MS Capstone Proposal: Research Thesis

Thesis Title: TESTING ROBUSTNESS AND STABILITY OF EMBEDDINGS AND CLUSTERIZATION

By Ashish Karikere

Thesis advisor:  Prof. Dr. Alex Wolpert

Committee members: Prof. Dr. Evgeny Dantsin

Prof. Dr. Jerry Schnepp

Abstract

We aim to propose metrics and methods to measure the robustness of embeddings created by Large Language Models. This study builds upon the previously establish methods of measurement, metrics and frameworks to measure robustness which are primarily qualitative. Using the above established metrics, we test the different types of embedding representations using foundational models of GPT-2 and BERT. This lays the necessary groundwork for testing clusterization on a dataset containing texts from a number of different topics.

GitHub: [CST-485-Thesis-Project-Research](https://github.com/AshishKarikere/CST-485-Thesis-Project-Research)

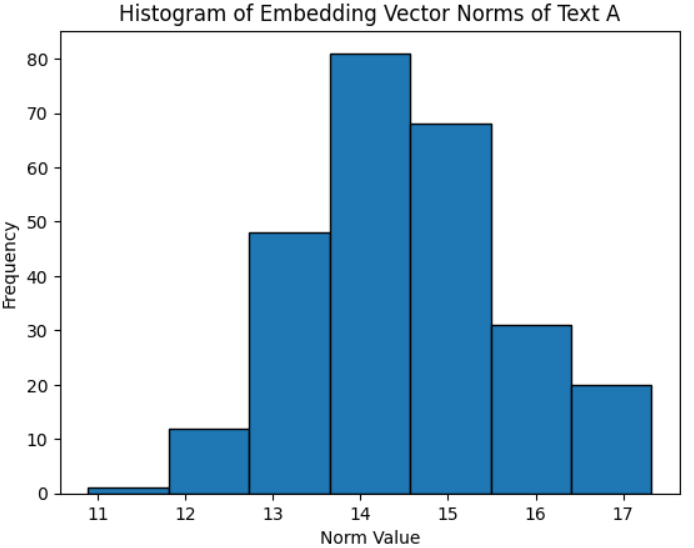
INITIAL EXPERIMENTS

The preview of the following Thesis involved a set of experiments to learn about the internal structure of the embeddings for foundational models of BERT and GPT2. This involved experiments to objectively learn about the distributions of the norms (fig 1.1), pairwise angles (fig 1.2) of the embedding vectors and the distribution of distances in the In-Between distance matrixes from the embeddings vectors.

To get a general idea of how the embeddings matrices would be spread in N-dimensional space experiments were performed using the Multi-Dimensional Spectrum and Mean Vector Calculations for the embeddings post attention. Post this experiments were performed using Bhattacharya distance on the distribution of distances in distance matrices to learn about the optimal length for the texts, given that it is necessary to perform the following experiments over a large number of texts. These above experiments set the groundwork for future experiments in giving a basic understanding of the embeddings, attention mechanism and optimal text length.

Following this it was necessary to learn how foundational models performed when there were perturbations in the input text sequences. To do so perturbations in the form of Similar and Dissimilar substitutions upto varying extents (1-10) were introduced into texts to learn how the Euclidean and Hausdorff Distance changed. During this experiment it was observed that both the distances increased as Edit Distance increased signalling the possibility of outliers (Fig 1.3).

The above experiment setup the stage for the division of texts into two categories General and Specialized as it was hypothesized that General texts have a more uniform distribution compared to Specialized texts. To test this hypothesis a set of experiments were designed to study how the embeddings were structured in the above texts and how they responded to perturbations/substitutions by studying the spread of distances in the distance matrices, which included a comparative study between generalized and specialized, generalized and altered generalized, and specialized and altered specialized (Fig 1.4). Though these set of experiments did not yield any conclusive differences between general and specialized texts it proved that distributions of distance in the distance matrices were not the ideal way to explore how the LLM’s reacted to input perturbations indicating the necessity to explore other metrics to study them objectively.

