



Analyzing Large Language Model Behavior via Embedding Analysis

*Sourya Dey*¹, Michael Robinson², Shauna Sweet,
Andrew Lauziere¹, Jonathan Daugherty¹, Cait Burgess¹

¹Galois, Inc., USA

²American University, USA

AHFE Hawaii 2024

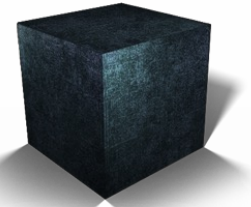
The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

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).

Highlights

Large language models (LLMs) and generative AI form a disruptive technology that is changing how we lead our lives.

But, we don't really understand how LLMs work internally!



We aim to develop a foundational understanding of LLMs from their token embeddings. Key findings are:

- Pre-trained LLMs learn to *differentiate between tokens* based on syntax and semantics at the token embedding level.
- Token embeddings lie in a *stratified manifold* subspace.
- Topological properties help explain *overfitting* and *math performance* of LLMs.

Token embeddings

Prompt

I have 3 responsibilities.

Tokenize

Tokens

_I
_have
_
3
_respons
ib
ilities
.

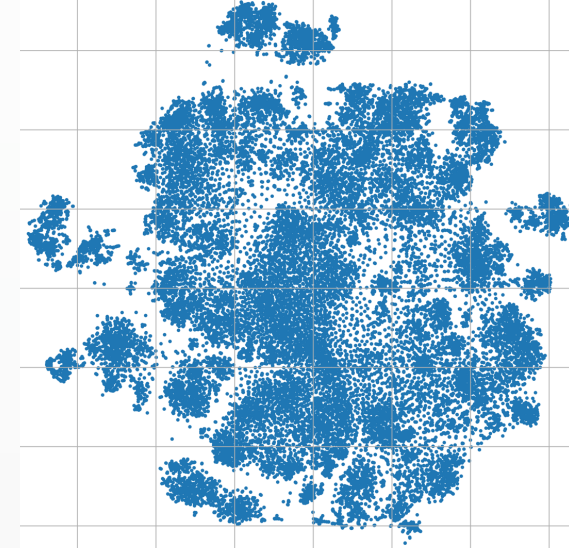
Token embeddings

<E-dim vector>
<E-dim vector>
<E-dim vector>
<E-dim vector>
<E-dim vector>
<E-dim vector>
<E-dim vector>
<E-dim vector>

Live in an E-dimensional **latent space**.

Only the part of this containing the token embeddings is linguistically meaningful – known as the **token subspace**.

Latent space and token embedding vectors of Llemma-7B shown in 2D.





