Using embeddings' geometric similarity for in-context learning

Om Mehta
omehta@umass.edu

Tanush Savadi
tsavadi@umass.edu

Abstract

Recent work argues that large language models (LLMs) share geometric regularities in their embedding spaces, including similar global orientations across models from the same family and locally low-dimensional structure correlated to semantic coherence scores. In this project, we propose to test whether prompt effectiveness is partly explained by the geometry of token embeddings associated with a prompt. We will evaluate the correlation between geometrical prompt embeddings and performance on prompt-response datasets such as TruthfulQA and gsm8k. Results could offer better insight into prompt design/engineering.

1 Introduction

The Platonic Representation Hypothesis [HCWI24] argues that AI systems are becoming increasingly homogeneous in both their architectures and their capabilities. In particular, it conjectures that there is an endpoint to this convergence of representations. Recent work by Lee et al. [LWVW25] has demonstrated that token embeddings across large language models share numerous "global" and "local" geometric similarities: including high within-family cross-model correlation of pairwise token orientations, shared locally-linear neighborhoods and low "intrinsic dimensions" for tokens that cluster semantically.

Prompt engineering has emerged as an indispensable technique for extending the capabilities of large language models [SSS+24]. Subtle choices in prompt design such as wording, tone and structure can dramatically influence model outputs [Dje25]. A natural question to ask is, therefore, whether there is a link between the geometric nature of prompt embeddings and predictability/factuality of model generations (i.e., quality of output!).

2 Related work

Our project sits at the intersection of *representation geometry* and *prompt engineering*. We briefly cover both and include details on the gap we aim to fill.

Geometry of LM representations. A number of papers argue that modern models share common structure in their internal spaces. The Platonic Representation Hypothesis claims model representations are converging across architectures, which helps explain cross-model similarities [HCWI24]. In that spirit, [LWVW25] show strong *global* alignment (correlated pairwise token orientations within model families) and *local* similarity (shared LLE neighborhoods) in token (un)embeddings, plus evidence that some tokens live in lower-dimensional, more coherent regions. In short: not all parts of embedding space are equal; some neighborhoods look tidier and more semantically organized than others.

Prompt engineering through a geometric lens. Recent work connects prompting choices to changes in internal geometry. [KCCC25] compare instructions, demonstrations, and soft prompts, and find that these strategies trigger different mechanisms of task adaptation in the hidden space. [TS25] analyze prompt-based text embeddings and report task-dependent geometric signatures (e.g., lower intrinsic dimensionality and reduced isotropy for classification vs. higher for retrieval/STS). This suggests a simple testable idea for us: prompts that rely on tokens in "well-organized" regions (stable neighbors, low local ID) might be more reliable.

Soft prompts as geometric objects and transfer. Prompt tuning treats prompts as continuous vectors. [VLC+22] (SPoT) show that soft-prompt

vectors can transfer across tasks, effectively behaving like task embeddings. While SPoT operates on learned vectors rather than plain text, it supports the broader view that geometric similarity between prompt representations and a task can predict how well the model will adapt.

Practice and responsible use. Surveys like [SSS+24] summarize practical prompt techniques (instruction formatting, few-shot examples, verifier prompts, schema constraints) and their trade-offs in accuracy, cost, and robustness. Work on responsible prompt engineering [Dje25] reminds us to evaluate not only accuracy but also faithfulness and safety—metrics we will report alongside performance.

Our gap. Putting these threads together: prior work (i) documents convergent geometry across LMs [HCWI24, LWVW25] and (ii) shows that prompt formats change internal representations and outcomes [KCCC25, TS25, SSS+24]. What is missing, to our knowledge, is a direct, controlled test of whether *prompt-token geometry* (e.g., neighborhood stability via LLE and local intrinsic dimension) *predicts* in-context performance and faithfulness across models and tasks. Our study targets this gap with simple correlations and small interventions (geometry-aware synonym swaps), framed as a low-compute, empirical evaluation.

3 Your approach

A key question we wish to answer is whether the embedding-level metrics described in [LWVW25] such as local neighborhood stability, intrinsic dimensionality, and semantic coherence scores are useful predictors of the output quality of an LLM generation.

- We obtain the embedding of every token using a pretrained model from the Llama3 family of models. We will compute the following metrics, as described in [LWVW25]:
  - Local Linear Embedding (LLE) error is a metric that quantifies how well each token's neighborhood is linearly reconstructable.
  - Local Intrinsic Dimension (LID) represents the number of top (defined as explaining 95% or equivalent percentage of the variance) components identified using PCA
  - Semantic Coherence Score (SCS): average shortest distances from token x to nearest neighbors in ConceptNet [SCH17]. These provide external "cleanliness" measures.
- We will query the model on all benchmark datasets and record the performance on the given task using metrics as accuracy, perplexity, etc
- We will compute Pearson/Spearman correlation metrics between geometry metrics (LLE, LID, SCS) and outcome quality. These metrics will be produced, controlling for token length and frequency.
- Finding low-LID prompts to yield higher performance will support our hypothesis.
- Within prompt templates, we conduct word-swap experiments
  - we will identify words with high ID, and replace them with semantically closer words that will reduce the ID. following, we will re-evaluate performance on the task and measure the average difference in scores
- We repeat all analyses on Gemma2 and compare correlation patterns. It will be interesting to note any inherent correlations shared across model architectures.

