Extended Abstract Track

Neural Representational Geometry of Concepts in Large Language Models

Editors: List of editors' names

Abstract

Despite tremendous successes of large language models (LLMs), their internal neural representations remain opaque. Here we characterize the geometric properties of language model representations and their impact on few-shot classification of concept categories. Our work builds on Sorscher et al. (2022)'s theory, previously used to study neural representations in the vision domain. We apply this theory to embeddings obtained at various layers of a pre-trained LLM. We mainly focus on LLaMa-3-8B, while also confirming their applicability to OpenAI's text-embedding-3-large. Our study reveals geometric properties and their variations across layers that are unique to language models, and provides insights into their implications for understanding concept representation in LLMs.

1. Introduction and Methodological Background

Large language models (LLMs) have recently demonstrated impressive capabilities in manipulating human language (Brown et al., 2020). Elucidating how they internally represent various linguistic concepts is one of the central questions to understanding their mechanisms.

A recent study by Sorscher et al. (2022) proposed a geometric theory of few-shot learning and applied it to the visual domain, demonstrating that the performance of prototype-based classifiers can be predicted by four key geometric properties of concept manifolds in the representational space of neural networks trained for image classification.

Here we apply this theory to study concept representations in LLMs. By analyzing the geometry of concept embeddings and its implications for few-shot category classification performance, we aim to provide a theoretical foundation for understanding geometrical representations of concepts in language models.

Method. Our work builds on Sorscher et al. (2022)'s theory of representational geometry of concepts, which states that few-shot learning performance of the prototype-based classifier is governed by four simple and readily measurable geometric quantities. Given two concepts (a, b), with m example representations of dimension N for each concept, their theory predicts that the average classification error of a test example of concept a is given by $E_a = H(\mathrm{SNR}_a)$, where $H(\cdot)$ is the tail function of the standard Gaussian distribution. The quantity SNR_a is the signal-to-noise ratio (SNR) for concept manifold a, whose dominant terms are given by:

$$\mathrm{SNR} = \frac{1}{2} \frac{R_a^{-2} \|\boldsymbol{x}_0^a - \boldsymbol{x}_0^b\|^2 + (R_b^2 R_a^{-2} - 1)/m}{\sqrt{(mD_a)^{-1} + m^{-1} R_a^{-4} \|(\boldsymbol{x}_0^a - \boldsymbol{x}_0^b)\boldsymbol{U}_b\|^2 + R_a^{-4} \|(\boldsymbol{x}_0^a - \boldsymbol{x}_0^b)\boldsymbol{U}_a\|^2}} \tag{1}$$

where $\boldsymbol{x}_0^a$ and $\boldsymbol{x}_0^b$ are the (unnormalized) centroids of the respective representations, and

$$\boldsymbol{U}_a = [R_{a,1}\boldsymbol{u}_1^a, \ldots, R_{a,N}\boldsymbol{u}_N^a] \quad ; \quad R_a^2 = \sum_{i=1}^{N} (R_{a,i})^2 \tag{2}$$

$(\boldsymbol{u}_1^a, \ldots, \boldsymbol{u}_N^a)$ are principal axes, and $(R_{a,1}, \ldots, R_{a,N})$ are the corresponding radii. We refer to Sorscher et al. (2022) for further technical details and derivations (note that equations in their main text contain several typos).

Both the SNR and generalization error for few-shot learning is governed by the following key quantities.

- Signal: $R_a^{-2}\|\boldsymbol{x}_0^a - \boldsymbol{x}_0^b\|^2$ represents the pairwise distance between the manifolds' centroids.
- Bias: $R_b^2 R_a^{-2}$ represents the average bias of the linear classifier.
- Dimension: $D_a = \left(\sum_{i=1}^N R_{a,i}^2\right)^2 / \left(\sum_{i=1}^N R_{a,i}^4\right)$ measures the number of dimensions along which the concept manifold varies significantly (called "participation ratio").
- Signal-noise overlap: $\|(\boldsymbol{x}_0^a - \boldsymbol{x}_0^b)\boldsymbol{U}_a\|^2$ and $\|(\boldsymbol{x}_0^a - \boldsymbol{x}_0^b)\boldsymbol{U}_b\|^2$ represents the overlap between the signal direction and the manifold axes of variation.