Figure 1.1

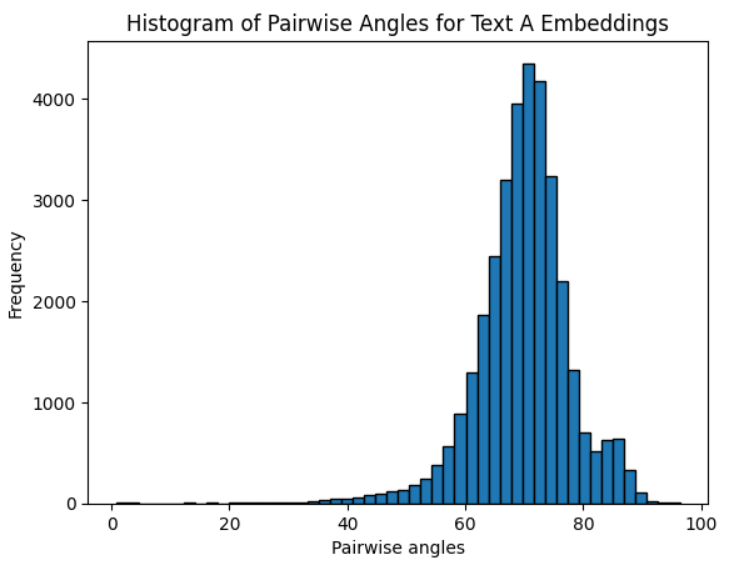


Figure 1.2

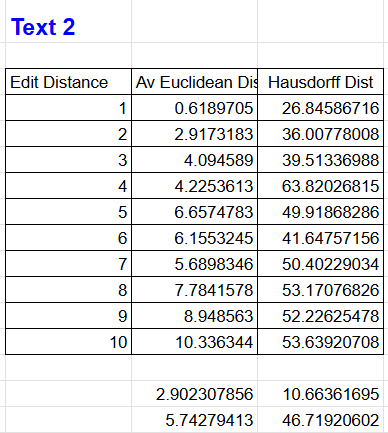


Figure 1.3

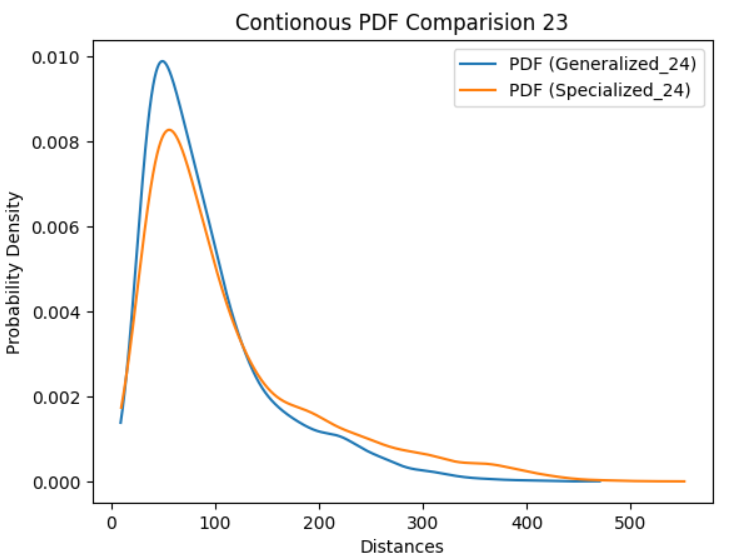


Figure 1.4

RELATED WORK

Making sure large language models (LLMs) are resilient to different threats has become a crucial research topic as these models are used in a growing number of applications. According to Esman et al. (2024), robustness in the context of LLMs refers to the models' capacity to continue performing well under a variety of inputs, distributions, and perturbations without sacrificing dependability or safety.

Numerous recent studies have looked at how robust LLMs are, pointing out major risks and suggesting ways to strengthen them. Esman et al. (2024) conducted a thorough survey that classifies the primary challenges to LLM robustness, such as distribution shift, adversarial attacks, and safety/reliability concerns. There are several methods to improve resilience, including rapid engineering, out-of-distribution (OOD) detection, and adversarial training.

With a focus on distribution change, Gao et al. (2022) assess how resilient well-known LLMs—such as GPT-3, BERT, and T5—are to "natural" distribution shifts, which are more indicative of deployment scenarios in the real world. Significant performance degradation can still happen even with larger and more recent models, especially for certain distribution shift types as diachronic and cross-lingual shifts. Comprehensive robustness evaluation is emphasized, as is the necessity of additional developments to increase LLM's resilience to various distribution alterations.

Samanta et al. (2023) investigate methods like data augmentation, domain adaptation, adversarial training, and uncertainty-aware prediction to lessen the effects of distribution shift. Their empirical analysis shows how well these strategies work together to improve LLMs' resilience to different distribution shift benchmarks.

Researchers have also looked at other facets of LLM robustness in addition to distribution shift. In order to increase the robustness of these models, Esman et al. (2024) stress the significance of interpretability and transparency in addition to the necessity of addressing the ethical and societal ramifications. Sharma et al. (2024) also stress the significance of creating comprehensive robustness frameworks that incorporate several robustness factors, such as safety and dependability.

