Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 1565–1575
https://doi.org/10.54941/ahfe1005724

Analyzing Large Language Model Behavior via Embedding Analysis

Sourya Dey[1], Michael Robinson[2], Shauna Sweet[1], Andrew Lauziere[1], Jonathan Daugherty[1], and Cait Burgess[1]

[1]Galois, Inc., Arlington, VA 22203, USA
[2]American University, Washington, DC 20016, USA

ABSTRACT

The usage of large language models (LLMs) as a generative artificial intelligence tool is becoming increasingly widespread, yet there is limited understanding of the mechanisms by which prompts in whole or in part influence their behavior, capabilities, and limitations. In this paper, the authors conduct a mathematical and topological analysis of token embeddings – the first step in the computational workflow of LLMs. This work shows that the subspace where token embeddings lie is a stratified manifold with varying local dimension, and in those cases where semantically related tokens are co-located on a submanifold, there are non-trivial implications for model behavior. These topological and geometric findings help to explain performance aspects of different LLMs such as why the Llemma model is more likely to overfit than the GPT-2 model, yet the latter does worse at mathematical queries than the former. To the best of the authors' knowledge, this paper is among the first to conduct such research into the topological characterization of the token embedding space and analyze LLM behavior starting from first principles.

Keywords: Large language models, Generative artificial intelligence, Machine learning, Emerging technologies

INTRODUCTION

As people increasingly rely on large language models (LLMs) to inform and enrich key business workflows, it is critical that users and consumers of generative output understand the use cases for which LLMs are appropriate and the conditions under which they can be reasonably expected to be performant. Despite continuing advances in model development (Minaee, 2024), there is little mechanistic understanding of when and why LLMs perform well, or the conditions under which they do not. The internal structure and dynamics of most LLMs are difficult to examine and interpret, either because the models are proprietary (OpenAI, 2023) or because such an analysis would be computationally prohibitive due to the model's complexity (NVIDIA, 2024). Absent a clear understanding of when, why, and how these models should be used, users (and consumers) of LLMs are being asked to trust the outputs of a technology that is essentially a 'black box.'

The aim of this research is to unbox these black boxes by understanding how they work mathematically. To that end, this paper presents a

low-cost mathematical analysis of pre-trained LLMs, and connects estimated topological and geometric properties to model behaviors. The process of using a pre-trained model for text generation starts by providing to it a textual prompt, which is parsed into tokens and consequently the corresponding token embeddings are processed by the model to generate output text. Therefore, in order to develop a foundational, first principles understanding of the behavior, capabilities, and limitations of LLMs, it is crucial to understand the structure and topology of the token subspace. This is what this work aims to do. Concretely, contributions of this paper are as follows:

- Token embeddings are visualized and it is illustrated how syntactically and semantically related tokens such as numbers and capitalized words may form their own distinct clusters. This shows how pre-trained LLMs learn to differentiate between different types of tokens at the level of token embeddings.
- The dimensionality of the token subspace is estimated and it is shown that it is not a manifold, instead it is a stratified manifold.
- The dimension of the token subspace is found to be significantly lower than the latent space dimension, which explains why some models exhibit overfitting behavior.
- Manifold properties are used to explain why some models perform better at mathematical queries than others.

BACKGROUND

This work is at its core a multidisciplinary endeavor which spans topological analysis and generative artificial intelligence. It is thus important to provide a common conceptual framework and associated definitions to introduce and motivate the analysis.

Each LLM has an associated vocabulary which contains strings of characters known as *tokens*. Tokens – which can be whole words, sub-words, numbers or symbols – are obtained through a process known as byte pair encoding of a training corpus (Sennrich, 2016). Commonly used LLMs typically have tens of thousands of such tokens in their vocabulary; and this quantity will be denoted as the vocabulary size V. Each token has a *token embedding* associated with it, which is a vector that lives in an E-dimensional *latent space*. The values of the E components, or weights, of each of the V token embedding vectors are optimized during the training phase of an LLM, then frozen when the model is used in inference mode for text generation. However, not every point in this latent space is linguistically meaningful; only a subspace of it, denoted as the *token subspace*, contains the pre-trained token embeddings of the learned vocabulary.

