CSC2517 Proposal: Topological Analysis of Hidden Structures in BERT

Devan Srinivasan
University of Toronto
devan@cs.toronto.edu

1 Introduction

It is well understood that Large Language Models (LLMs) compete with (and sometimes exceed) humans in various tasks such as mathematics DeepSeek-AI et al. (2025); Chen et al. (2025b) [1], question answering (Kamalloo et al., 2023), and writing (Gómez-Rodríguez and Williams, 2023), all which require a sufficient understanding of human knowledge and language. While it seems LLMs have effectively modelled this, little is known about their internal mechanisms in latent space. There are many ways to analyze these latent representations, but one interesting approach comes from the vantage point of topology. Prior work has proposed that latent state representations in DNNs form manifolds (Fitz et al., 2024; Benfenati et al., 2025), on which we can utilize tools from topology to study.

Seldom work has been done to justify the existence of such a manifold, and these ideas have yet to be reconciled with emerging results from mechanistic interpretability describing this space with the linear representation hypothesis (LRH) (Park et al., 2024). LRH posits that learned features are stored as directions, pairwise nearly orthogonal, in high dimensional latent space, utilizing high sparsity in co-occurence to do so reliably. This phenomena of storing more vectors than the rank of the vector space is called superposition. Under this principle in the context of transformers (Vaswani et al., 2023), as each token is processed in parallel, the latent representation of each token is called it's residual stream, a sum of numerous features in superposition. From this viewpoint it can be shown mathematically that the transformer reads and writes information between token streams (Elhage et al., 2021).

Now given this hypothesis, we can create a new interpretation on the efficacy of prompting strategies (Zhou et al. (2023); Agarwal et al. (2024) are a few examples). Since LLMs are, by design, constrained to work with the features present in each token's stream, when different tokens are present in the prompt, it is possible that these tokens contain new features in their streams (held in superposition) that the transformer now has access to compute with, yielding a different (hopefully better) answer. Colloquially this phenomena has been described differently as "changing manifolds" where LLMs gain access to new manifolds previously unavailable. This suggests the structure of the residual manifold should change, and such a geometric change is a perfect problem to apply topological data analysis (TDA) to, which is what we intend to do.

Reconciling these three ideas - residual manifolds in transformers, LRH, and TDA is what this proposal aims to accomplish. Explicitly it strives to make the following inquisitions[2]:

1. Mathematically investigate residual "manifolds" and their structure
2. Formalize theoretical ideas behind "changing manifolds" under the linear representation hypothesis
3. Conduct experiments studying residual manifold topology under linguistic perturbation supporting the above theory

The details of the intended approach and experiments can be found in section 3.

2 Related Work

TDA and Machine Learning

While there has been plenty of work conducting topological data analysis in machine learning, we

[1] See https://epoch.ai/frontiermath for latest results

[2] Note we use "manifold" here where we have not yet justified such a mathematic object exists. We are following prior use of the term (Fitz et al., 2024)

present a subset most relevant to this work and it's related topics. One avenue of applying tools from topology to machine learning problems has demonstrated the relationship between the complexity of the problem and/or model, and it's topological features. Bianchini and Scarselli (2014) studied how the topological complexity of a neural network's input class regions reflects it's architecture and expressivity, relating theoretical results to the expressivity of deep versus wide neural networks. Similarly, Guss and Salakhutdinov (2018) analyzed the topological complexity of decision boundaries in the data and related it to the capacity of the network to learn the function, by empirically characterizing it's topological capacity. Rieck, Bastian Alexander et al. (2023) even developed a topology-based complexity measure for neural networks by applying persistent homology to the weights, and aggregating across the entire network.

TDA and Hidden Representations

Furthermore, there has been work even conducting topological data analysis on latent state representations, akin to what we seek to do. Świder (2024) does persistent homology with Vietoris-Rips complexes built from latent representation point clouds on vision datasets (with vision models). Valeriani et al. (2023) analyze the geometry of hidden representations of large transformer models (not necessarily LLMs) finding correlations between topological features and semantic content in the network. Balderas et al. (2025) use persistent homology on neurons in BERT to define a pruning criterion acheiving strong results. Gardinazzi et al. (2025) even use persistent homology across entire language models, finding results supportive of previous results in mechanistic interpretability.

