LLM Induced Embedding Spaces

Amanda Myntti, TurkuNLP, University of Turku

The aim of this project is to investigate the embedding spaces induced by LLMs. Below are different ways to approach this problem.

INVESTIGATE AND MODIFY SPACE

Investigating an embedding space present some difficulties, as the spaces consist of multidimensional vectors which encode information.

Dar et al. (2023) treat each layer output as the final output of the model and decode the embeddings using an inverse of the first layer embedder.

Tennenholtz et al. (2023) investigate a target embedding space by creating an adapter to input target embeddings to separate LLM and querying its properties using natural language.

These findings show that **mapping from an embedding space to a vocabulary space** is achievable and **interpolating** between two points in the embedding space yields semantically meaningful outputs.

Liu et a. (2023) create task specific vectors to steer the model layer outputs, effectively turning few-shot prompting to one query by modifying the model's latent space.

Li et al. (2023) linearly interpolate each attention head in a model to steer the model output to a more desired direction.

These results pave the way to **modifying the embedding** space in a purposeful way.

Subramani et al. 2022

Turner et al. 2024

PROPERTIES

Statistical analysis of word representations has shown that embedding spaces of popular LMs are highly anisotropic.

Thus, cosine similarity as a metric has been questioned.

Focus has mostly been on how to combat this issue.

Timkey et al. 2021

Li et al. 2020

Steck et al. 2024

Mickus et al. 2020

MODEL TRAINING

There's been an uptick in research about how training data affects model downstream performance.

Generally, this research is focused on statistical text metrics, with some exceptions, and the goal is to maximize benchmark performance.

This opens up questions about the effect of training data on the embedding space and about the qualities that correlate with model performance.

