Statistical Stability Analysis of Large Language Model Embeddings Across Prompt Variations and Model Architectures

[1]Pawar Abhijit
Assistant Professor, School of information Technology, Indira University,
Pune-411033, India.

[2]Thakare Sarika
Assistant Professor, School of information Technology, Indira University, Pune-411033, India.

[3]Ajagekar Pournima
Assistant Professor, School of information Technology, Indira University, Pune-411033, India.

DOI: 10.29322/IJSRP.16.02.2026.p17020
https://dx.doi.org/10.29322/IJSRP.16.02.2026.p17020

Paper Received Date: 11th January 2026
Paper Acceptance Date: 6th February 2026
Paper Publication Date: 12th February 2026

Abstract

Large language models (LLMs) create numerical representations of text, called embeddings. These are key to applications like search and recommendations. But, these embeddings are not always consistent, they can change significantly with minor rewording of the prompt or when using a different model.

This study systematically tests how much embeddings vary across prompts and models using statistical analysis. We find that both wording and model choice meaningfully affect the results, highlighting the need for stability testing in real-world systems.

Keywords: Embedding stability, prompt design, cosine similarity, LLM reproducibility.

I. INTRODUCTION

Embeddings from LLMs power many modern AI tools. But their consistency is often assumed, not proven. Factors like prompt phrasing and model architecture can alter the embedding for the same text, affecting reproducibility. This work shows why evaluating embedding stability statistically is essential for building reliable applications.

II. RESEARCH GAP

1. There is limited quantitative analysis of how small prompt variations influence embedding vectors.
2. Comparative studies across multiple LLM architectures with respect to embedding stability are scarce.
3. Statistical hypothesis testing is rarely applied to examine whether observed differences in embeddings are significant or merely random noise.

III. OBJECTIVES OF THE STUDY

The present study is undertaken with the following specific objectives:

1. To examine the effect of prompt variations on the statistical properties of Large Language Model embeddings.
2. To compare the stability of embeddings generated by different LLM architectures.
3. To quantify the variability in embeddings using descriptive statistical measures such as mean, variance, cosine similarity.
4. To test whether differences in embeddings across prompt formats and model architectures are statistically significant using appropriate inferential statistical techniques.
5. To assess the reproducibility and reliability of LLM embeddings for use in downstream analytical and decision-support applications.

IV. METHODOLOGY

1. Research Design

The study adopts a quantitative experimental research design to analyze the statistical stability, variability, and reproducibility of Large Language Model embeddings across prompt variations and model architectures.

2. Data Source and Corpus Construction

A dataset of **198 base prompts** was obtained from an Excel file prepared by the researcher. The prompts span multiple academic domains including Science, Life Science, Geography, History, English, Mathematics, Physics, Chemistry, and Biology. Each base prompt is treated as a semantically stable concept.

3. Prompt Engineering Strategy

For each base prompt, three controlled prompt variations are generated:

1. **Neutral Prompt:** Original base prompt.
2. **Elaborative Prompt:** The base prompt rewritten with a detailed instructional context.
3. **Concise Prompt:** A shortened keyword-style version of the base prompt.

4. Model Selection

Three different Large Language Model architectures are selected:

1. Model A – all-MiniLM-L6-v2 (Lightweight MiniLM, 6 layers)
2. Model B – all-mpnet-base-v2 (Large MPNet-base embedding model)
3. Model C – paraphrase-MiniLM-L12-v2 (Medium-depth MiniLM, 12 layers, paraphrase-tuned)

5. Embedding Generation

For every base prompt, embeddings are extracted for all combinations of prompt types and model architectures.

Total embeddings generated:

198 base prompts × 3 prompt types × 3 models = **1,782 embedding vectors**.

All embeddings are of fixed dimensionality and generated under identical system settings.

6. Descriptive Statistical Measures

To analyze variability, the following metrics are computed: mean, variance, minimum and maximum of cosine similarity scores between prompt variations.

7. Hypothesis Testing Framework

Null Hypothesis (H_0): There is no statistically significant difference in cosine similarity scores across prompt variations and model architectures.

8. Inferential Statistical Techniques

1. Paired sample t-test is applied to compare embeddings obtained from different prompt types for the same base prompt.
2. One-way ANOVA is used to compare cosine similarity values across different prompt-pair types within each model architecture.

A significance level of **0.05** is used.

9. Reproducibility and Reliability Assessment

To assess reproducibility and reliability:

1. Embedding extraction is repeated under identical experimental conditions.
2. Intraclass Correlation Coefficient (ICC) is calculated to quantify consistency.
3. Stability thresholds are defined using cosine similarity benchmarks.