The E-dimensional latent space of large language models is usually ascribed the Euclidean metric, which induces a topology on both the latent space and the token subspace. The number of free parameters needed to locate a point within a general topological space is one of its most fundamental intrinsic properties – this is the *dimension*. A space for which the dimension cannot change abruptly without leaving the space is a *manifold*;

on the other hand, a space which can be decomposed into pieces (*strata*) with different dimensions is a *stratified manifold*.

The findings in this paper are the results of experimentation on two open source LLMs, which are tuned for slightly different purposes, feature different vocabularies, and also vary in their latent dimensions. These models are described below:

- GPT-2 (Radford, 2019) is a general purpose LLM for text generation. It has around 125 million trainable parameters, and is not optimized for any specific application. Its vocabulary size is $V = 50257$, and latent space dimension is $E = 768$.
- Llemma-7B is a much larger model containing around 7 billion trainable parameters. As per its authors (Azerbayev, 2023), the Llemma-7B model is optimized for mathematics. The authors of this paper also noticed via experimentation that the Llemma-7B model performed significantly better than the GPT-2 model in answering mathematical queries. Its vocabulary size is $V = 32016$, and latent space dimension is $E = 4096$.

The authors have performed experiments on several other LLMs such as Mistral-7B (Jiang, 2023); however, the details and results are excluded from this paper due to brevity considerations. Note that the results on GPT-2 and Llemma-7B that are presented in this paper offer a good representation of the research conducted at the time of writing.

VISUALIZING TOKEN EMBEDDINGS AND CLUSTERING

As mentioned in the previous section, all the V token embeddings of an LLM lie in an E-dimensional latent space. As a first step in the analysis, dimensionality reduction techniques are used to reduce the token embedding vectors to two dimensions (2D), which is suitable for visualization. As recommended in (van der Maaten, 2008), principal component analysis (PCA) is first used to reduce the dimension down to 50, then t-distributed stochastic neighbor embedding (t-SNE) is used with perplexity set to 40 to reduce the dimension down to two. The values 50 and 40 were chosen using a brief hyperparameter search procedure; see (Dey, 2020) for a more detailed analysis of hyperparameter search in deep learning. Note that t-SNE is a stochastic algorithm and so applying it on given data will not produce exactly reproducible results, however, the results were observed to be qualitatively similar across multiple runs.

Figure 1 shows a 2D visualization of all 50,257 token embeddings in GPT-2's vocabulary. While most of the tokens are part of a large, visually undifferentiated cluster, there are some smaller clusters that are separated. These segregated clusters correspond to specific token types as shown in Figure 1, suggesting that during its training phase an LLM is learning very different representations of specific kinds of semantically or syntactically related tokens. In particular, note that the cluster towards the right contains exclusively numeric tokens, indicating that the learned representation of numbers in the model, in terms of the components of token embedding vectors, is quite different from that of non-numeric characters.

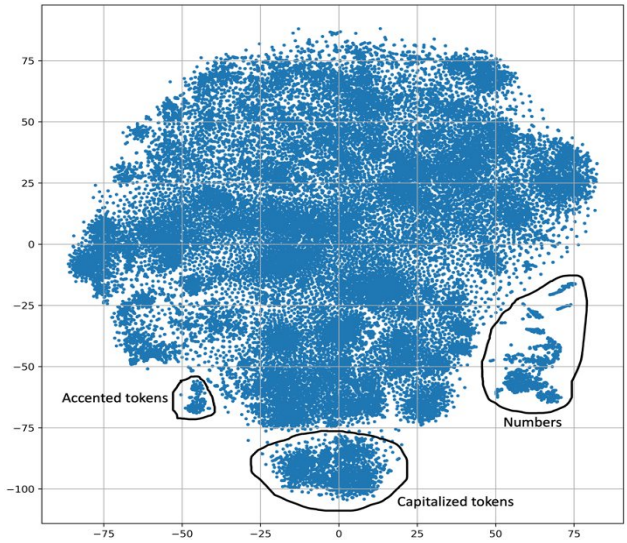

Figure 1: Visualizing GPT-2 token embeddings in 2D, with some clusters of specific token types annotated with black boundaries.