In our study, the "representations" are obtained by embedding sentences representing the corresponding concept using a large language model. We provide further details in the following section.

2. Experimental Results

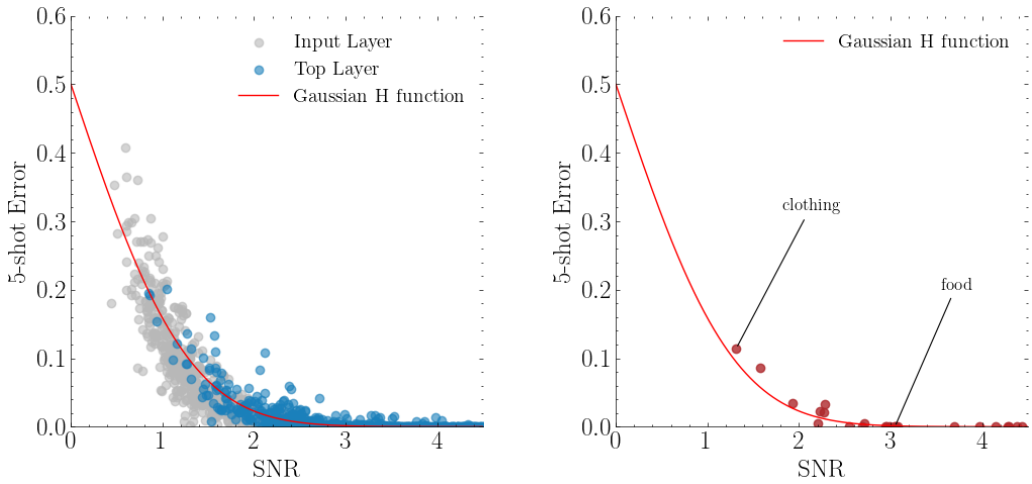
Basic Settings. Here we describe how we apply the theory above to a language model. We conduct our main experiments using LLaMa-3-8b (Touvron et al., 2023), which is a 32-layer auto-regressive model with the embedding dimension of $N = 4096$.

We first define "concepts" by curating a set of 25 categories (e.g., 'plants' and 'beverages') and use GPT-4 (Achiam et al., 2023) to generate 200 representative sentences for each category (e.g., 'The poppy flowers sway in the breeze with their delicate petals.' for the category 'plants'). Further details can be found in Appendix A.1.

Each example representation of the concept is obtained by feeding one of these sentences to the language model; we first extract token embeddings from all the positions at the given layer (at the output of the feedforward block; Vaswani et al. (2017)), and then average them to obtain a single vector. The resulting collection of sentence embeddings for a given category forms what we call a "concept manifold".

Few-shot Concept Classification. For pair-wise classification, we employ the simple prototype learning rule as in Sorscher et al. (2022), i.e., we compute the mean of the training examples (called prototypes), and a test sentence is classified by comparing the distance of its embedding to each prototype and assigning it to the nearest one.

Here we apply this method to LLaMa-3-8b (Touvron et al., 2023). Our experiments on 25 categories show that prototype learning achieves an average test accuracy of 98.6% using the top-layer representations, using only $m = 5$ training examples per concept. Notably, the method achieves an average test accuracy of 92.6% using the *input* layer representation, suggesting that even early representations contain substantial information about the concepts they encode. This contrasts with results in the visual domain, where representations at the input layer are much less representative of the abstract concept.