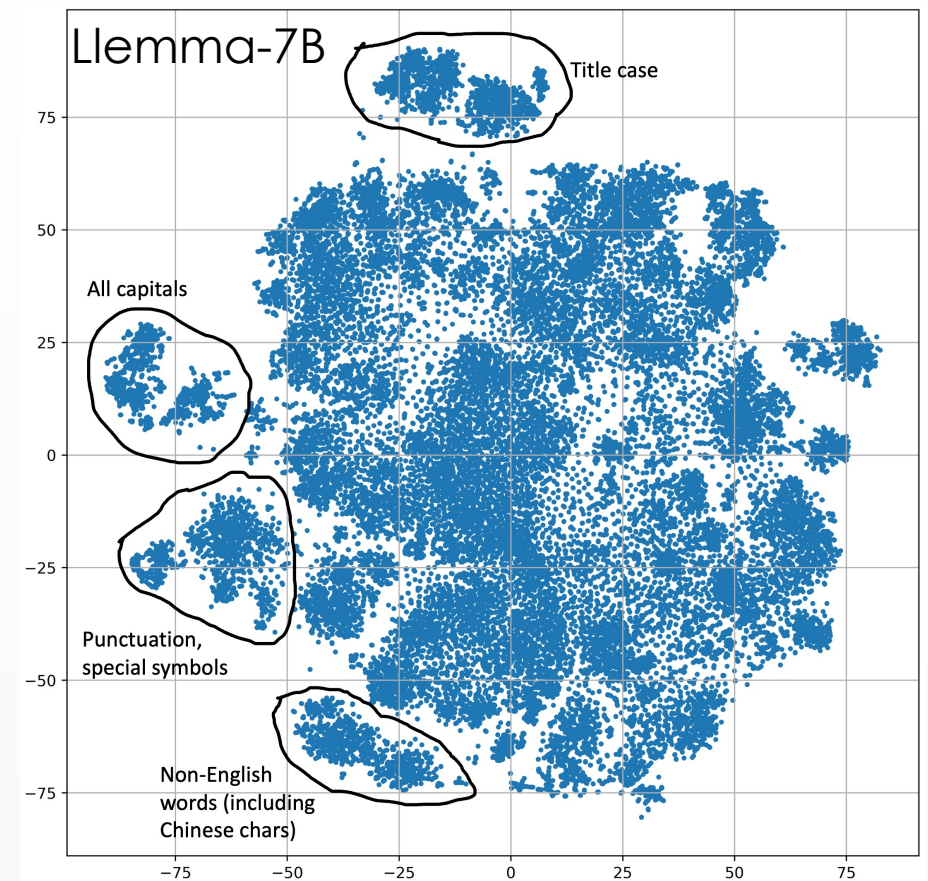
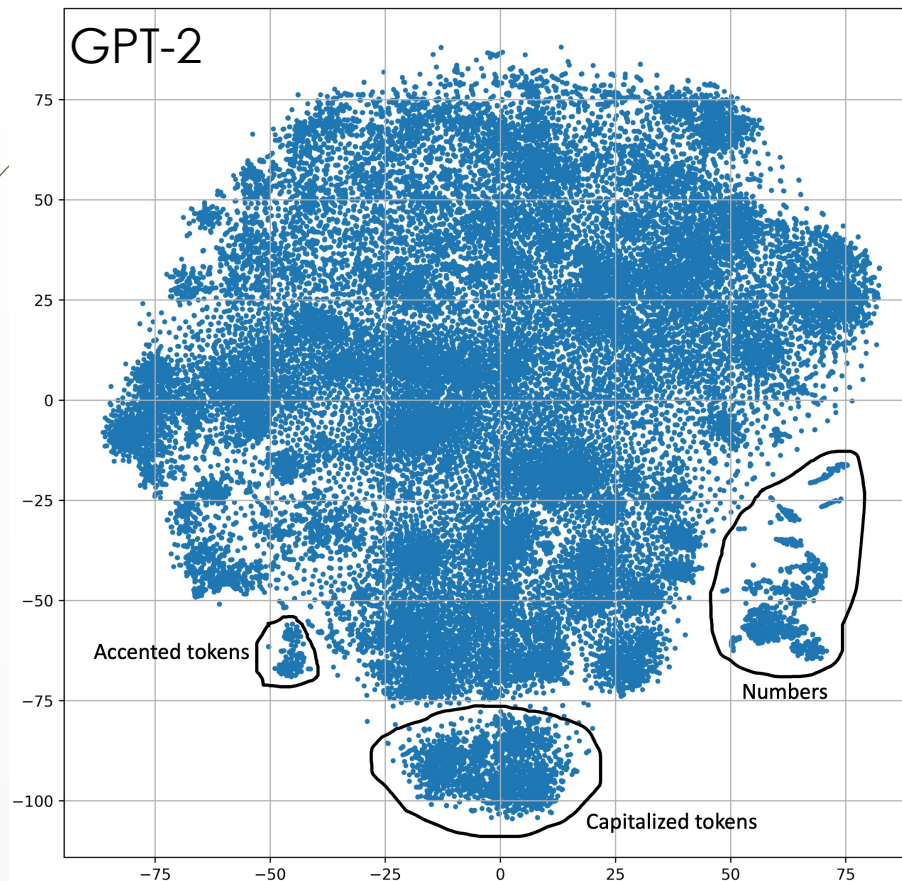
Models for experimentation

- GPT-2:
 - General purpose model
 - 125 million parameters
 - Vocabulary size = 50257 tokens
 - Latent space dimension: $E = 768$
- Llemma-7B:
 - Model specialized for math and code
 - 7 billion parameters
 - Vocabulary size = 32016 tokens
 - Latent space dimension: $E = 4096$

Our current results also include Mistral and Pythia models (omitted here for brevity).

