Skoltech

MASTER'S THESIS

Analyzing Transformer Language Models with Interpretative Techniques

Master's Educational Program: Data Science

Student: _____ Pavel Shtykov
signature

Research Advisor: _____ Serguei Barannikov
signature

PhD, Leading research scientist

Moscow 2025

Skoltech

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Анализ трансформерных языковых моделей с помощью интерпретационных методов

Магистерская образовательная программа: Науки о данных

Студент: _____ Павел Штыков
подпись

Научный руководитель: _____ Сергей Баранников
подпись

к.м.н., ведущий научный сотрудник

Москва 2025

Analyzing Transformer Language Models with Interpretative Techniques

Pavel Shtykov

Submitted to the Skolkovo Institute of Science and Technology on June 10, 2025

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords: keyword1, keyword2, keyword3, keyword4

Research advisor:
Name: Serguei Barannikov
Degree, title: PhD, Leading research scientist

Contents

Chapter 1
Introduction

Relevance. Transformer-based language models have revolutionized natural language processing, yet they largely remain ”black boxes” with limited interpretability. Understanding these models is crucial for reliable development, bias mitigation, and architectural innovation. As transformers become integrated into critical applications, interpretability becomes both a technical challenge and ethical necessity. This thesis investigates how transformer models process information, focusing on the relationship between internal representations and output probabilities.

Main purpose of the research. This research aims to develop interpretative techniques for transformer language models that illuminate their internal mechanisms. Specifically, we investigate:

- The relationship between probability distributions from the language modeling head and geometric properties of internal representations
- Alternative formulations of the language modeling objective that enhance interpretability
- A theoretical framework conceptualizing transformer operation as inertial movement in embedding space
- Experimental validation of these interpretative approaches

Scientific novelty. This research contributes novel elements to transformer interpretability:

- Analysis of geometric relationships between final hidden states and vocabulary token embeddings
- A training approach replacing the traditional linear language modeling head with a learnable Gaussian kernel over k-nearest neighbors
- A theoretical framework interpreting transformer operation as inertial movement in embedding space
- Empirical validation through targeted experiments isolating specific aspects of transformer behavior

Unlike previous work focusing on isolated components, this research examines the holistic relationship between internal representations and model outputs.

Statements for defense.

1. Probability distributions from the language modeling head correlate strongly with distances between final hidden states and token embeddings, suggesting transformers learn to navigate embedding space.
2. Transformers trained with Gaussian kernels over k-nearest neighbors can achieve comparable performance while providing more interpretable representations.

3. Transformer operation can be conceptualized as inertial movement in embedding space, with attention mechanisms and feed-forward networks acting as guiding forces.
4. The proposed interpretative techniques provide actionable insights for developing more efficient transformer architectures.

Chapter 2
Author contribution

In this thesis, I have made the following contributions to the research presented:

- **Conceptualization:** I formulated the research questions and hypotheses regarding the relationship between transformer hidden states and token embeddings. I proposed the interpretation of transformer operation as inertial movement in embedding space, providing a novel framework for understanding these models.
- **Methodology:** I designed the experimental methodology to test the relationship between probability distributions from the language modeling head and distances between hidden states and token embeddings. Additionally, I developed the approach for replacing the traditional linear language modeling head with a learnable Gaussian kernel over k-nearest neighbors.
- **Implementation:** I implemented all code necessary for the experiments, including modifications to transformer architectures to support alternative language modeling heads and to extract internal representations for analysis. This implementation work included developing custom PyTorch modules for the Gaussian kernel approach and metrics for comparing probability distributions.
- **Experimentation:** I conducted all experiments described in this thesis, including training modified transformer models and analyzing the geometric properties of transformer hidden states in relation to token embeddings.
- **Analysis:** I performed the analysis of experimental results, including the calculation of NDCG scores between probability distributions and the interpretation of these results in the context of the proposed theoretical framework.

All aspects of this research were conducted under the supervision of Serguei Barannikov, who provided guidance on the theoretical foundations and experimental design. I acknowledge the use of open-source libraries and pre-trained models that served as the foundation for the experimental work.