Manifolds and LLMs

There is also a body of more work pertaining specifically to manifolds and LLMs. Mamou et al. (2020) conduct manifold analysis with word representation models (like BERT) finding separability in class manifolds. Benfenati et al. (2025) interpret transformer layer processing as deformations of the input manifold, and through Jacobian-based algorithms find equivalence classes within the input manifold. Fitz et al. (2024) uses various tools from TDA on point cloud residual manifolds in transformer language models, studying topological change through training. Zhang and Dong (2025) propose alignment between LLM internal representations (across layers) and hierarchial semantic categorization. Modell et al. (2025) theoretically justify the existince of feature manifolds in transformers, which are programmatically found by Tiblias et al. (2025). In fact Chen et al. (2025a) incorporate manifold alignment as a constraint in fine-tuning, finding models perform successfully whilst preserving geometric structure.

LRH

Finally, apart from manifolds and topology, there is the work related to mechanistic interpretability. Elhage et al. (2022) empirically demonstrate the phenomena of superposition, and Elhage et al. (2021) derive the theory behind transformers computing by moving information between token residual streams. Park et al. (2024) formalize these ideas as the "Linear Representation Hypothesis" (LRH).

3 Plan

Motivated by prior work we seek to investigate the theoretical mathematics behind residual manifolds, and "changing" as it relates to topology and the linear representation hypothesis. We also wish to empirically study some of this topological structure, seeing how different perturbations affect the topology and complexity of the residual geometries.

3.1 Theory

This work is exploratory in nature, but these are some keen mathematic ideas we seek to investigate. We do not expect to answer all of these, but the breadth of exploration should incite enough interesting work for the course of this project.

3.1.1 Background & Definitions

Given a language model L with latent dimension d and input tokens $T = \{t_1, t_2, \dots, t_n\}$ we will denote their latent representations after layer i as $\{z_1^{(i)}, z_2^{(i)}, \dots, z_n^{(i)}\} \subseteq \mathbb{R}^d$. For simplicity[3] we will focus on $i = 0$ and omit the $z^{(0)}$ notation. Under the linear representation hypothesis (LRH) we will denote the set of all features learned by the model as $F = \{f_1, f_2, \dots, f_m\} \subseteq \mathbb{R}^d$. We can re-write our vector sets as matrices $Z = [z_1, z_2, \dots, z_n]^\top \in \mathbb{R}^{n \times d}$ and $F = [f_1, f_2, \dots, f_m] \in \mathbb{R}^{m \times d}$. Then by using superposition, each latent vector z_i is a sparse linear combination of features in F, or formally, $Z = AF$ where $A \in \mathbb{R}^{n \times m}$ is a sparse matrix of

[3] For now we will mainly pertain to latent vectors before the first layer, after embedding. Given sufficient time, we intend to perform layer-wise experiments as well

coefficients.

When assessing the point cloud topology we will build simplicial complexes using Vietoris-Rips construction (Vietoris, 1927; Hausmann, 1995). Given a point cloud $Z \subseteq \mathbb{R}^d$ and a distance parameter $\epsilon > 0$, we can construct a simplicial complex $K_\epsilon(Z)$ whose simplices correspond to sets of points in Z that are pairwise within distance ϵ of each other, where simplices are added once their faces are in the complex. We then can compute the homology of $K_\epsilon(Z)$ to obtain Betti numbers β_k which count the number of k-dimensional holes in the complex. By varying ϵ we can build a filtration of complexes and compute persistent homology, capturing topological features across scales. This is how we intend to compute topological features of the residual manifolds.

3.1.2 Residual "Manifolds"

We wish to investigate the topological structure of Z as a point cloud. We begin with the following questions:

1. Under the LRH, how do we justify manifold structure in Z? That is, what assumptions are required to prove $\forall z \in Z, \exists \epsilon > 0$ such that $N_\epsilon(z)$ is homeomorphic to $\mathbb{R}^k$ for some $k \leq d$? $N_\epsilon(z)$ denotes the open ball of radius ϵ around z.

2. Can we relate the sparsity of A to the intrinsic dimension of Z? Note that we do not know A.

3.1.3 Jumping "Manifolds"

Under LRH we can formalize this notion of "changing manifolds" as follows. Given a pair of point cloud of latent representations Z_1, Z_2 based on token sets T_1, T_2 respectively, (where perhaps T_2 is a "better" prompt than T_1) a "jump" would correpond to more features being surfaced in Z_2 than Z_1. Formally, if we define the feature support of a point clouds as $F_1, F_2 \subseteq \mathbb{R}^d$ where $Z_1 \subseteq \text{span}(F_1)$ and $Z_2 \subseteq \text{span}(F_2)$, and $F_2 \neq F_1$, this should be reflected topologically.