Baselines

- text-level baseline: predict performance solely from surface statistics (prompt length, token frequency, sentiment).
- random synonym swaps without checking geometry

3.1 Schedule

We plan to work on all subtasks together.

1. Week 1: Reproduce LLE & ID computations on Llama 3 (sanity plots vs. Lee et al.). *Deliverable:* verification notebook.
2. Weeks 2–3: Build evaluation results from datasets (TruthfulQA, GSM8K, toxicity). *Deliverable:* score tables per prompt template.

3. Weeks 4–5: Compute geometry metrics per prompt; run correlation analysis; start synonym-rewrite search. *Deliverable:* correlation figures.

4. Week 6: Cross-family replication on Gemma 3; *Deliverable:* cross-model plots.

5. Week 7: Error analysis

6. Week 8: Final report

4 Data

We will evaluate on several publicly available benchmarks.

TruthfulQA [LHE22] tests factual accuracy and resistance to false statements. **GSM8K** [CKB+21] evaluates multi-step arithmetic reasoning. **XSum** [NCL18] provides summaries for news articles, useful for assessing coherence and faithfulness. **RealToxicityPrompts** [GGS+20] measures how prompt phrasing affects toxicity and tone.

All datasets are open-source and will be accessed via Hugging Face's `datasets` library. No manual annotation is needed. All datasets have fewer than 10k entries, barring xsum which has 200k prompt-response pairs. We will evaluate if the latter is computationally feasible, else scale down by random sampling.

5 Tools

Embeddings will be generated by yet-undecided members of the **Llama 3** and **Gemma 2** families of models, both of which are also available publicly. Performance metrics will be correlated with geometric features on our datasets. We will use standard ML/NLP libraries in Python such as **Transformers**, **scikit-learn**, etc. We do not anticipate training deep learning models, since our study is empirically focused. The models will be loaded on GPUs hosted on Google Colab, a free service offereed by Google that allows limited access to `https://colab.research.google.com`

6 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - Yes. We used GPT-5 served via ChatGPT.

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - Section 1 (Introduction): "Draft a 2–3 paragraph introduction that motivates our idea in plain language. Connect [HCWI24] (representation convergence) and [LWVW25] (global/local geometry) to our question about whether prompt-token geometry predicts performance. Keep it proposal-style with no results claims."
  - Section 1 (Introduction): "Rewrite this introduction for clarity and shorter sentences (student tone). Keep the citations to [HCWI24] and [LWVW25], and end with 1–2 lines that state our hypothesis and why it is testable this semester."
  - Section 1 (Introduction): "Turn these bullet notes into a LaTeX intro section. Define the problem in one sentence, briefly summarize prior observations about LLE/ID, and add a final 'our plan in one line' sentence."
  - Section 1 (Introduction): "Give me 3 candidate research questions and a single-sentence contribution statement that we can place at the end of the introduction."
  - Section 2 (Related Work): " Please tell me about this research paper by summarizing it and linking it to our project proposal. Is this relevant to what we are proposing? "
  - Section 2 (Related Work): " here are links to research papers relating to our project proposal, please use this and the uploaded research papers to write a .bib latex file "
  - Section 3 (Your approach): "Turn our outline into a clean 'Your Approach' section with A1 (correlation), A2 (geometry-aware synonym swaps), and A3 (cross-model checks). Keep it student-like, avoid claiming any results, and make the flow easy to follow."

- Section 3 (Your approach): "Explain LLE neighborhood stability, local intrinsic dimension (ID), and semantic coherence score (SCS) in plain language (3–4 sentences each). Include how we compute them from token embeddings and why they matter for prompts."
- Section 3 (Your approach): "Write a short baseline subsection with B0–B3 (length-matched control, frequency-only swaps, random swaps, family-transfer). One or two sentences per baseline."
- Section 4 (Data): "Write a concise Data section that names TruthfulQA, GSM8K, XSum (small slice), and RealToxicityPrompts. Explain why each fits our geometry-prompt study, how we will access them via the Hugging Face `datasets` library, and that we will subsample XSum if needed."

- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - We used an LLM mainly to bounce ideas and draft rough text. Most outputs needed editing: we rewrote for clarity/length and verified claims/citations ourselves. It was helpful for structure and phrasing, but we relied on our own judgment for the final content.

References

[CKB+21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[Dje25] Christian Djeffal. Reflexive prompt engineering: A framework for responsible prompt engineering and ai interaction design. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1757–1768, 2025.

[GGS+20] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models, 2020.

[HCWI24] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

[KCCC25] Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1855–1888, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[LHE22] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[LWVW25] Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. In *Second Conference on Language Modeling*, 2025.

[NCL18] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.

[SCH17] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2017.

[SSS+24] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[TS25] Hayato Tsukagoshi and Ryohei Sasano. Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25915–25930, Vienna, Austria, July 2025. Association for Computational Linguistics.

[VLC+22] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland, May 2022. Association for Computational Linguistics.