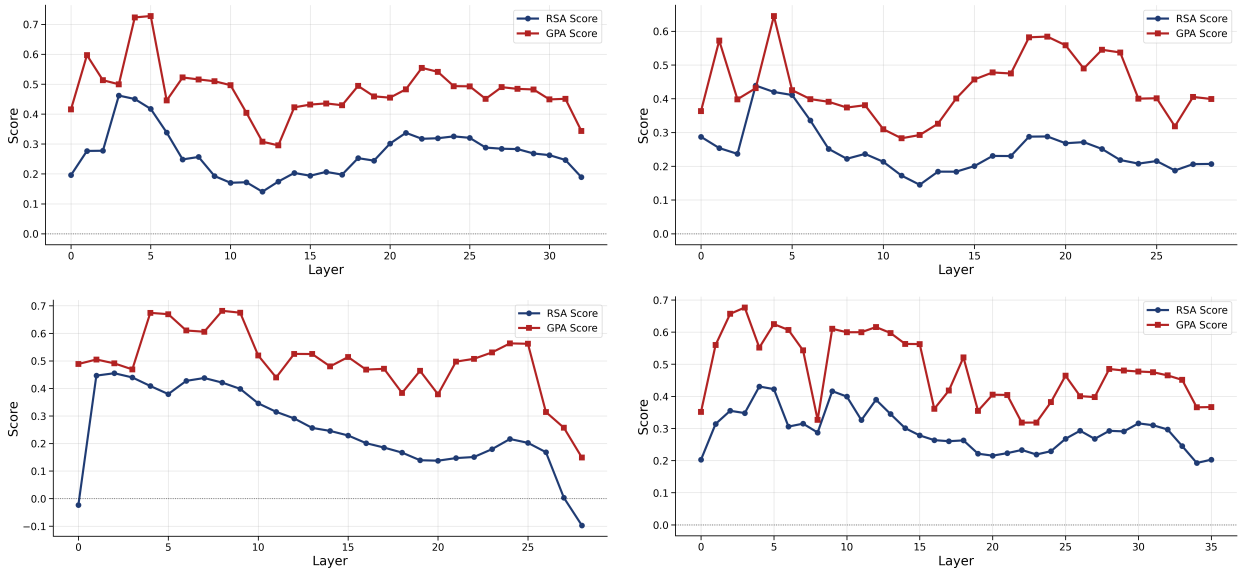

Figure 23. Layer-wise metric profiles for Taste across four models: (Llama-3-8B) model 1, (Llama-3.2-3B) model 2, (Gemma-7B) model 3, and (Qwen-3-4B) model 4.

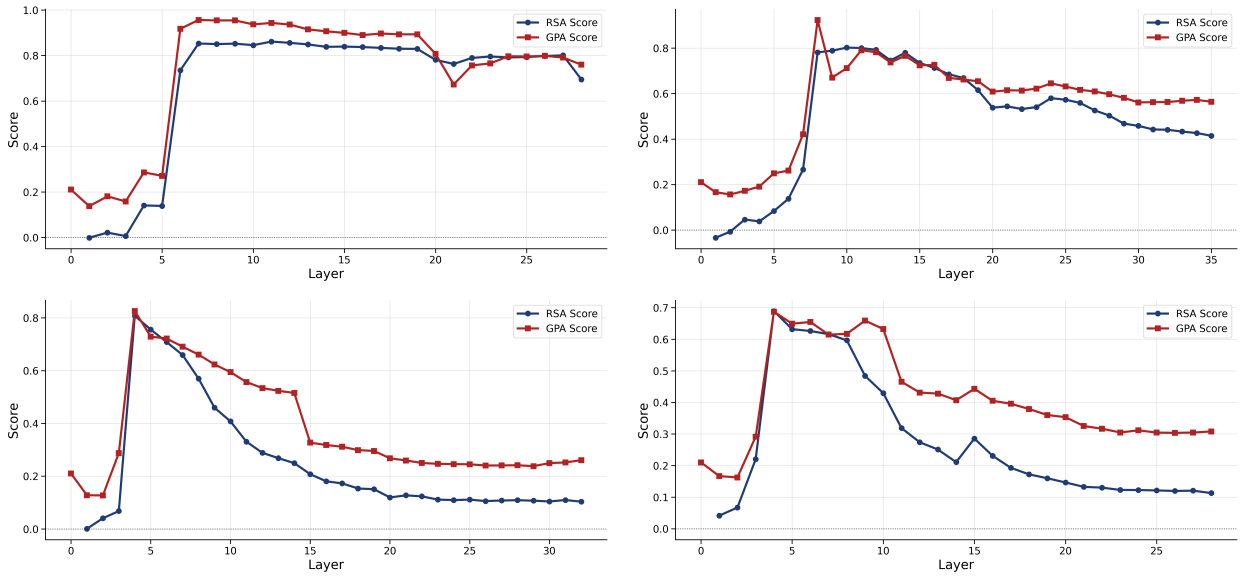

Figure 24. Layer-wise metric profiles for Pitch across four models: (Gemma-7B) model 1, (Qwen-3-4B) model 2, (Llama-3-8B) model 3, and (Llama-3.2-3B) model 4.