Chapter 3
Literature review

3.1 Transformer Models and Interpretability

Transformer-based language models have revolutionized natural language processing [10], yet they largely remain ”black boxes” with limited interpretability. The field of transformer interpretability has grown significantly, with approaches like VisBERT by Aken et al. [1] providing visualization techniques for hidden states that offer insights into the model's internal processing.

3.2 Geometric Interpretations of Transformer Representations

A promising direction involves analyzing transformers through geometry in embedding space. Dar et al. [3] present a framework where transformer parameters are interpreted by projecting them into embedding space, demonstrating that both pretrained and fine-tuned models can be understood through this lens.

Singh [7] extends this by framing transformer dynamics as movement through embedding space, establishing a theory where intelligent behaviors map to paths in embedding space and context vectors are composed by aggregating token features. This aligns with our thesis's focus on interpreting transformers as inertial movement in embedding space.

The geometric properties of transformer hidden representations have also been explored by Valeriani et al. [9], providing insights into information encoding across transformer layers.

3.3 Alternative Formulations of Language Modeling

Traditional transformers use a linear language modeling head, but alternative formulations enhance interpretability and performance. Geva et al. [4] showed that feed-forward networks in transformers function as key-value memories related to embedding space, suggesting a connection between internal representations and vocabulary token embeddings.

The relationship between k-nearest neighbors and transformer attention has been explored by Haris [5], who provides a theoretical framework for k-NN attention using Gaussian sampling for efficient approximation. Kernelized transformer variants have been investigated by Chowdhury et al. [2] and Simpson et al. [6], exploring Gaussian kernels and their integration with transformer architectures.

3.4 Embedding Space Dynamics and Token Relationships

The relationship between transformer hidden states and token embeddings has been studied from various angles. Song and Zhong [8] investigate hidden geometry in transformers by disentangling position and context. The concept of distances between hidden states and token embeddings, central to our research, has been touched upon by Aken et al. [1] and Dar et al. [3], providing a founda-

tion for our investigation into the relationship between probability distributions and embedding distances.

3.5 Research Gaps and Opportunities

Despite these advances, several gaps remain:

1. The specific connection between probability distributions from the language modeling head and distances between hidden states and token embeddings remains underexplored.
2. Alternative formulations using learnable Gaussian kernels over k-nearest neighbors to token embeddings have not been thoroughly investigated for their impact on interpretability.
3. The conceptualization of transformer operation as inertial movement in embedding space has not been fully developed into a comprehensive framework with empirical validation.

Our research addresses these gaps by developing a more complete understanding of transformer operations in embedding space, focusing on the relationship between internal representations and output probabilities.

Chapter 4
Problem statement

4.1 Notation and Preliminaries

Transformer Language Models

We consider transformer-based language models as defined by [10]. Let $\mathcal{V}$ be a vocabulary of tokens, and let $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ be the embedding matrix, where d is the embedding dimension. For a sequence of tokens $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ with $x_i \in \mathcal{V}$, the transformer model processes the sequence through L layers, producing hidden states $\mathbf{h}_i^l \in \mathbb{R}^d$ for each token i at each layer l.

The final hidden state for token i is denoted as $\mathbf{h}_i^L$. The language modeling head, typically a linear layer, transforms this hidden state into logits over the vocabulary:

$$\mathbf{z}_i = \mathbf{W}\mathbf{h}_i^L + \mathbf{b} \tag{4.1}$$

where $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$ are the weight matrix and bias vector of the language modeling head, respectively. The probability distribution over the next token is then computed using the softmax function:

$$P(x_{i+1} = v | x_1, \ldots, x_i) = \frac{\exp(z_{i,v})}{\sum_{v' \in \mathcal{V}} \exp(z_{i,v'})} \tag{4.2}$$

Embedding Space

We define the embedding space as the d-dimensional vector space in which token embeddings and hidden states reside. For any token $v \in \mathcal{V}$, its embedding is denoted as $\mathbf{e}_v \in \mathbb{R}^d$, which corresponds to the v-th row of the embedding matrix $\mathbf{E}$.