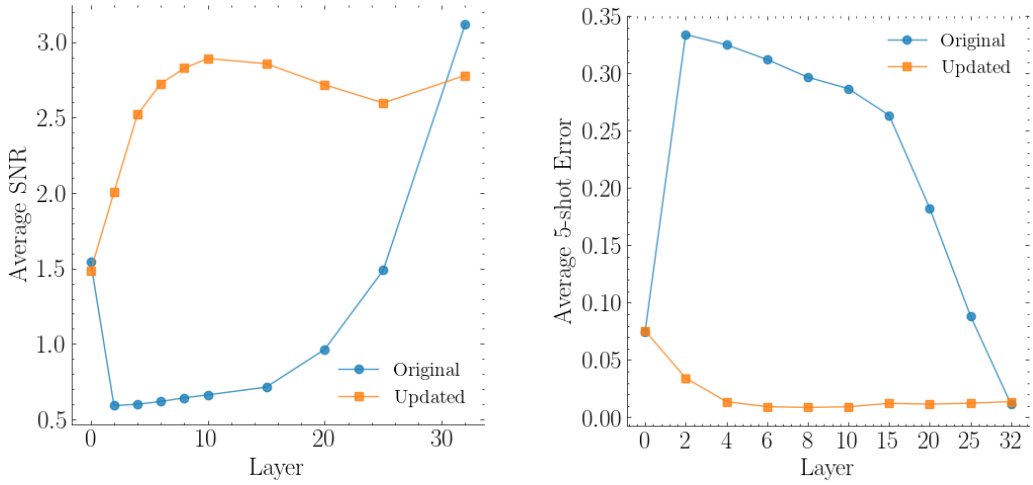
(a) All manifold pairings (b) Focus on 'Jewelry' manifold

Figure 1: Few-shot Concept Classification Error vs SNR

Error Prediction by the Geometric Theory. Here we confirm that the empirical classification errors show strong agreement with the theory. The corresponding results are shown in Figure 1. Figure 1(a) displays the generalization error as a function of SNR for all concept pairs, at the input and top layers. Each point represents the average generalization error of one pair of concepts. In Figure 1(b), we focus on the generalization error for one category, 'Jewelry', against all others, at the top layer. We find that discriminating 'jewelry' concept from 'clothing' is easier than from 'food', which is intuitive. Analogous figures for more concept categories can be found in Appendix A.2. In Appendix A.3, we provide ablation studies on the effect of number of training examples and total samples.

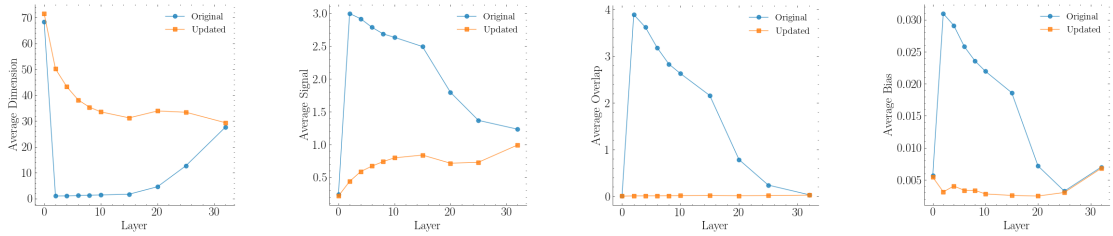
Analyzing Representations across Layers. Given the observed increase in SNR (and respective reduction in prediction error) from the input to the top layer, it might be expected that these trends would exhibit monotonic behavior across intermediary layers. However, our analysis reveals a more complex pattern: SNR initially decreases in the intermediate layers before rising sharply, while error undergoes an initial surge before subsequently dropping (Fig. 2(a), "Original"). To investigate this non-monotonic behavior, we examine each underlying geometrical component of the data manifolds (Fig. 3).

Our findings indicate a marked reduction in *dimensionality* (Fig. 3(a), "Original") within the initial layers, followed by a significant expansion in the final layers. This shift in dimensionality provides insight into the observed fluctuations in SNR and error. To explain this effect, we explore the correlation between sentence length—a very generic property of language data—and the principal component scores, focusing on the component associated with the largest eigenvalue of the entire dataset. We identified a significant Pearson correlation score of over 0.97 in the intermediate layers, suggesting that sentence length is a major factor contributing to the variance in these layers. We observed a smaller correlation of 0.35 at the input layer and 0.62 at the top layer (Figure 8 in the appendix). To mitigate the impact of sentence length, we generate adjusted embeddings by subtracting the projection



(*a*) SNR (*b*) Error

Figure 2: SNR and Error as a function of depth/layer



(*a*) Dimension (*b*) Signal (*c*) Overlap (*d*) Bias

Figure 3: Geometrical quantities as a function of depth. Before ('Original') and after ('Updated') substraction of the first principal component (Sec. 2).

of the original embeddings onto the primary principal component. With these adjusted embeddings, we see a much smaller reduction in dimensionality (Figure 3(*a*), 'Updated'). Note that changes in SNR and error across layers are still not monotonic (Figure 2(*a*)); finding an explanation for this trend requires further exploration.