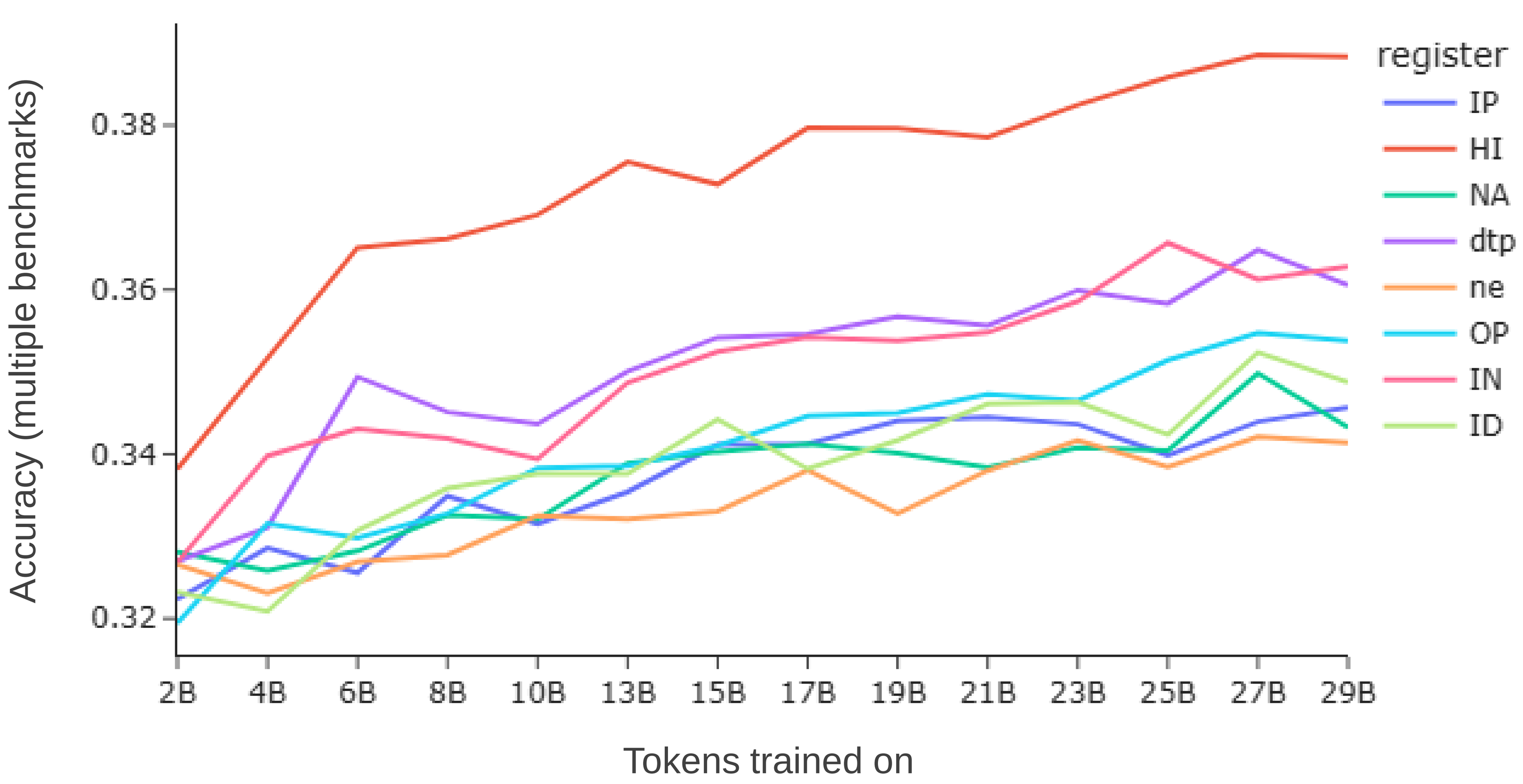
Ostendorff et al. 2024

Penedo et al. 2023

Longpre et al. 2024

Current work

Current work is focused on the effect of training data on LLMs. Following Penedo et al. (2024), I train multiple LLMs with different datasets, filtered with linguistic criteria: by register (text genre). Figure below presents preliminary findings.



The models trained and evaluated in this first publication can be used later to analyse the qualities of embedding spaces that are affected by linguistic features of training data.

Research Questions

Research Questions for this project are

RQ1 Can you build a system to get the sense of semantic meaning for an arbitrary point in an embedding space?

RQ2 What happens in the neighbourhood of an arbitrary point and what effects steering has?

RQ3 How can we affect the space using the training data of the model and are there qualities of the space that correlate with model performance?

UNIVERSITY OF TURKU

TURKUNLP .ORG

AI-DOC
Finnish Doctoral Program Network in Artificial Intelligence

Dar et al. (2023) Analyzing Transformers in Embedding Space. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.
Li et al. (2020) On the Sentence Embeddings from Pre-trained Language Models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).
Li et al. (2023) Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. Advances in Neural Information Processing Systems.
Liu et al. (2023) In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering. Preprint.
Longpre et al. (2024) A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
Mickus et al. (2020) What do you mean, BERT? Proceedings of the Society for Computation in Linguistics 2020.
Ostendorff et al. (2024) LLM-Datasets: An Open Framework for Pretraining Datasets of Large Language Models. Preprint.
Penedo et al. (2023) The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data only. Proceedings of the 37th International Conference on Neural Information Processing Systems.
Penedo et al. (2024) The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, forthcoming.
Steck et al. (2024) Is Cosine-Similarity of Embeddings Really About Similarity? Companion Proceedings of the ACM Web Conference 2024.
Subramani et al. (2022) Extracting Latent Steering Vectors from Pretrained Language Models. Findings of the Association for Computational Linguistics: ACL 2022.
Tennenholtz et al. (2023) Demystifying Embedding Spaces using Large Language Models. Proceedings of The Twelfth International Conference on Learning Representations: IRLC 2024, forthcoming.
Timkey et al. (2021) All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).
Turner et al. (2024) Steering Language Models with Activation Engineering. Preprint.