10. Tools and Software

All statistical computations are conducted using Python libraries including NumPy, Pandas, and SciPy.

V. RESULTS AND ANALYSIS

1. Descriptive Statistics of Cosine Similarities

Method

Mean, standard deviation, minimum, and maximum values of cosine similarity were computed for the three prompt pairs: Neutral–Elaborative, Neutral–Concise, and Elaborative–Concise, separately for Model A, Model B, and Model C.

Criteria

High mean cosine similarity and low standard deviation indicate greater semantic stability of embeddings across prompt variations.

Table No. 1: Descriptive Statistics of Cosine Similarities

Model	Prompt Pair	Mean	Std. Dev.	Min	Max
Model A	Neutral–Elaborative	0.8761	0.0367	0.7249	0.9358
Model A	Neutral–Concise	0.9550	0.0153	0.9008	0.9842

Model A	Elaborative–Concise	0.8898	0.0291	0.7900	0.9441
Model B	Neutral–Elaborative	0.8966	0.0328	0.7244	0.9484
Model B	Neutral–Concise	0.9412	0.0197	0.8599	0.9754
Model B	Elaborative–Concise	0.9180	0.0231	0.8216	0.9606
Model C	Neutral–Elaborative	0.8539	0.0339	0.7506	0.9271
Model C	Neutral–Concise	0.9286	0.0249	0.8032	0.9710
Model C	Elaborative–Concise	0.8427	0.0363	0.7154	0.9197

Comparative Interpretation of Mean Cosine Similarity Graphs

The bar charts for Models A, B, and C display the mean cosine similarity for three prompt-pair types: Neutral–Elaborative, Neutral–Concise, and Elaborative–Concise.

Model A

1. The Neutral–Concise pair shows the highest mean similarity (≈ 0.955).
2. Neutral–Elaborative and Elaborative–Concise pairs record lower means (≈ 0.876 and 0.890 respectively).
3. The gap between Neutral–Concise and the other two pairs is visually prominent, indicating strong stability for concise rephrasing.

Model B

1. Mean similarity for Neutral–Concise is ≈ 0.941, which is again the highest.
2. Neutral–Elaborative and Elaborative–Concise pairs show moderate similarity (≈ 0.897 and 0.918).
3. The separation among the three bars is smaller than in Model A, indicating relatively higher robustness.

Model C

1. Mean Neutral–Concise similarity is ≈ 0.929, still the highest among the three prompt pairs.
2. Neutral–Elaborative and Elaborative–Concise pairs drop to ≈ 0.854 and 0.843, respectively.
3. The larger drop in these bars shows that Model C is more sensitive to prompt variation.

2. Prompt Variation Stability Results

Method

Cosine similarity was computed for each of the 198 base prompts between the following prompt-pair types: Neutral–Elaborative, Neutral–Concise and Elaborative–Concise

The analysis was performed separately for Model A, Model B, and Model C.

Criteria

Cosine similarity values closer to 1 indicate greater semantic stability. Values equal to or above 0.90 were considered to represent high semantic similarity.

Result Output

The mean ± standard deviation of cosine similarity values for each prompt-pair type are shown below.

Table No. 2: Prompt Variation Stability Results

Model	Neutral–Elaborative	Neutral–Concise	Elaborative–Concise
Model A	0.8761 ± 0.0367	0.9550 ± 0.0153	0.8898 ± 0.0291
Model B	0.8966 ± 0.0328	0.9412 ± 0.0197	0.9180 ± 0.0231
Model C	0.8539 ± 0.0339	0.9286 ± 0.0249	0.8427 ± 0.0363

3. Effect of Prompt Pair Type within Each Model (One-Way ANOVA)

Method

One-way Analysis of Variance was conducted separately for Model A, Model B, and Model C to compare mean cosine similarities across the three prompt-pair types: Neutral–Elaborative, Neutral–Concise, and Elaborative–Concise.

Criteria

A p-value less than 0.05 indicates a statistically significant difference in mean cosine similarity across prompt-pair categories.

Table No. 3: One-way Analysis of Variance

Model	F-statistic	p-value
Model A	433.9078	0.0000
Model B	148.2102	0.0000
Model C	420.2904	0.0000

4. Hypothesis Testing Results

Method

Paired sample t-tests were applied to compare cosine similarity scores between the following prompt-pair combinations for each model: Neutral–Elaborative vs Neutral–Concise and Neutral–Elaborative vs Elaborative–Concise.