Given their increasing relevance and influence, the literature now in publication emphasizes how vital it is to guarantee the stability of big language models. Although a number of robustness challenges, such as distribution shift, adversarial attacks, and safety/reliability issues, have been better understood and addressed, more research is required to create comprehensive and scalable solutions for the responsible and dependable deployment of LLMs in real-world applications.

GOAL OF THE THESIS

The primary goal of the Thesis is to establish the relation between LLM-base embeddings and meaning as it is understood by humans. . In the precursor of the experiments that were described above a rather general but vague idea has been established as to how embeddings function quantitatively. This opens up more opportunity to study how embeddings capture the meaning in the texts and how close they are to the way humans understand Natural Language. The importance of this study in its expected contribution to understanding of the degree to which LLM’s create and process embeddings that are in alignment to Human Understood Semantics.

The above objective is achieved by introducing possible metrics that reflect the robustness of LLM’s and measuring how they respond to different types of perturbations. This would give us a good understanding as to how LLM’s have created embeddings and processes them. The experiments on robustness are designed to establish that embeddings are not only stable and continuous but also provide metrics to measure the robustness of embeddings quantitatively.

The Thesis can be further extended to study Clustering using stable continuous embeddings and intrinsic space representations to reflect human understanding of topics. Different types of representation for embeddings such as Average, [CLS], Attention heads, Maximum and so on can be used to represent the overall context of the texts. But the study aims to find which of the embeddings can be used to best differentiate the texts based on topics like Physcology, Children Books, News, etc. Experiments will be designed with Foundational Models like GPT2 and BERT and different representation of embeddings with the pursuit to learn how each type of embedding will represent the topic under consideration and differentiates it from other topics. This can be the basis for future studies in developing specialized LLM’s based on different subject topics.

In summary, the Thesis is expected to string together the primary objective of establishing that embeddings correspond to human notion of meaning by studying the continuity and robustness of embeddings and extending this to differentiate topics as interpreted by humans by using the appropriate metrics and best type of embedding representation for this context.

FIRST EXPERIMENTS

In the pursuit to quantify aspects related to the Robustness of LLM’s. Initial experiments were designed to devise metrics to measure such properties. The first experiment involves taking short texts and introducing perturbations in the form of Similar and Dissimilar substitutions at the token level and measuring how LLM’s response is captured in terms of these metrics. The Hypothesis of this experiment is for small perturbations in the source texts, the initial difference (ro) is larger than the final observed difference(epsilon). More exactly, let be the initial text sequence, and let be a text sequence obtained from initial sequence by a small perturbation. After tokenization and initial embedding layer each of the text sequences is transformed into corresponding sequences of vectors in . We define distance function Let also and be sequences of vectors in obtained by embedding after attention (the last layer of encoder). Then we define distance function between after attention embeddings similarly to the distance function above, i.e., We can now measure the effects of attention. A new measure is called Attenuation Factor (af = epsilon/ro). It measures the effective disturbance observed relative to the initial induced disturbance due to perturbation.

The results (Fig 2.1) obtained from the experimental setup over a small number of texts give a positive sign that the hypothesis may hold up. It involves measuring the ro, epsilon and af metrics for these texts for both Similar Substations and Dissimilar Substitutions. Other observations indicate the possibility that perturbation behave like noise where the peak value of disturbance in the final embeddings is more often than not is observed at the position where the substitution happens.