Analysis using another model. We also conducted similar analysis using OpenAI's text-embedding-3-large (OpenAI, 2024). Results can be found in Appendix A.4.

3. Conclusion

We studied neural representations of concepts in large language models using tools from the geometrical manifold theory of Sorscher et al. (2022). We demonstrated that this theory accurately predicts classification errors of a prototype-based classifier based on sentence embeddings from a pre-trained language model, and we characterized representations in different layers through key geometrical quantities provided by the theory. We hope to extend this study to classification of "novel" linguistic concepts using LLMs in future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *Preprint arXiv:2303.08774*, 2023.

Tom B Brown et al. Language models are few-shot learners. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Virtual only, December 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. North American Chapter of the Association for Computational Linguistics on Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, MN, USA, June 2019.

OpenAI. New embedding models and API updates. [Available Online] : https://openai.com/index/new-embedding-models-and-api-updates/, 2024.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proc. National Academy of Sciences (PNAS)*, 119 (43), 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *Preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, Long Beach, CA, USA, December 2017.

Appendix A. Extra Experimental Details and Results

A.1. List of concepts

The full list of 25 concepts used in our study is: animals, furniture, food, sports, clothing, professions, plants, electronics, jewelry, transportation, music, beverages, literature, countries, buildings, work tools, body parts, games, weather, mythical creatures, natural phenomena, historical events, celestial bodies, art movements, and culinary techniques.

Representative sentences about a specific category were generated by using GPT-4 with the following prompt: "Write 200 varied sentences about [category] (either about a specific example of [category] or about [category] in general, but do not use the word [category]). Make these normal, simple sentences that can easily classified to be about [category]".

These initial sets of 200 sentences were refined to 170 sentences per category, by ensuring no sentence repetition and removing any sentences that were ambiguous or could reasonably belong to multiple categories. In the current work, we did this process manually to ensure the data quality; however, in future work, we expect this could be automated by using some large language model too.

An illustration for these manifolds are presented in Figure 4.

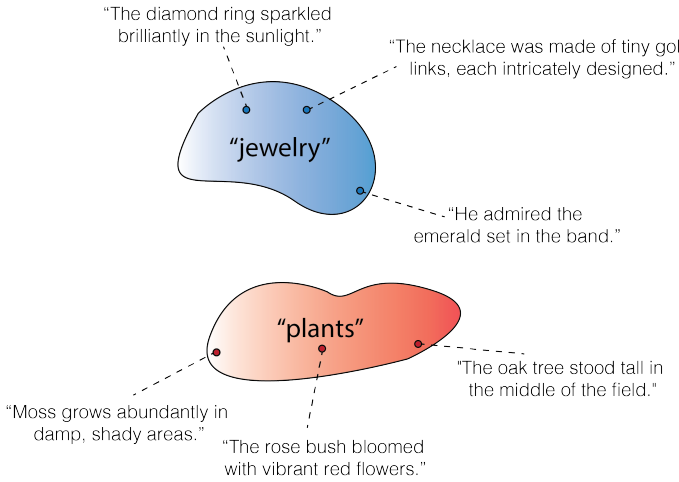


Figure 4: An illustration of a concept manifold

A.2. Error vs SNR plots for more concepts

In Figure 5, we show plots displaying error vs SNR for two more concepts.

A.3. Effect of number of training examples and total samples

Here we present the effect of changing the number of sentences used as a few training samples and the total number of sentences from which they are sampled. We find that increasing the total sample size P does not significantly alter the signal; with error, bias, and overlap following similar patterns (Figure 6). However, the dimensionality of the embedding space

exhibits a monotonic increase with P (Fig 6(b)). Based on these findings, we use all $P = 170$, which appears near the inflection point where dimensionality begins to level off.

As the number of training examples m increases, the estimated prototypes converge more closely to the true centroids of the underlying manifold. However, even with precise centroid estimation, the inherent signal-to-noise ratio remains non-negligible, leading to a generalization error that asymptotically approaches a finite value rather than vanishing entirely. This behavior aligns well with the empirical results presented in Figure 6(c). For all the few-shot classification experiments, we use $m = 5$.

