
Manifold Recovery as a Benchmark for Text Embedding Models¹

Lennart Finke
University of Göttingen & ETH Zürich

With
Bo Li & Apart Research

Abstract

Inspired by recent developments in the interpretability of deep learning models and, on the other hand, by dimensionality reduction, we derive a framework to quantify the interpretability of text embedding models. Our empirical results show surprising phenomena on state-of-the-art embedding models and can be used to compare them, through the example of recovering the world map from place names. We hope that this can provide a benchmark for the interpretability of generative language models, through their internal embeddings. A look at the meta-benchmark MTEB [1] suggest that our approach is original.

Keywords: *Model Evaluations, Text Embeddings, Manifold Learning*

1. Introduction

Transformer models are capable, which motivates us to understand their inner workings and hopefully align them. During the hackathon, we tried to answer the question: How can we benchmark how well a transformer internally represents concepts us humans intuitively grasp?

To answer this, we look to recent developments in interpretability from an evaluations point of view. In particular, we note that the residual stream of language models can yield mechanistic understanding and are therefore inspired to benchmark embedding models. This is because looking at the residual stream, that is, stopping

¹ Research conducted at the AI Security Evaluation Hackathon, 2024

a forward pass at an intermediate step in a transformer, yields a broad class of (albeit sparse) text embedding models which are likely valuable to have benchmarks for.

Our challenge then becomes to find a non-mechanistic, model-agnostic way to quantify whether an embedding model captures human-intelligible concepts.

2. Methods

The above problem can be solved by calculating how well the text embedding model can recover manifolds from discrete data. What we mean by this is, in the following, illustrated by an interesting example, to fit within the constraints of one weekend. Consider the strings “Palo Alto,” “San Francisco,” “Berkeley,” “Chicago” and “Amsterdam”. If one were to ask us which ones belong together, which ones are close or far, which is the odd one out etc. we would have a clear answer to all those questions. This is of course because we recognize the strings to be places; we keep them cognitively embedded on the manifold that is the surface of the earth. While it would be reductive to frame all of human understanding this way, many kinds of word clouds can be organized geometrically in some way, and each of those provides a test manifold which will be useful in our method. Our benchmark wants to check whether the embedding model does the same by organizing the strings into a neat structure in its output space.

To briefly state this mathematically, we consider an embedding model $\Gamma: \Omega \rightarrow R^n$, and a function Λ that maps a Manifold M to the input space of the model Ω . The benchmark then asks: How nice does the composition $\Lambda \circ \Gamma$ look? Does it approximate a smooth function? Can it ever be close to linear? How many output dimensions do we need to completely recover the input? Generally, how well can the text embedding recover the manifold across the discrete barrier?

Returning to our example and demonstrating our idea, we embed ~ 10000 addresses from across the world, as obtained from the GeoNames [2] dataset, with OpenAI’s text-embedding-3-small.

After confirming that our idea is sensible, we think about ways to give quantitative answers to the above questions. For smoothness, which we will focus on in this report for brevity, a natural way might be to endow both the original manifold M and the embedding output space R^n with metrics d_M, d_{R^n} and calculate the correlation between $d_{R^n}(\Lambda \circ \Gamma(y), \Lambda \circ \Gamma(x))$ and $d_M(y, x)$.

3. Results

The simplest hypothesis one can make about the embedding $\Lambda \circ \Gamma$ is that it is in some sense continuous. To get a visual answer to this, we plot the address coordinates on a world map colored by a one-component Principal Component Analysis obtain Figure 1 (see Appendix). Indeed we are positively surprised. The model gives similar embeddings to close by places and there seem to be few hard edges, even across country and state borders. To take the burden off our simple dimensionality reduction technique (PCA), we take Italy as a case study with equally spaced addresses and obtain a curious result in Figure 2. The first component runs from North to South

and second runs somewhat orthogonally from East to West. On this text subspace, a model as complicated text-embedding-3 is more or less linear. It had learned the manifold.

