ROOSEVELT
UNIVERSITY

Ashish Karikere

Measuring Robustness of Embeddings and Evaluating Text Similarity with Attenuation Factor

MS Thesis

Submitted to the Department of Computer Science

in partial satisfaction of requirements for the degree

of Master of Science in Computer Science

Spring 2025

Faculty Committee:

05/07/2025

Prof. Alex Wolpert
(Thesis Advisor)

EDantsin

Prof. Evgeny Dantsin

05/07/2025

Prof. Jerry Schnepp

Published: May 2025

ABSTRACT

Advancements in the field of Generative AI with LLMs have had an unprecedented impact, along with this there has been significant press on the need for AI safety development, for which robustness of LLM embeddings is integral. How to measure Robustness? We develop three new metrics to validate robustness in LLMs and measure the impact of word-level perturbations on short texts. Attenuation Factor, one of these metrics measures the impact of perturbations on the output compared to the input. We experiment with similar and dissimilar perturbations using different LLM embedding models (Bert-based), to quantify the effect of the perturbations with the above metrics and establish their consistency. Through Linear Regression and KL Divergence studies on the Attenuation Factor measure, we establish a new type of embedding which we call Sequences that provide better separation between Similar and Dissimilar substitutions compared to commonly used Center and CLS embeddings and thereby showing better alignment with Human Understood Semantics.

ACKNOWLEDGEMENTS

I thank my Thesis Advisor: Professor Alex Wolpert for accepting to supervise my work through the Thesis and guide me through the past one year. It was only through your help that not only was I able to gain knowledge in this field but also importantly learned the ways of research and take the initial steps in becoming an Independent Researcher.

I thank Professor Evgeny Dantsin and Professor Jerry Schnepp for being part of the committee and taking time to go through my Thesis. I hope my work interests you.

Lastly, to my family: My mom, my dad, and my sister. Thank you for the time and support you have given me not only during this Thesis but also during my entire Masters journey.

Measuring Robustness of Embeddings and Evaluating Text Similarity with Attenuation Factor

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 PROBLEM OVERVIEW

Current Progress in the field of Generative AI in terms of models' sizes and capabilities of Large Language Models has demonstrated an increase in their ability to tackle multiple Natural Language Processing Tasks and Problem Solving. This has implications in various sectors from IT to Education where we have seen an increase in the applications that rely or are centered around LLMs. These LLMs are often trained with clean and high-quality data that is often different compared to the large amounts of unstructured and disorganized data they have to process in real world settings. Due to this there is large scale demand to make these AI systems more adaptable in production environments where their performance will not be compromised.

Robustness in LLMs refers to the ability of these models to perform normally in various scenarios. Current research in this area is focused on multiple aspects from understanding different challenges such as distributional shifts to creating several methods to improve LLM resilience such as rapid engineering to creating datasets that simulate real world input data. Progress made in this area will help us in bridging the gap by making AI more safe, reliable and aligned to Human Values.

The thesis aims to expand the notion of Robustness measurement in Foundational Models. We introduce three new robustness metrics rho, epsilon and Attenuation Factor which measure the extent of perturbations and their relative impact at the output and input. Experimentation with BERT for multiple types of embeddings and distance measurements have helped in validating consistency of the above metrics, Robustness of BERT and establishing Center Embeddings as the most robust among the above embeddings. Further empirical analysis has led us to establish Sequence embeddings to be the best in separating different types of perturbations and hence having better representation of Human Understood Semantics. Efforts to separate different types of texts such as Generalized and Specialized texts using techniques such as Clusterization and In-Between distance distributions on their embeddings have been made but no conclusive results were obtained.

The Thesis contributes to the studies in LLM Robustness and Text Similarity with an aim to make AI Systems more safe, reliable and more aligned with Human Values. It lays the foundation for future work on Robustness and helps expand our notion of Text Similarity.

1.2 RELATED WORK

As the use of large language models (LLMs) in practical applications grows, ensuring their robustness has emerged as a crucial area of research. Robustness is usually defined as the capacity of LLMs to continue operating consistently under a variety of inputs, distributions without sacrificing dependability or safety (see e.g., [1]) .

According to recent assessments [1], the main obstacles include safety concerns, adversarial attacks, and distribution shifts. A number of mitigation techniques, such as out-of-distribution (OOD) detection, prompt engineering, and adversarial training, have been put forth [2] authors assesses how natural distribution variations, such as diachronic and cross-lingual shifts on models like GPT-3, BERT, and T5, find that even in large language models, performance degrades significantly. [4] proposes techniques including data augmentation, domain adaptation, and uncertainty-aware prediction, which collectively improve resilience across standard benchmarks.

Model resilience is also impacted by representation-level problems in addition to input-level perturbations. It has been shown in [3] that popular embedding models (e.g., RoPE, APE) display large positional biases—where alterations in the beginning of a document disproportionately affect the embedding—highlighting fragility in downstream tasks like information retrieval. This shows that embedding-level robustness should be an intrinsic part of overall model evaluation.

Methodology of model evaluation is another area of intensive study. In particular, it relies heavily on construction of data sets that most often is based on selection of synonyms/ antonyms to be used in substitutions necessary to introduce perturbations in robustness studies. Promising method to address this issue is proposed in [8] where authors introduce the Distiller, a unique approach to categorize synonyms and antonyms by projecting pre-trained embeddings onto specialized subspaces. Their model enforces semantic features like symmetry and transitivity without needing external corpora, obtaining excellent accuracy and demonstrating that subspace modifications can boost embedding fidelity for fine-grained semantic reasoning.

In addition to methodological advances, many studies (e.g. [1], [6]) underline the relevance of interpretability, ethical considerations, and the development of integrated robustness frameworks. These perspectives emphasize the greater problems involved in adopting LLMs reliably and ethically.

Overall, the research underlines the growing necessity for rigorous robustness evaluation—both at the input and representation levels. Adversarial Robustness and Distribution generalization are mainly focused, addressing structural bias and semantic interpretability are important for improving the real-world reliability of LLMs.