Figure 2 investigates these clusters of semantically related tokens more closely by highlighting all the numeric tokens in GPT-2's vocabulary in purple. Numeric tokens are defined to comprise a) tokens made up of only numeric characters (including superscripts, subscripts, and hexadecimal ASCII codes for numbers), e.g., '354', or b) numeric characters preceded by space, e.g., 'Ġ354', where 'Ġ' is a special character GPT-2 uses in its vocabulary to denote a leading space. There are 1694 such numeric tokens in the vocabulary. As seen in Figure 2, the vast majority of numeric tokens are in the 'archipelago' cluster towards the right.

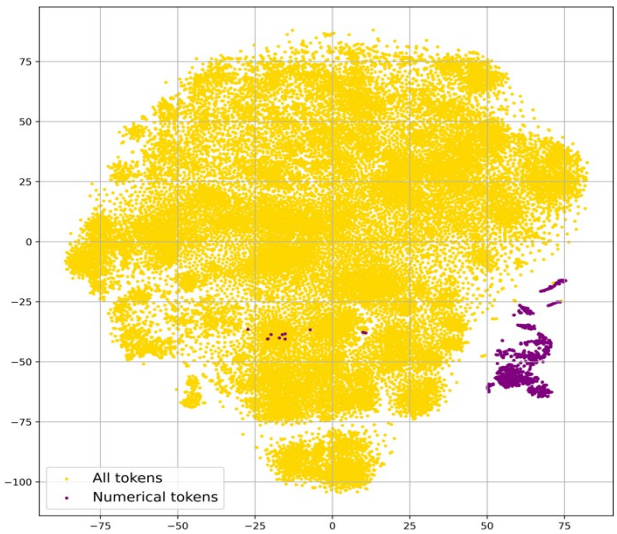

Figure 2: Visualizing GPT-2 token embeddings in 2D, with numeric tokens in purple.

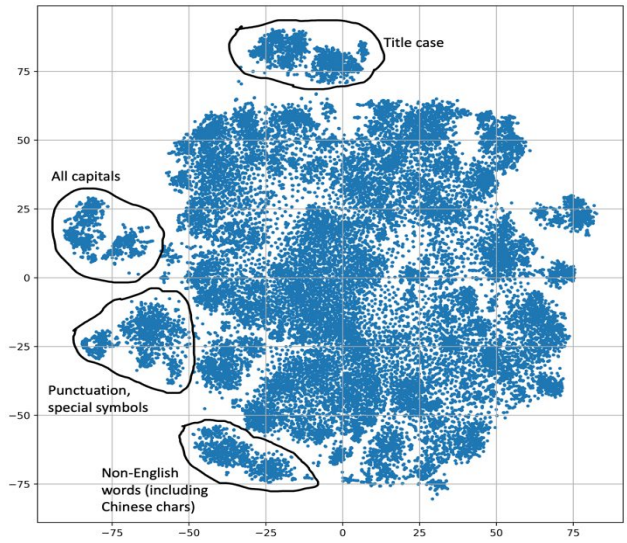

Figure 3: Visualizing Llemma-7B token embeddings in 2D, with some clusters of specific token types annotated with black boundaries.

Figure 3 is the corresponding plot to Figure 1 for the Llemma-7B LLM. This has 32,016 tokens overall, and several segregated clusters are again visible for specific types of tokens such as capitalized tokens, special symbols, or non-English characters. However, there is no separate cluster for numeric tokens. The vocabulary for Llemma-7B only has 44 numeric tokens, which are shown colored purple in Figure 4. With significantly fewer numeric tokens that are not clustered together, it is notable that Llemma-7B is designed for and achieves better performance against mathematical queries than GPT-2. The next section further explores the relationship between topological and geometric properties of the token subspace and model behavior.