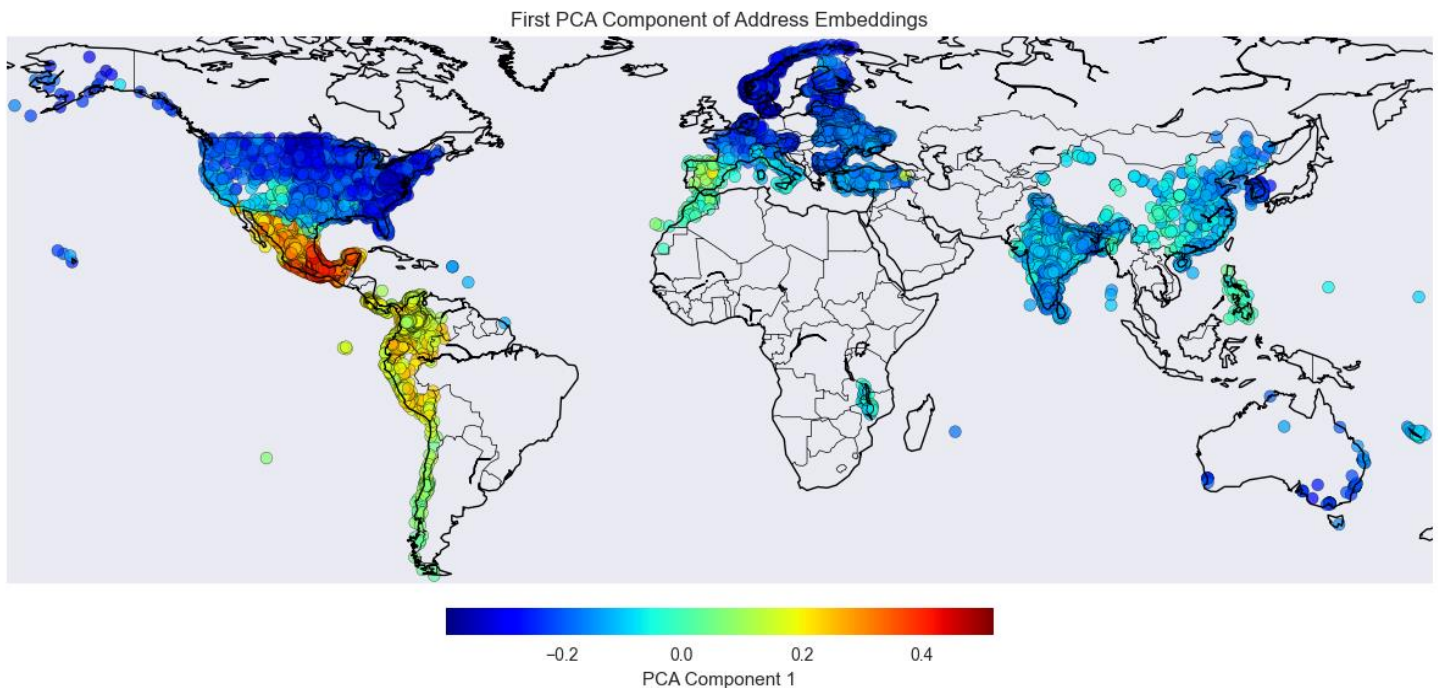


Figure 1 – Text-embedding-3 forward passes of addresses across the globe in countries where we had data. The reason that Mexico is so significant in the PCA seems to be that most of the addresses in the dataset are there. We observe that the color gradient is seemingly continuous in space.

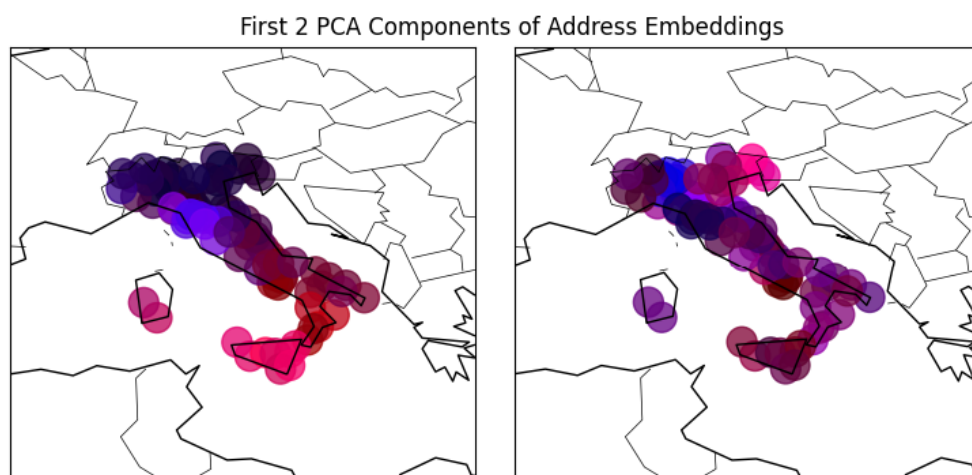


Figure 2 – As in Figure 1, but using only data from Italy and with 2 PCA components. The red and blue RGB channels are the first and second component respectively. On the left, text-embedding-3. On the right, Jina AI V2 embeddings.