A.4. Results for the OpenAI Embedding Model

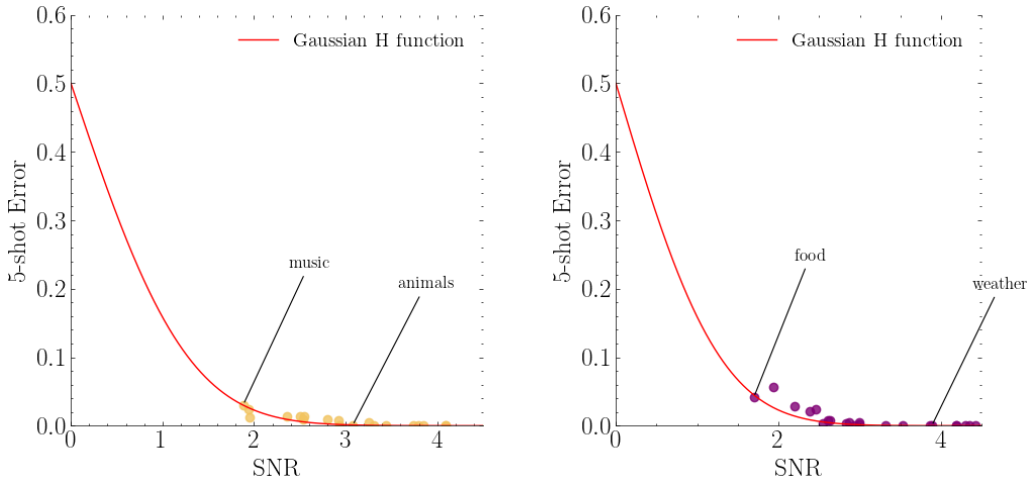
We selected the LLaMa-3-8b model for the main study due to its widespread adoption and open-source availability. As it is an autoregressive language model, it is not specifically optimized to generate "sentence embeddings" in the same manner as models such as BERT (Devlin et al., 2019) or other dedicated embedding models. To account for this limitation, we also compared the performance of LLaMa-generated embeddings with those from OpenAI's text-embedding-3-large model (OpenAI, 2024). Both models were used to encode identical sentences across the same categories, followed by an evaluation in the few-shot classification setting. Interestingly, the embeddings produced by OpenAI's model did not exhibit a substantial performance improvement over those generated by LLaMa (Fig. 7); both mean error and mean SNR were found to be smaller for OpenAI's model embeddings.

A.5. Exploring alternative embedding/manifold construction methods

In the main text, we defined the sentence embedding by averaging token embeddings across all positions in the sentence. This is a reasonable yet somewhat arbitrary choice. For the sake of completeness, here we present two additional methods for constructing embeddings (and therefore the manifolds), which we studied in our preliminary experiments. Our sole goal is to show their qualitative trends (they differ in too many details to be compared side by side).