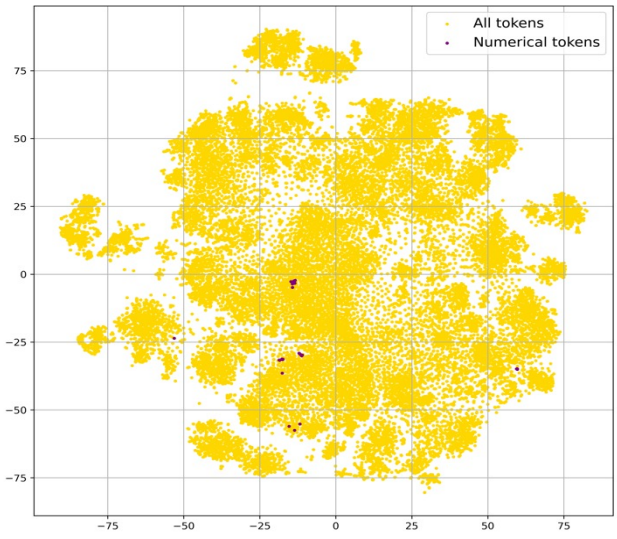

Figure 4: Visualizing Llemma-7B token embeddings in 2D, with numeric tokens in purple.

TOPOLOGICAL ANALYSIS OF THE TOKEN SUBSPACE

Estimating Dimension

As noted above, one of the most fundamental properties of the token subspace is its dimension. Because the dimensionality of the token subspace is not directly observable, the first key objective is to estimate this quantity. To do so, because the token subspace of an LLM is comprised of the embeddings of individual tokens in its vocabulary, the authors propose the following method of estimating the local dimension at different tokens to yield a distribution of local dimensionality.

In the token subspace, the volume v of a ball of radius r, in the limit of small r, is (Gray, 1974, Yomden, 2004):

$$v = Kr^n \left(1 - \frac{1}{6\,(n+2)} Ric\, r^2 + O\left(r^4\right)\right) \tag{1}$$

where n is the dimension of the token subspace, Ric is the Ricci scalar curvature, and K is the constant of proportionality. Taking the natural logarithm of both sides of the above equation yields the following asymptotic series for small r:

$$\log v = \log K + n \log r - \frac{1}{6\,(n+2)} Ric\, r^2 + O\left(r^4\right) \tag{2}$$

From here, the dimension n (and also $\log K$ and Ric) can be solved via a linear regression against pairs of radius-vs-volume values. To do so, volume needs to be estimated first. In particular, the volume of the portion of the token subspace surrounding any token is of interest. This can be done using a Monte Carlo estimation, according to which the volume v of a ball of radius r centered at token j is proportional to the number of tokens within a distance r to j. Letting $N(\cdot)$ denote cardinality of a set, this becomes:

$$v\left(r, j\right) \approx MN\left(\left\{i : \left\|i - j\right\| \leq r\right\}\right) \tag{3}$$

where M is the constant of proportionality, and $\|\cdot\|$ is a distance metric. As a point of clarification, note that for the purposes of volume and local dimension calculation, the distances considered are in the original E-dimensional space of token embeddings, not the t-SNE reduced 2D values.

Since there are V (vocabulary size) tokens in total, for a given token j, a sequence $r_{1,j} < r_{2,j} < \cdots < r_{V,j}$ of distances to the other tokens can be obtained such that $v\left(r_{k,j}; j\right) \approx Mk$. The set of $r_{i,j}$ values can be arranged in a matrix, in which rows are denoted as i and columns as j. Notice that each column of the $r_{i,j}$ matrix is sorted in ascending order. Since the sequence of volumes corresponding to each token (column) is the same, namely the sequence of integers from 1 to V, the log-linear portion of Equation (2) can be rewritten as a matrix equation for token j:

$$\begin{pmatrix} \log 1 \\ \log 2 \\ \vdots \\ \log V \end{pmatrix} + \log M \approx \begin{pmatrix} 1 & \log r_{1,j} \\ 1 & \log r_{2,j} \\ \vdots & \vdots \\ 1 & \log r_{V,j} \end{pmatrix} \begin{pmatrix} \log K_j \\ n_j \end{pmatrix} + O\left(r^2\right) \tag{4}$$

This can be readily solved via the least squares regression to obtain an estimate for the dimension n_j.