Next, we demonstrate the quantify these results to obtain a benchmark, comparing text-embedding-3 to Jina AI V2 embeddings. To get multiple functions Λ out of the map example, we separate out the countries and respectively compute the correlation between the cosine distance on longitude / latitude vs. on embeddings. These are shown in Figure 3. Text-embedding-3 has a higher correlation on 27/37 countries.

4. Discussion and Conclusion

We demonstrated that manifold recovery might be a sensible evaluation technique for embedding models, which we hope can be used for quantification of how interpretable intermediate layers of language models are.

A range of natural extension questions present themselves. Firstly, what other choices of manifold-to-text embeddings might be useful to examine models with? A couple examples that come to mind are mapping 1) $\mathbb{R}^3$ to color names according to RGB + xkcd color names, 2) places in the night sky to star names / formations, 3) numbers as floats to numbers as strings etc. The reader is invited to suggest some others.

The next pressing question is if the correlation between distances is really the canonical metric to use here. For example, we could look at the Lipschitz constant of the map $\Lambda \circ \Gamma$. We guess that principled ways to do this might involve comparing graphs induced by the point neighborhoods (like the Čech complexes used in the UMAP algorithm).

We might also be able to leverage more sophisticated dimensionality reduction techniques for visualization. Here, we use PCA because we find it remarkable that it gives such clean results already.

Lastly, we thank the Apart Research team, the judges and Bo Li for the opportunity to participate and learn so much about alignment.

5. References

- [1] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, & Nils Reimers. (2023). MTEB: Massive Text Embedding
- [2] Marc Wick. (2005). GeoNames. Available: <https://www.geonames.org>.

6. Appendix

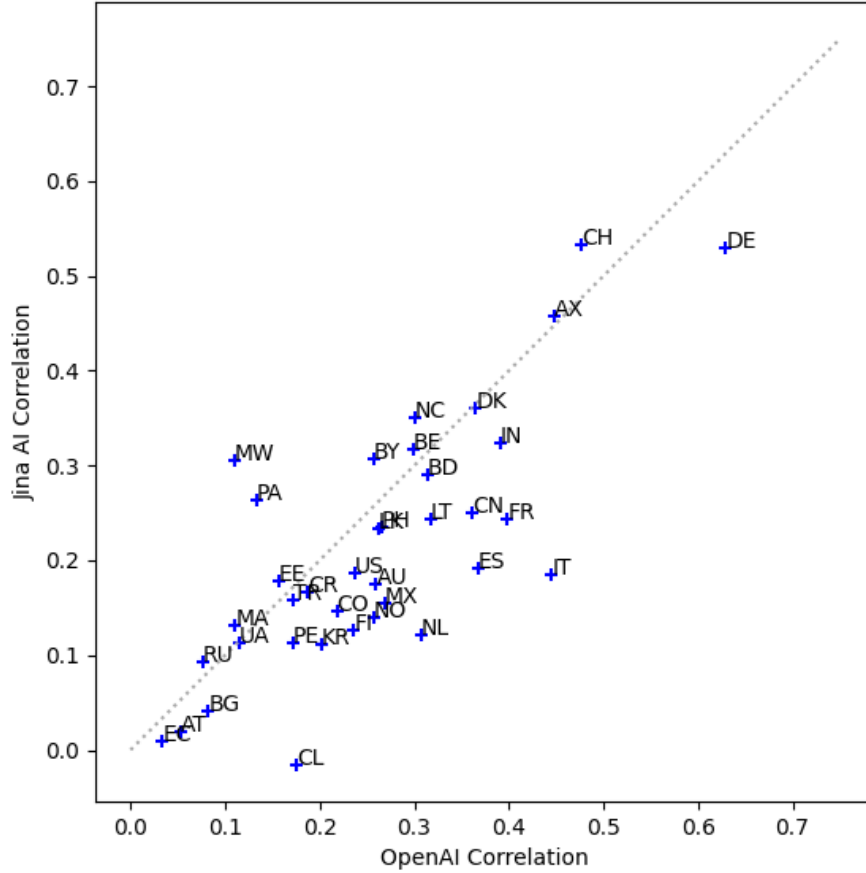


Figure 3 – Correlations between the cosine distances of two coordinates on the surface of earth vs. the cosine distances of two embeddings of addresses corresponding to that coordinate, by country. The countries are indicated by their ISO code, see (https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2). The dotted line is simply the bisector of the axes, to judge which model has better correlations.