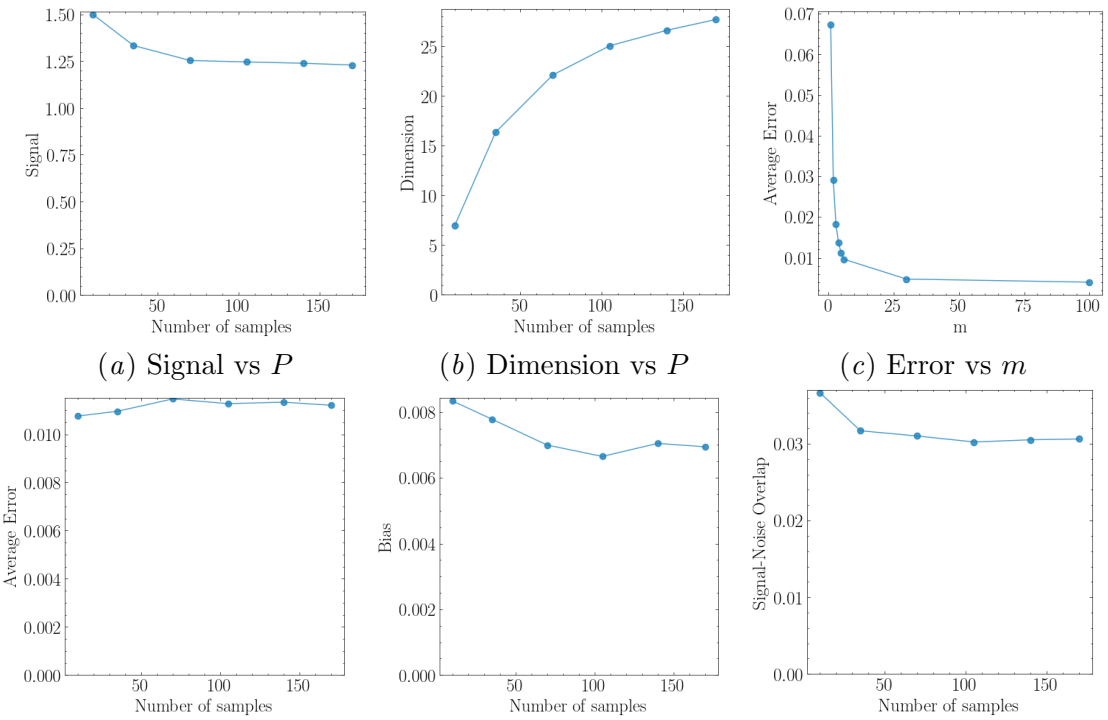
Next Token Manifold (NTM). The Next Token Manifold is built from categories from six categories: *dog*, *apple*, *painting*, *soccer*, *earring*, *fork*. Each category is made up of sentences that are incomplete, ending just before the category word (e.g., *He always keeps a leash for his*). The categories and sentences are distinct from those used to define the concept manifolds in the main text. The embedding of the last token is used to represent the sentence. The corresponding results are shown in Figures 9(a) and 10. The NTM displays a distinct geometry, characterized by a nearly monotonic decrease in error and a monotonic increase in SNR, aligning closely with trends observed in the vision domain (Sorscher et al., 2022).

Masked Average Manifold (MAM). The Masked Average Manifold uses the same categories as NTM but in sentences for MAM, the category word is masked (e.g., *The loyal [MASK] wagged its tail.*), and embeddings are averaged across all tokens. The results are shown in Figures 9(b) and 11. The MAM setting exhibits significantly lower top-layer SNR and higher error compared to both the NTM and "category manifold" (used in main text) settings, with similar non-monotonic changes in geometry across layers as the category manifold setting, suggesting that token averaging may cause this effect.



(*a*) 'Games' manifold (*b*) 'Beverages' manifold

Figure 5: Generalization Error vs SNR for 'Games' and 'Beverages' manifolds



(*a*) Signal vs P (*b*) Dimension vs P (*c*) Error vs m

(*d*) Error vs P (*e*) Bias vs P (*f*) Overlap vs m

Figure 6: Effect of number of total number of sentences P and training examples m used for few-shot classification on the geometrical quantities (see Sec. A.3).