Log–log curves of volume vs radius of a ball centered at particular tokens within the token subspace are subsequently shown. The slope of any such curve, which may or may not be constant, is the local dimension at that token.

Table 1 shows the computed local dimension at all tokens present in the token subspace of the GPT-2 and Llemma-7B models. The tokens are split according to their type – either numeric or non-numeric, where the numeric tokens are as described in the previous section and shown in Figures 2 and 4 for GPT-2 and Llemma-7B, respectively. For each token type, Table 1 lists the quartile values of the calculated local dimension.

Table 1. Quartile values for local dimension at different token types for different LLMs.

Model	Token Type	Number of Tokens	Dimension		
			Q1	Q2	Q3
GPT-2	Non-numeric	48563	384	498	566
	Numeric	1694	10.7	15.8	22.1
Llemma-7B	Non-numeric	31972	9.44	10.5	11.2
	Numeric	44	4.92	6.84	9.07

If the local dimensionality of the token subspace were constant throughout, it can be concluded that the token subspace is a manifold. However, the key conclusion to be drawn from Table 1 is that the local dimension is not constant across the token subspace for either LLM. Thus, the authors posit that the *token subspace is not a manifold, instead it is a stratified manifold*. Deeper analyses for each model are presented in the following subsections.

Analysis of GPT-2 Results

Figure 5a plots volume vs radius for a few sample tokens in GPT-2. Notice that the plots have a few different slope values, i.e., there are 'knees' between straight segments that are prominently visible. This is evidence of the token subspace being a stratified manifold. The points where slope changes are marked as the stratification boundaries.

Figure 5b plots histograms of estimated local dimensions for GPT-2. For non-numeric tokens, the histogram appears to be bimodal with some tokens having low dimension close to zero while most others have dimension around 400-600. In contrast, the numeric tokens in GPT-2 have significantly lower dimension than non-numeric tokens. This is also evidenced by Figure 2 where it is shown that numeric tokens form their own cluster, i.e., they have fewer neighbors, which leads to lower dimension (although note as a word of caution that Figure 2 was plotted for the t-SNE reduced 2D embeddings while the dimensional analysis in this section is done in the original E-dimensional space).

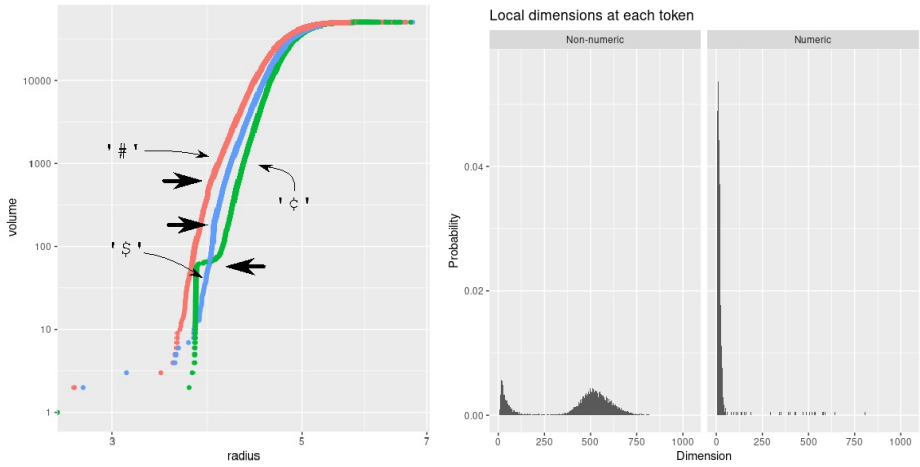

Figure 5: (a, Left): Volume vs radius for several tokens in GPT-2: # (red), $ (blue), ¢(green). Visible stratification boundaries are marked with arrows. **(b, right):** Histogram of estimated local dimensions for GPT-2 for non-numeric tokens and numeric tokens.