Table No. 4: Paired sample t-tests

Model	Comparison	t-Statistic	p-Value
Model A	Neutral–Elaborative vs Neutral–Concise	−38.03	0.0000
Model A	Neutral–Elaborative vs Elaborative–Concise	−10.67	0.0000
Model B	Neutral–Elaborative vs Neutral–Concise	−27.02	0.0000
Model B	Neutral–Elaborative vs Elaborative–Concise	−13.83	0.0000
Model C	Neutral–Elaborative vs Neutral–Concise	−35.64	0.0000
Model C	Neutral–Elaborative vs Elaborative–Concise	9.12	0.0000

5. Reproducibility and Reliability Metrics

Method

Intraclass Correlation Coefficients were computed using six models: ICC(1), ICC(2), ICC(3) for single raters and ICC(1k), ICC(2k), ICC(3k) for average raters, separately for Model A, Model B, and Model C.

Criteria

An ICC value above 0.75 was interpreted as excellent reliability, values between 0.50 and 0.75 as moderate reliability, and values below 0.50 as poor reliability.

Interpretation of ICC Reliability Table:

ICC Type	Meaning
ICC1 / ICC1k	Absolute agreement of single / average ratings
ICC2 / ICC2k	Consistency of random raters
ICC3 / ICC3k	Consistency of fixed raters (your experimental setup)

Table No. 5: Intraclass Correlation Coefficients

Model	ICC Type	ICC	95% CI
Model A	ICC3k	0.8580	[0.82, 0.89]
Model B	ICC3k	0.8350	[0.79, 0.87]
Model C	ICC3k	0.8907	[0.86, 0.91]

Table No. 6: ICC1 Table (Single Raters – Absolute Agreement)

Model	ICC1
Model A	−0.025
Model B	0.253
Model C	0.010

VI. CONCLUSIONS:

Neutral–Concise prompts produced the most stable embeddings across all models. They showed the highest cosine similarity with very little variation. Other prompt pairs demonstrated notably lower stability, with Model C being particularly sensitive to changes in phrasing.

Preliminary descriptive statistics, including mean and standard deviation, effectively captured these differences.

Hypothesis testing methods such as One-way ANOVA and paired tests, consistently rejected the null hypothesis for each model, confirming that prompt wording significantly influences embedding similarity.

Reliability analysis revealed high intraclass correlation ICC(3,k) scores. This indicate that consistency in aggregated results across prompt variations. In contrast, low ICC(1) values suggest that individual prompt embeddings are inherently unstable and should not be used in isolation for reliable measurement.

This study shows that prompt wording significantly affects embedding behavior. To make sure a reliable evaluation of embedding stability, it is necessary to adopt standardized prompts and to base conclusions on aggregated statistical measures rather than individual outputs.

IMPLICATIONS

Small changes in prompt wording can change embedding similarity values. This can affect search results in retrieval systems. It can also change cluster formation in analytical pipelines.

Using short and fixed prompt templates improves stability. Different models react differently to prompt changes. Therefore, stability should be tested for each model before use.

For long-term systems, prompt formats and model versions should be documented. Periodic re-indexing helps maintain consistency in stored embeddings.

LIMITATIONS

1. Embedding results may slightly change due to API nondeterminism.
2. The effect of temperature on embeddings was not directly studied.
3. Hidden updates to language models may affect reproducibility.
4. Only English academic prompts were used in this study.
5. Results may not fully apply to other languages or domains.

RECOMMENDATIONS

1. **Standardization of Prompt Design** Researchers and practitioners should adopt standardized concise prompt templates when comparing semantic embeddings across models to minimize unintended variability.
2. **Architecture-Specific Evaluation** Embedding stability benchmarks should be performed separately for each model architecture, as prompt sensitivity varies significantly across models.

3. **Future Research Directions** Future studies should incorporate domain-wise analysis and explore the effects of multilingual prompts on embedding stability.

REFERENCES

1. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.
2. Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163.
3. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
4. Fisher, R. A. (1925). Statistical methods for research workers. Oliver and Boyd.
5. Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.
8. Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. Proceedings of NAACL-HLT, 1073–1094.
9. Ethayarajh, K. (2019). How contextual are contextualized word representations? Proceedings of EMNLP-IJCNLP, 55–65.
10. Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of EMNLP, 6894–6910.
11. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of ICLR.
12. Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842–866.
13. Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305.
14. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of AAAI, 3207–3214.
15. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT, 610–623.