We wish to explore topological measures we can use to detect, and quantify this change, or "jumps". We wish to formally define these and mathematically prove why they should measure this phenomena, before experimenting with them. Currently we consider the following:

1. Using simplicial and/or persistent homology, we can build Vietoris-Rips complexes on Z and measure Betti numbers β_k to capture global topological structure. Can we develop a theoretical suggestion to these point cloud topologies, and their changes?

2. Intuitively we would like to measure change in geometric structure. We would like to develop or use a measure of curvature on Z and relate it to these jumps

3. Prior work has undoubetdly demonstrated semantic correlation in how tokens are embedded (Mikolov et al. (2013) is one example), can we relate this to how semantically correlated point clouds may present topologically? Can we conjecture about their homology class and persistent homology?

3.2 Experiments

To test this theory empirically, we intend to collect a dataset of long-form text (atleast 100 tokens) of math & science concepts. We choose these over the more natural choice of stories or articles as they are less likely to contain named entities and proper nouns that may complicate tokenization. We will use a list of elementary and secondary math and science concepts from Wikipedia as our source (subsubsection A.1.1) and scrape them using the Wikipedia API[4]. This will give us samples of long-form text explaining one concept.

Then we will paraphrase these samples into different linguistic styles (academic, poetic, child-like, etc.) using the OpenAI API[5]. We will use GPT-4 to do this given it's efficacy in various writing tasks[6]. We will filter the paraphrases to ensure they are within ±5% of the original token count to ensure similar information content.

This will leave us with a dataset of concept explanations in different linguistic styles, representing the same semantic content. For each sample $S = \{w_1, w_2, \ldots, w_n\}$, we will pass the sample through BERT (Devlin et al., 2019) and collect the latent representations $\{z_1, z_2, \ldots, z_n\}$ after the embedding layer (before any transformer layers). This will give us a point cloud in $\mathbb{R}^d$ representing the residual manifold of this sample. We will then conduct topological analysis implementing the

[4] https://pypi.org/project/Wikipedia-API/

[5] https://platform.openai.com/

[6] See https://openai.com/research/gpt-4

measures described in subsection 3.1 to materialize those ideas.

3.3 Visualizations

Perhaps a minor part, but worth mentioning, we intend to use PCA on top $k \leq 3$ components of our residual manifolds, to visualize topology. We will also investigate (should time permit) using UMAP (McInnes et al., 2020) as well. This will be more dependent on the quality of our empirical results.

4 Logistics

4.1 Compute

Given the relative small size of BERT it's expected that all experiments can be conducted on the available compute of a Macbook Pro M1 series laptop. There is also GPUs available through the `sahitya` server which will more than certainly suffice for these experiments. Libraries like `gudhi` and `torch` will be used as they have an API for efficient vectorized C++ implementations of persistent and simlicial homology.

4.2 Timeline

From Nov 3 to Dec 23 we intend to roughly adhere to the following schedule (7 weeks):

1. Week 1 to 3: Work on the theory behind this subject, hopefully derive concrete theoretical results to test

2. Week 2 to 4: Collect data and build dataset, and conduct experiments

3. Week 4 to 6: Interpret experiments & work on mathematic explanations

4. Week 7: Finalize results and write final paper

Please note that the theoretical side of this work is exploratory as we desire for part of the proposed project to be the development of theory at the intersection of these subjects. It is not to be interpreted that the project is underspecified.

References

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. *Preprint*, arXiv:2404.11018.

Luis Balderas, Miguel Lastra, and José M. Benítez. 2025. A green ai methodology based on persistent homology for compressing bert. *Applied Sciences*, 15(1):390.

Alessandro Benfenati, Alfio Ferrara, Alessio Marta, Davide Riva, and Elisabetta Rocchetti. 2025. Unveiling transformer perception by exploring input manifolds. *Preprint*, arXiv:2410.06019.

Monica Bianchini and Franco Scarselli. 2014. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1553–1565.

Dexia Chen, Qianjie Zhu, Weibing Li, Yue Yu, Tong Zhang, and Ruixuan Wang. 2025a. Preserve and sculpt: Manifold-aligned fine-tuning of vision-language models for few-shot learning. *Preprint*, arXiv:2508.12877.