Visualizing token embeddings

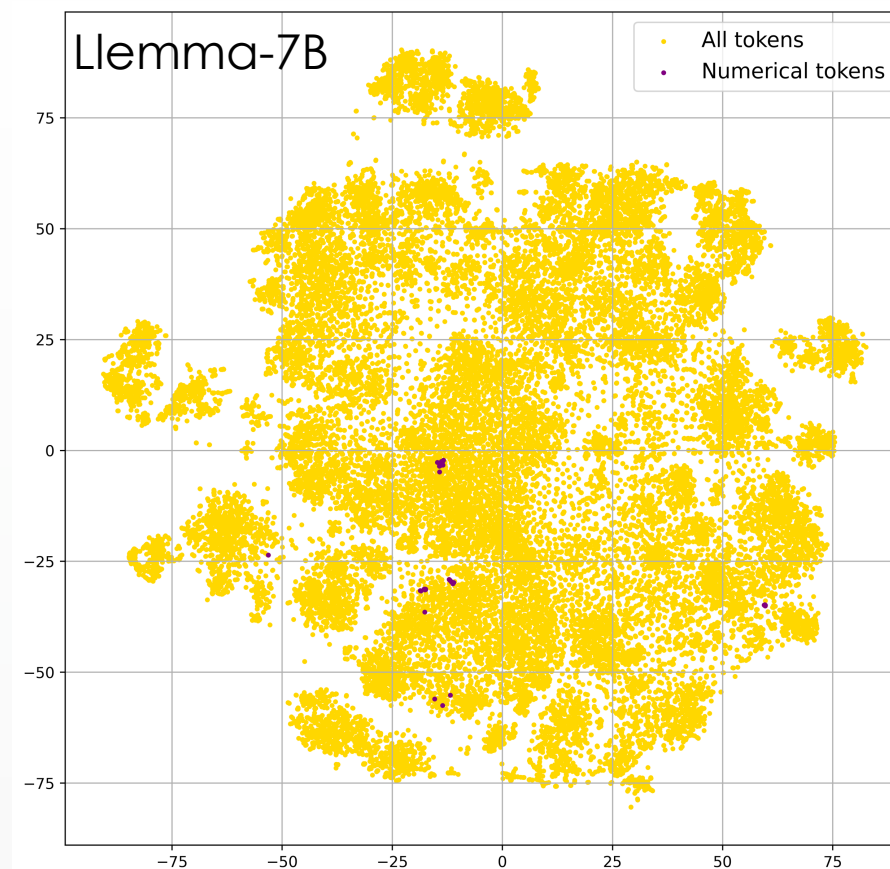
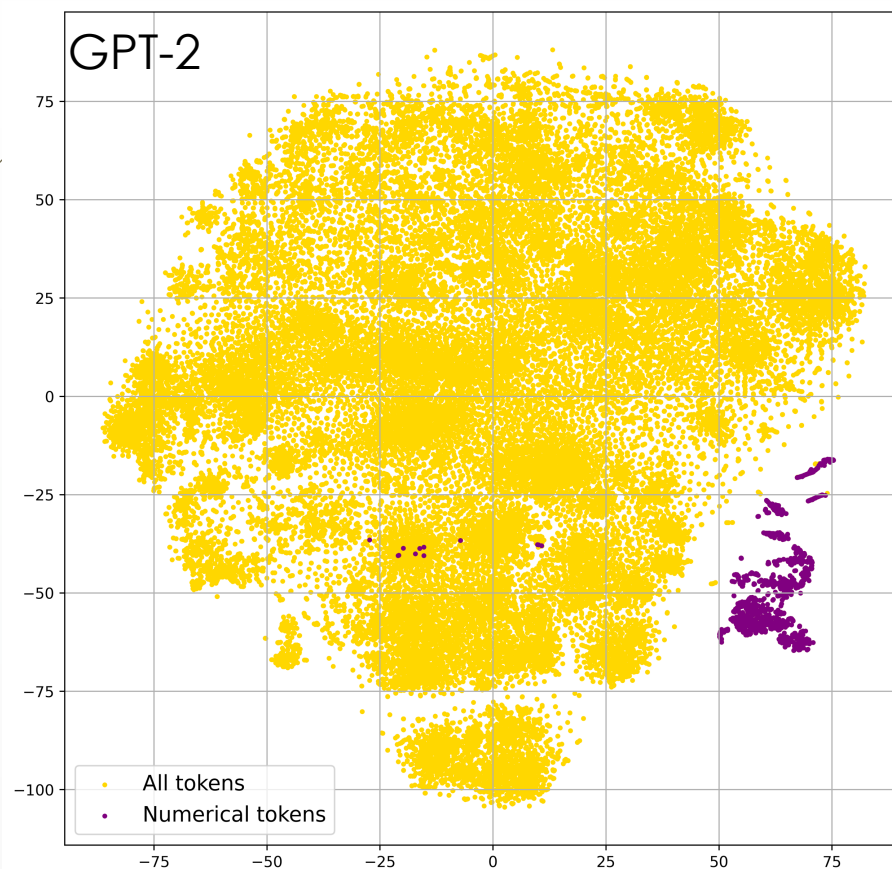
Key finding: Distinct **clusters** of syntactically and semantically related tokens are observed when visualizing pre-trained token embeddings of different models.