B.3. Bootstrap Confidence Intervals for Layer-wise Alignment

To assess the statistical stability of the observed layer-wise alignment profiles, we estimate bootstrap confidence intervals for both RSA and GPA scores by resampling stimuli with replacement at each layer. For each perceptual domain and each layer, we perform 1000 bootstrap iterations. In every iteration, the stimulus set is resampled with replacement, the corresponding human and model dissimilarity submatrices are extracted, and RSA and GPA are recomputed on the resampled set. We report 95% percentile bootstrap confidence intervals for both metrics, allowing us to evaluate whether the observed layer-wise trends are robust to the particular choice of stimuli rather than driven by a small subset of labels.

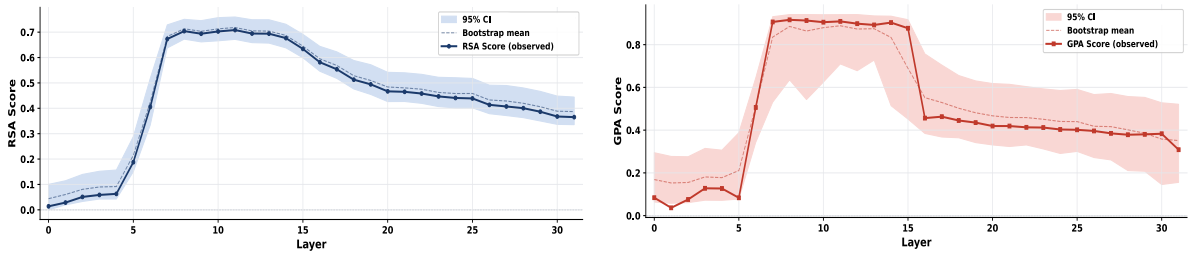

Figure 25. Bootstrap confidence intervals for color in LLaMA-3-8B, showing layer-wise RSA (left) and GPA (right) profiles with 95% confidence bands.

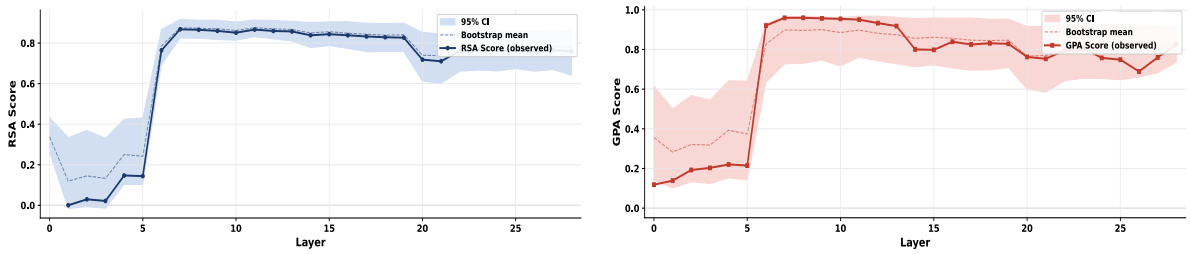

Figure 26. Bootstrap confidence intervals for Pitch in Qwen-3-4B, showing layer-wise RSA (left) and GPA (right) profiles with 95% confidence bands.

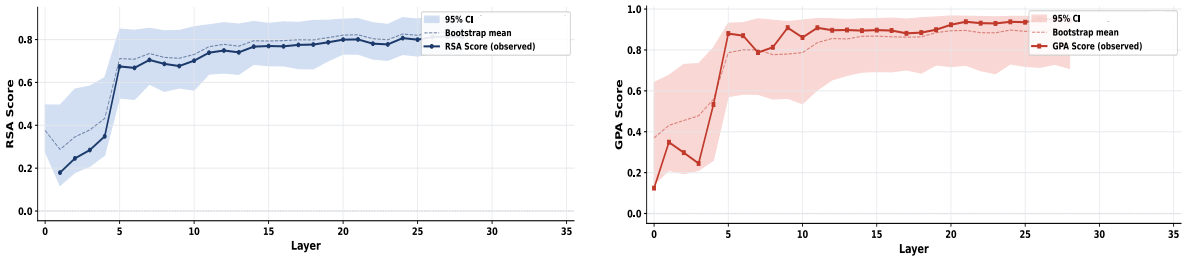

Figure 27. Bootstrap confidence intervals for emotion in Gemma-7B, showing layer-wise RSA (left) and GPA (right) profiles with 95% confidence bands.

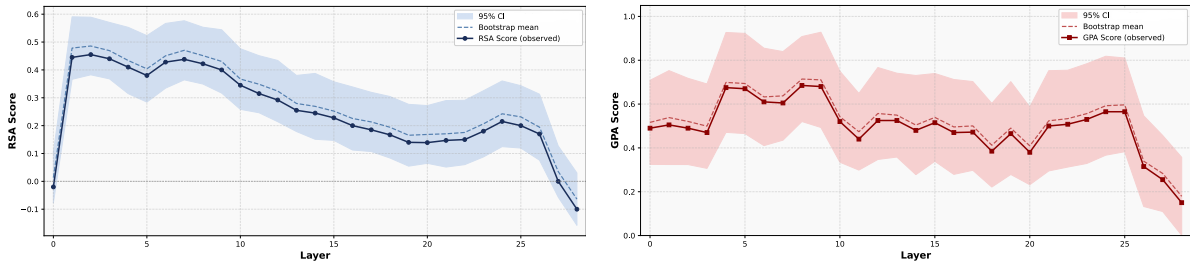

Figure 28. Bootstrap confidence intervals for taste in Gemma-7B, showing layer-wise RSA (left) and GPA (right) profiles with 95% confidence bands.