In addition to the overall lower dimensionality, it is found that the numeric tokens are confined to a constant dimension submanifold, thereby limiting the expressivity of any continuous dynamic map – such as the transformer layers (Vaswani, 2017) of the GPT-2 LLM – that act upon them. Note that the dynamic map induced by a transformer is continuous – a result which follows from the continuity of the activation functions in a deep neural network (Hendrycks, 2016). This implies that any set of numeric tokens that are near each other will tend to be taken to other tokens that are also near each other. Briefly, the topology of the token subspace suggests that GPT-2 will tend to treat all numeric tokens as interchangeable because they are co-located on a submanifold. Of course, mathematics requires that numeric tokens have distinct meanings, and they are definitely not interchangeable! Therefore, it can be immediately hypothesized, simply from the finding that the numeric token subspace is a (largely connected) manifold, that *GPT-2 will be a poor performer on mathematical queries*.

Analysis of Llemma-7B Results and Comparison to GPT-2

The Llemma-7B token subspace is also a stratified manifold, however, obtaining this conclusion is a bit more subtle than in GPT-2. Figure 6 shows the local dimension of different tokens in the t-SNE visualization. Slicing this figure along the value of 3 on the x-axis yields violin plots for the dimensionality of different tokens as the y-axis varies. None of these dimensions are constant, which is evidence of stratification.

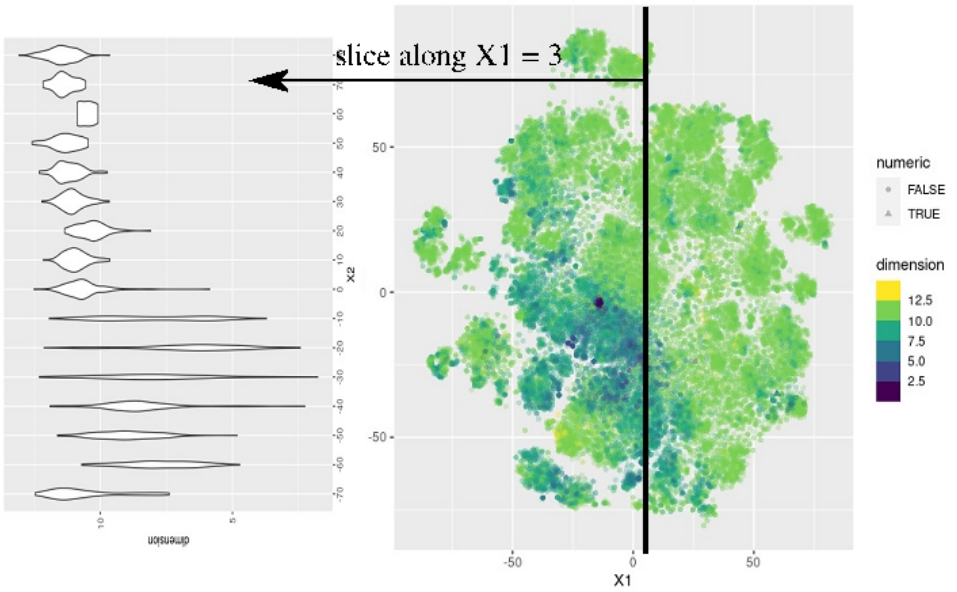

Figure 6: Estimated local dimension for Llemma-7B tokens, plotted using t-SNE reduced 2D token embeddings on the right. Dimension is indicated by the depth of color; darker points have lower local dimension. Dimension of individual tokens along X1=3 is shown on the left.

The bigger finding from the computed local dimensions for Llemma-7B is how low their values are. As seen in Table 1, most tokens have local dimension less than 10. This has implications for overfitting, which occurs whenever the dimension of the latent space (E) exceeds what is necessary to capture the structure of the token subspace. According to the Whitney embedding theorem (Lee, 2003), if the token subspace dimension is less than half of E, overfitting may occur. For Llemma-7B, the value of E is 4096, which is orders of magnitude bigger than twice the local dimension of any token. This is evidence of *overfitting in Llemma-7B*. Moreover, the excess dimension present in Llemma-7B's latent space is potentially wasteful. Such issues are not present in GPT-2, where most tokens have a local dimension that exceeds half the latent space dimension of 768.