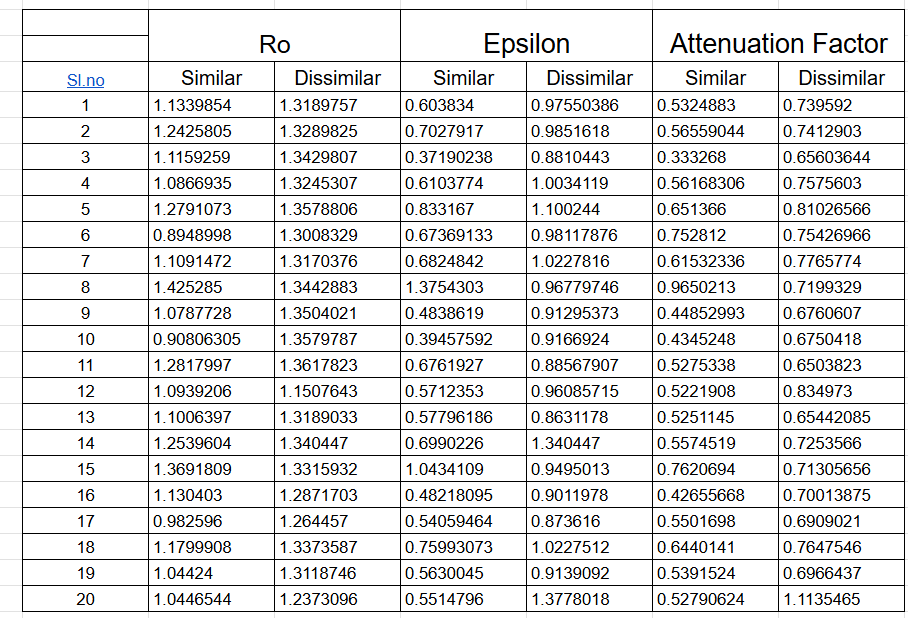


Fig 2.1

Experimental DESIGN

This research will pursue a quantitative approach to learn about the Robustness of Embeddings in LLM’s and robustness of cauterization based on robust embeddings. We plan to examine experimentally a few meaningful (known and new) notions of robustness in embeddings and in clusterization. The research methodology is centred around designing experiments related to LLM’s to collect data, analyse and reach conclusion or generate hypothesis. It follows through a chain of experiments which build upon the learning from the previous experiments to establish the objective.

The experimental research design involves building on previously established theories and hypothesis derived from previous work in the field of LLM’s. Each experiment is designed to study both GPT-2 and BERT, considering a hypothesis with an objective to either establish or debunk it. They are setup in Jupyter Notebooks using libraries like HuggingFace Transformers for using pre-trained models, ML libraries like PyTorch to perform operations, Scikit-learn and pandas to analyse data and so on. Many rounds of experiments are conducted to refine the data, correct biases and also involve error handling and logging to ensure experiments can be performed on a large scale.

The data collected from experiments forms the cornerstone for analysis and arriving at conclusions. This data is presented on Spreadsheets which are inherently generated as a part of the experiment. Data Analysis tools from Scikit-learn, Numpy and Pandas library are used to manipulate and interpret both the intermediary and final data which is collected. Additionally, tools from spreadsheets such as functions, charts are used for better presentation of data. Pre experiment data collection mainly involves collection of different type of texts such as general and non-general texts from online articles, repositories, etc. which are accessed either manually or directly during experimentation.

The analysis of data as an important step is performed at different stages of the experiment. Intermediary data is mainly derived from LLM outputs which are in general manipulated using algebraic methods like vector operations, matrix multiplications and also algorithms like PCA. Final data is collected from intermediate data and involves cleaning of data and used mainly as the results of the experiments. Many statistical methods like Normal Distribution, Power Law, Probability distribution, Hypothesis Testing, Bhattacharya Distance, Hellinger Distance and so on are used to create a better understanding.

Meetings conducted by the Thesis Supervisor on weekly or bi-weekly basis are the backbone of the Research Thesis. Meetings generally involve discussion of the results from current experiments, to arrive at conclusions, generate hypothesis, brainstorm ideas and create the basis for new experiments to improve the understanding from current experiments.

REFERENCES

1. Guanting Dong, Jinxu Zhao, Tingfeng Hui, Daichi Guo, Wenlong Wan, Boqi Feng, Yueyan Qiu1, Zhuoma Gongque, Keqing He, Zechen Wang, Weiran Xu: Revisit Input Perturbation

Problems for LLMs: A Unified Robustness Evaluation Framework for Noisy Slot Filling Task. (2023)

1. Zhouxing Shi1, Huan Zhang2, Kai-Wei Chang2, Minlie Huang1, Cho-Jui Hsieh2: Robustness Verification for Transformers. (2020)
2. Xiaojing Fan, Chunliang Tao: Towards Resilient and Efficient LLMs: A Comparative Study of Efficiency, Performance, and Adversarial Robustness. (2024)
3. Haoyu Wang, Guozheng Ma, Cong Yu, Ning Gui, Linrui Zhang, Zhiqi Huang, Suwei Ma, Yongzhe Chang, Sen Zhang, Li Shen, Xueqian Wang, Peilin Zhao, Dacheng Tao2: Are Large Language Models Really Robust to Word-Level Perturbations? (2023)
4. Nura Aljaafari, Danilo S. Carvalho, André Freitas: Interpreting token compositionality in LLMs: A robustness analysis. (2024)
5. Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, Kamalika Das: SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models. (2024)