Luoxin Chen, Jinming Gu, Liankai Huang, Wenhao Huang, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Kaijing Ma, Cheng Ren, Jiawei Shen, Wenlei Shi, Tong Sun, He Sun, Jiahui Wang, Siran Wang, Zhihong Wang, Chenrui Wei, and 17 others. 2025b. Seed-prover: Deep and broad reasoning for automated theorem proving. *Preprint*, arXiv:2507.23726.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Stephen Fitz, Peter Romero, and Jiyan Jonas Schneider. 2024. Hidden holes: topological aspects of language models. *Preprint*, arXiv:2406.05798.

Yuri Gardinazzi, Karthik Viswanathan, Giada Panerai, Alessio Ansuini, Alberto Cazzaniga, and Matteo Biagetti. 2025. Persistent topological features in large language models. *Preprint*, arXiv:2410.11042.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

William H. Guss and Ruslan Salakhutdinov. 2018. On characterizing the capacity of neural networks using algebraic topology. *Preprint*, arXiv:1802.04443.

Jean-Claude Hausmann. 1995. On the vietoris-rips complexes and a cohomology theory for metric spaces. In F. Quinn, editor, *Prospects in Topology: Proceedings of a Conference in Honor of William Browder (Annals of Mathematics Studies, Vol. 138)*, pages 175–188. Princeton University Press, Princeton, NJ, USA.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. *Preprint*, arXiv:2305.06984.

Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2020. Emergence of separable manifolds in deep language representations. *Preprint*, arXiv:2006.01095.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. 2025. The origins of representation manifolds in large language models. *Preprint*, arXiv:2505.18235.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Rieck, Bastian Alexander, Togninalli, Matteo, Bock, Christian, Moor, Michael, Horn, Max, Gumbsch, Thomas, and Borgwardt, Karsten. 2023. Neural persistence: A complexity measure for deep neural networks using algebraic topology.

Federico Tiblias, Irina Bigoulaeva, Jingcheng Niu, Simone Balloccu, and Iryna Gurevych. 2025. Shape happens: Automatic feature manifold discovery in llms via supervised multi-dimensional scaling. *Preprint*, arXiv:2510.01025.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2023. The geometry of hidden representations of large transformer models. *Preprint*, arXiv:2302.00294.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Leopold Vietoris. 1927. Über den höheren zusammenhang kompakter räume und eine klasse von zusammenhangstreuen abbildungen. *Mathematische Annalen*, 97(1):454–472.

Yukun Zhang and Qi Dong. 2025. Multi-scale manifold alignment for interpreting large language models: A unified information-geometric framework. *Preprint*, arXiv:2505.20333.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Paweł Świder. 2024. Characterization of topological structures in different neural network architectures. *Preprint*, arXiv:2407.06286.

A Example Appendix

A.1 Dataset Creation

A.1.1 Source

We seek to paraphrase (reword) long form text samples where the underlying semantic content remains unchanged. To this end we chose elementary and secondary level math and science concept explanations, which can be easily sourced from Wikipedia. This is advantageous over something like stories or historical events as there are less named entities and proper nouns that may complicate tokenization. Additionally, we are interested in residual geometry not necessarily the actual distribution over the type of text samples, so this should suffice.

Using the Wikipedia API [7] we can scrape lots of text from paragraph explanations concepts mathematics (and possibly other disciplines). As of now we can use the following resources:

1. `https://en.wikipedia.org/wiki/List_of_physics_concepts_in_primary_and_secondary_education_curricula`

[7] https://pypi.org/project/Wikipedia-API/

2. https://en.wikipedia.org/wiki/List_of_calculus_topics

3. https://en.wikipedia.org/wiki/Outline_of_geometry

4. https://en.wikipedia.org/wiki/Outline_of_arithmetic

We will use the OpenAI API[8] to facilitate paraphrasing, with more details below.

A.1.2 Prompts

Note that "role" below is one of "Academic", "Poet", and "Child".

Prompt:

```
You are an expert {role} writer. Your task is to
reword the following text while preserving the
underlying concept explained. Ensure that the
reworded text explains the exact same thing,
with equal level of detail and no new information,
but is expressed in a {role} style. Ensure that a
similar amount of words are used (don't summarize or
elaborate too much).
```

We will filter the paraphrases to retain samples within $\pm 5\%$ of the original token count.

[8] https://platform.openai.com/docs/overview