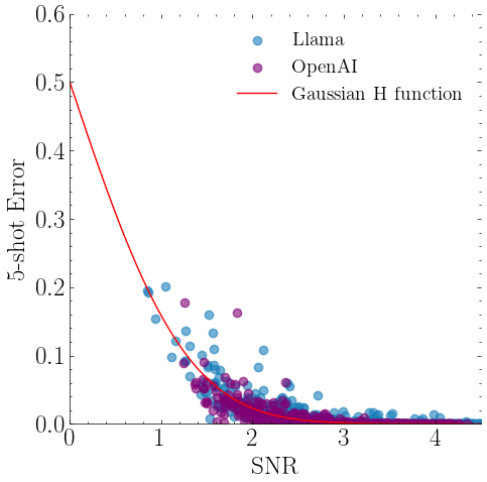


Figure 7: Error vs SNR comparing LLaMa (top layer) and OpenAI embeddings

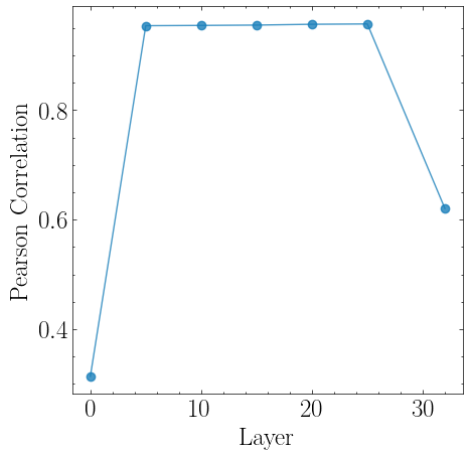
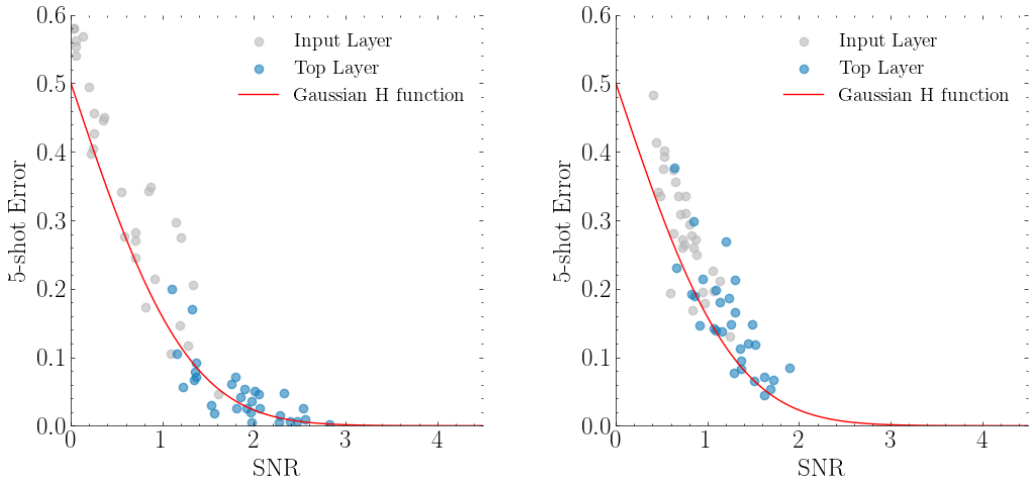
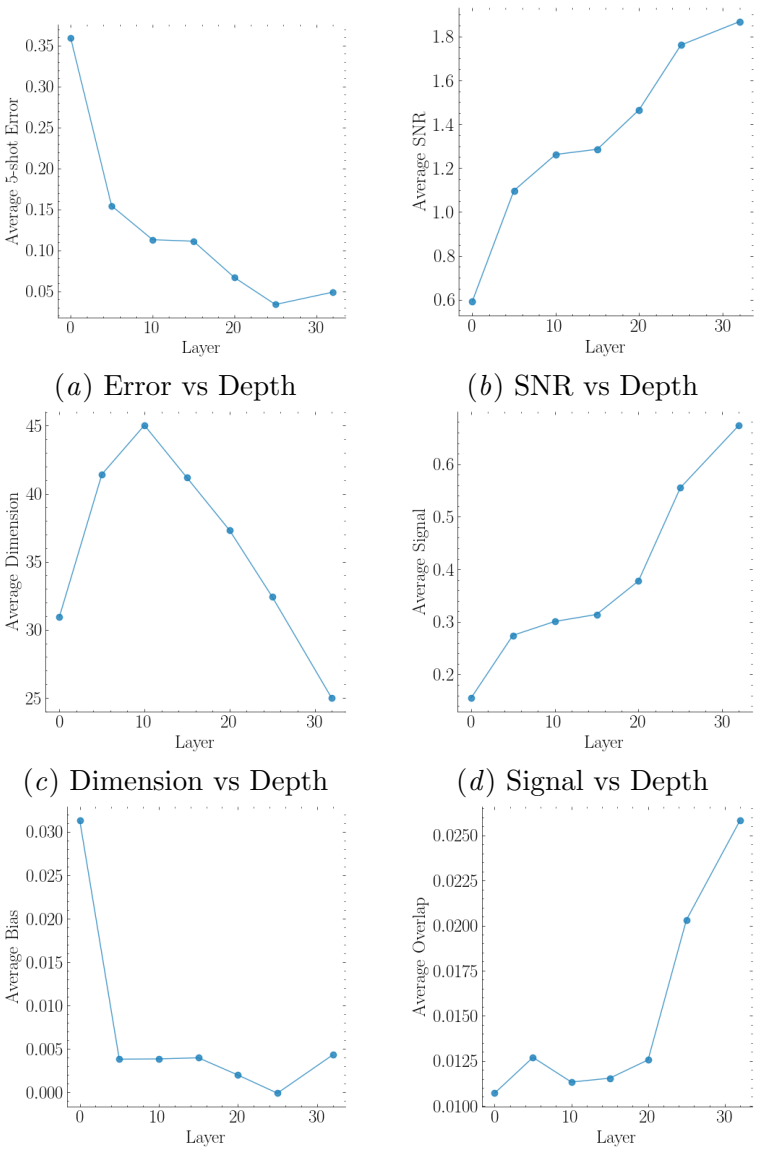