The distance between two vectors $\mathbf{a}$ and $\mathbf{b}$ in the embedding space is measured using the Euclidean distance:

$$d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2 = \sqrt{\sum_{j=1}^{d} (a_j - b_j)^2} \tag{4.3}$$

4.2 Research Problems

This thesis addresses the following research problems:

Relationship Between Probability Distributions and Embedding Distances

We investigate the relationship between the probability distribution produced by the language modeling head and the distances between the final hidden state and token embeddings in the vocabulary. Specifically, we examine whether there exists a correlation between the probability assigned to a token and its proximity to the final hidden state in the embedding space.

For a given hidden state $\mathbf{h}$ and a token $v \in \mathcal{V}$, we study the relationship between $P(v|\mathbf{h})$ and $d(\mathbf{h}, \mathbf{e}_v)$, where $P(v|\mathbf{h})$ is the probability assigned to token v by the language model given hidden state $\mathbf{h}$, and $d(\mathbf{h}, \mathbf{e}_v)$ is the Euclidean distance between $\mathbf{h}$ and the embedding of token v.

We hypothesize that there exists an inverse relationship between these quantities, such that tokens with higher probabilities tend to have smaller distances to the hidden state in the embedding space. To quantify this relationship, we use the Normalized Discounted Cumulative Gain (NDCG) metric:

$$\text{NDCG} = \frac{\text{DCG}}{\text{IDCG}} \tag{4.4}$$

where DCG (Discounted Cumulative Gain) is computed based on the ranking of tokens according to their probabilities and inverse distances, and IDCG is the ideal DCG.

Gaussian Kernel Language Modeling

We propose an alternative formulation of the language modeling objective that replaces the traditional linear language modeling head with a learnable Gaussian kernel over k-nearest neighbors to token embeddings. The probability of the next token is computed as:

$$P(x_{i+1} = v|x_1, \ldots, x_i) = \frac{\exp\left(-\frac{d(\mathbf{h}_i^L, \mathbf{e}_v)^2}{2\sigma_v^2}\right)}{\sum_{v' \in \mathcal{V}} \exp\left(-\frac{d(\mathbf{h}_i^L, \mathbf{e}_{v'})^2}{2\sigma_{v'}^2}\right)} \tag{4.5}$$

where σ_v is a learnable parameter that controls the width of the Gaussian kernel for token v. This formulation makes the connection between hidden states and token embeddings more explicit and potentially more interpretable.

Inertial Movement in Embedding Space

TODO: This section will be filled in later with a description of the theoretical framework that interprets transformer operation as inertial movement in embedding space.

4.3 Research Objectives

Based on the research problems defined above, this thesis aims to:

1. Empirically verify the relationship between probability distributions from the language modeling head and distances between hidden states and token embeddings in the vocabulary.

2. Develop and evaluate a transformer model that uses a learnable Gaussian kernel over k-nearest neighbors to token embeddings instead of a traditional linear language modeling head.

3. Formulate and validate a theoretical framework that interprets transformer operation as inertial movement in embedding space.

4. Derive insights from these investigations that can guide the development of more interpretable and efficient transformer architectures.

Through these objectives, we seek to contribute to a deeper understanding of how transformer language models process and represent information, with a particular focus on the geometric properties of their internal representations in embedding space.

Chapter 5
Methodology

5.1 Analyzing Probability Distributions and Embedding Distances

Our approach to analyzing the relationship between probability distributions and embedding distances involves several key methodological steps:

Data Processing Pipeline

We implement a processing pipeline that extracts the necessary information from transformer models during inference:

1. For each input sequence, we capture the final hidden states from the last transformer layer. 2. We compute the full probability distribution over the vocabulary using the language modeling head. 3. We calculate the Euclidean distances between each hidden state and all token embeddings in the vocabulary. 4. We store these values in an efficient format that allows for subsequent analysis.