Note that Llemma-7B has far fewer numeric tokens than GPT-2, and they are not near each other in the token subspace. Therefore, Llemma-7B can distinguish between numeric tokens and is not hampered like GPT-2 is when responding to mathematical queries. This plays a key role in Llemma-7B's superior mathematical performance. As validation, the authors of this paper conducted experiments where both GPT-2 and Llemma-7B were fed >160,000 mathematical queries of varying difficulty, and Llemma-7B correctly answered 5 times as many queries as GPT-2.

CONCLUSION

This paper presents a mathematical and topological analysis of large language models (LLMs) focused on their token embedding space, the geometry of

which constrains the model's inferential behavior. In particular, the token embedding spaces of different LLMs have varying local dimension, i.e., they are stratified manifolds. Exploring this space and visualizing it shows how, in some cases, clusters of semantically and syntactically related tokens such as numbers are co-located on submanifolds. In particular, the GPT-2 LLM has more than a thousand numeric tokens clustered together in a largely connected manifold that is distinct from the manifolds containing non-numeric tokens. Mathematical queries which demand the model fluently generate across numeric and non-numeric strata consequently yield poor performance. In contrast, the Llemma-7B LLM has less than 50 numeric tokens, but they are not clustered together, a geometry with corresponding superior mathematical performance. However, the local dimension of tokens in Llemma-7B is orders of magnitude lower than the dimension of the latent space, an indication of overfitting and limited generalizability.

The goal of this research is to explain the behavior, capabilities, and limitations of LLMs using a foundational, first principles approach by starting from their internal structure. This paper contains initial findings from ongoing research in this domain. Future work will involve continued experimentation on both foundational and fine-tuned models of different sizes, consideration of position embeddings, and extensions from tokens to token sequences and associated semantics.

ACKNOWLEDGMENTS AND DISCLAIMER

The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Distribution Statement 'A' (Approved for Public Release, Distribution Unlimited).

REFERENCES

Azerbayev, Z., Schoelkopf, H., Paster, K., Dos Santos, M. (2023) "Llemma: An Open Language Model for Mathematics", arXiv preprint arXiv:2310.06786.

Dey, S., Kanala, S. C., Chugg, K. M., Beerel, P. A. (2020), "Deep-n-Cheap: An Automated Search Framework for Low Complexity Deep Learning", proceedings of the12th Asian Conference on Machine Learning. pp. 273–288.

Gray, A. (1974) "The volume of a small geodesic ball of a Riemannian manifold", Michigan Mathematics Journal, Volume 20, No. 4. pp. 329–344.

Hendrycks, D., Gimpel, K. (2016) "Gaussian Error Linear Units (GELUs)", arXiv preprint arXiv:1606.08415.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., El Sayed, W. (2023) "Mistral 7B", arXiv preprint arXiv:2310.06825.

Lee, J. (2003) "Smooth Manifolds", Springer.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J. (2024) "Large Language Models: A Survey", arXiv preprint arXiv:2402.06196.

NVIDIA (2024) "Nemotron-4 340B Technical Report", arXiv preprint arXiv:2406.11704.

OpenAI (2023) "GPT-4 Technical Report", arXiv preprint arXiv:2303.08774.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019) "Language Models are Unsupervised Multitask Learners".

Sennrich, R., Haddow, B., Birch, A. (2016) "Neural Machine Translation of Rare Words with Subword Units", proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1715–1725.

van der Maaten, L., Hinton, G. (2008) "Visualizing Data using t-SNE", Journal of Machine Learning Research, Volume 9. pp. 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017) "Attention is All you Need", proceedings of Advances in Neural Information Processing Systems 30.

Yomden, Y., Comte, G. (2004) "Tame geometry with applications in smooth analysis", Springer.