Figure 8: Pearson correlation scores between sentence length and the dominant principal component as a function of depth/layer



(*a*) **Next Token Manifold** (*b*) **Masked Average Manifold**

Figure 9: Generalization Error vs SNR using alternative manifold construction methods (Appendix A.5)

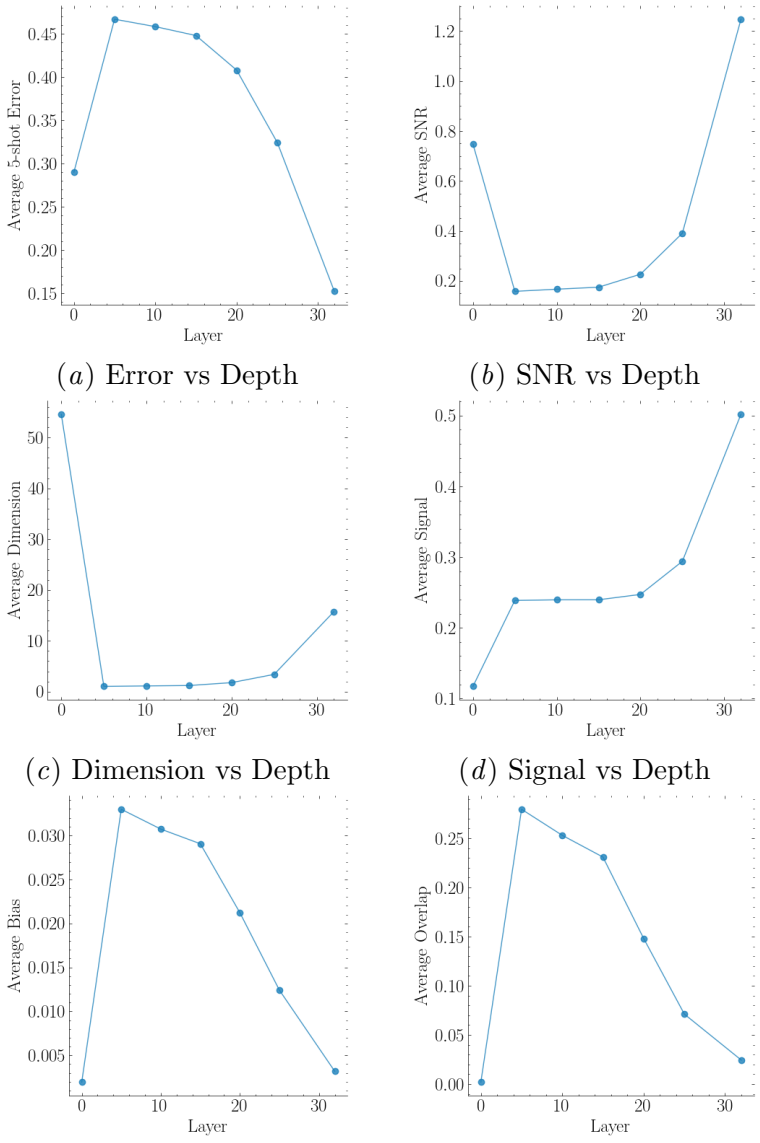


(*a*) Error vs Depth (*b*) SNR vs Depth

(*c*) Dimension vs Depth (*d*) Signal vs Depth

(*e*) Bias vs Depth (*f*) Overlap vs Depth

Figure 10: **Next Token Manifold**: Geometrical quantities as a function of depth



(*a*) Error vs Depth (*b*) SNR vs Depth

(*c*) Dimension vs Depth (*d*) Signal vs Depth

(*e*) Bias vs Depth (*f*) Overlap vs Depth

Figure 11: **Masked Average Manifold**: Geometrical quantities as a function of depth

Extended Abstract Track