Embedding numerical tokens

Key finding: GPT-2 has a distinct cluster containing >1600 numerical tokens. On the other hand, Llemma-7B only has 44 numerical tokens that do not cluster together.

Yet, Llemma performs better at math ... we will soon see why!



Topological Analysis of the Token Subspace

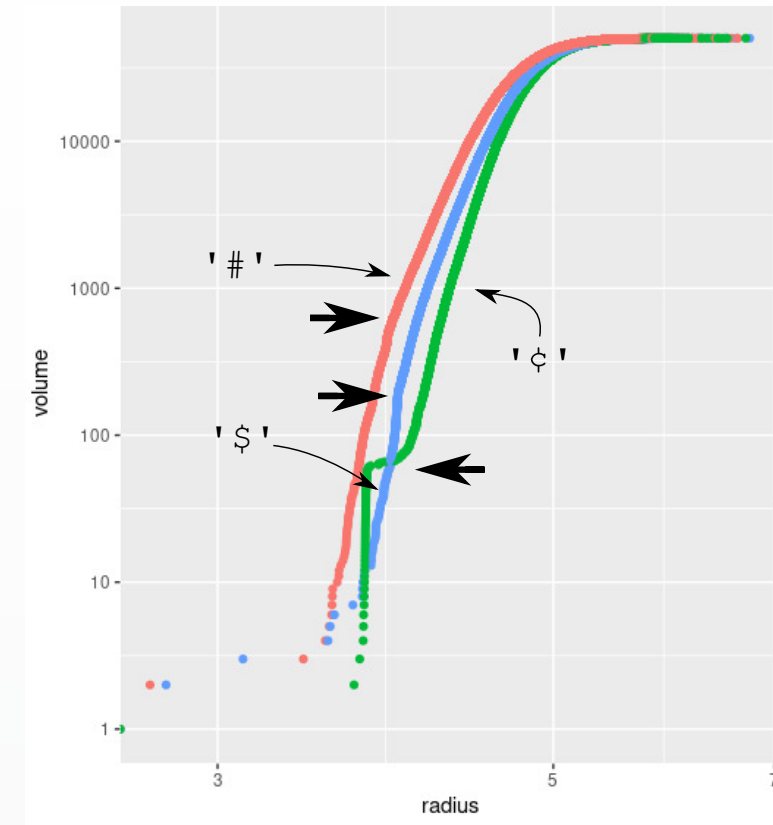
Dimension is a fundamental topological property.

We know the dimension of the latent space (E), but the quantity that can explain model behavior is the **dimension of the token subspace**, measured as the (local) dimension at different tokens.

Dimension at a token = slope of its volume vs radius plot.

- Radius = A distance metric between different tokens.
- Volume = How many tokens are within a given radius of a token.

Key finding: Slope of volume vs radius curve for various tokens is not constant, i.e. the **token subspace is a stratified manifold**, not a simple manifold.



Explaining math performance using 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

Key finding: While most tokens in GPT-2 have dimension in the hundreds, the numerical tokens lie in a submanifold of much lower dimension. Their clustering implies that **GPT-2 will treat all numerical tokens as interchangeable**. This makes GPT-2 a poor performer on mathematical queries.

Llemma-7B does not have this issue, making it superior at math than GPT-2.

Validation: We fed >160k math queries of varying difficulty to both models. Llemma-7B performed 5X better than GPT-2.

Explaining overfitting using 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

Key finding: Most tokens in Llemma-7B have dimension (n) around 10, but its latent space dimension (E) is 4096. By the Whitney embedding theorem¹, $E > 2n$ is evidence of overfitting. Thus, **Llemma-7B exhibits overfitting**, and the excess dimension present in its latent space is potentially wasteful.

GPT-2 does not have this issue.

¹ John M. Lee, "Introduction to Smooth Manifolds", Springer, 2003.

Conclusion and Future Work

Overall research goal: Explain the behavior, capabilities and limitations of LLMs and generative AI using a foundational, first-principles approach by starting from their internal structure.

Ongoing and future work:

- Experiments with more models (Mistral and Pythia families of models)
- Connecting topological structure to semantics such as parts of speech.
- Explore latent spaces at intermediate layers, i.e. attention heads.

More extensive follow-up paper:

<https://arxiv.org/abs/2410.08993>