Correlation Analysis Framework

To quantify the relationship between probabilities and distances, we develop a correlation analysis framework that:

1. Ranks tokens according to both their probabilities and their proximity (negative distance) to the hidden state. 2. Computes rank correlation metrics such as Spearman's rank correlation coefficient. 3. Calculates the Normalized Discounted Cumulative Gain (NDCG) to measure the agreement between probability-based and distance-based rankings, with special attention to the top-k tokens.

Statistical Validation

To ensure the robustness of our findings, we employ statistical validation techniques:

1. Bootstrap sampling to estimate confidence intervals for correlation metrics. 2. Permutation tests to assess the statistical significance of observed correlations. 3. Cross-validation across different text domains to verify the consistency of the relationship.

5.2 Gaussian Kernel Language Modeling

Our methodology for developing transformer models with Gaussian kernel language modeling heads involves several innovative approaches:

Architecture Modifications

We modify the standard transformer architecture by:

1. Preserving the transformer encoder stack up to the final hidden states. 2. Replacing the linear language modeling head with a distance-based Gaussian kernel module. 3. Implementing efficient computation of distances between hidden states and token embeddings. 4. Adding learnable parameters that control the width of the Gaussian kernel for each token.

Kernel Variants

We explore three variants of the kernel-based approach:

1. **Pure Gaussian Kernel**: Using learnable σ parameters for each token to control the width of the Gaussian kernel.

$$\text{score}(v, \mathbf{h}) = \exp\left(-\frac{d(\mathbf{h}, \mathbf{e}_v)^2}{2\sigma_v^2}\right) \tag{5.1}$$

2. **Context-Dependent Scaling**: Dynamically adjusting the scaling of distances based on the hidden state.

$$\text{score}(v, \mathbf{h}) = \exp(-d(\mathbf{h}, \mathbf{e}_v) \cdot f(\mathbf{h})) \tag{5.2}$$

where $f(\mathbf{h})$ is a small neural network that outputs a positive scalar.

3. **Hybrid Approach**: Combining token-specific and context-dependent scaling.

$$\text{score}(v, \mathbf{h}) = \exp(-d(\mathbf{h}, \mathbf{e}_v) \cdot \sigma_v \cdot f(\mathbf{h})) \tag{5.3}$$

Training Methodology

Our training methodology includes:

1. Initialization strategies for the learnable σ parameters, including uniform initialization and token frequency-based initialization. 2. Specialized learning rate schedules for the kernel parameters versus the transformer parameters. 3. Regularization techniques to prevent overfitting of the kernel parameters. 4. Curriculum learning approaches that gradually increase the complexity of the training data.

Evaluation Protocol

We establish a comprehensive evaluation protocol that assesses:

1. Language modeling performance using perplexity and token prediction accuracy. 2. Parameter efficiency compared to standard transformer models. 3. Interpretability of the learned kernel parameters through visualization and analysis. 4. Generalization capabilities to out-of-distribution text.

5.3 Inertial Movement in Embedding Space

TODO: This section will describe the methodology for developing and validating the theoretical framework that interprets transformer operation as inertial movement in embedding space.

5.4 Implementation Framework

Software Architecture

Our implementation is structured around a modular framework that facilitates:

1. Easy modification of transformer architectures through a component-based design. 2. Efficient computation of distances and kernels using optimized tensor operations. 3. Comprehensive logging and analysis of model behaviors during training and inference. 4. Visualization tools for exploring the embedding space and model predictions.

Optimization Techniques

To handle the computational challenges of our approach, we implement several optimization techniques:

1. Approximate k-nearest neighbors algorithms to reduce the computational cost of distance calculations. 2. Caching mechanisms for token embeddings to avoid redundant computations. 3. Mixed-precision training to reduce memory requirements and increase throughput. 4. Parallelization strategies for distributing computations across multiple GPUs when necessary.

Reproducibility Framework

We establish a reproducibility framework that includes:

1. Deterministic initialization through fixed random seeds. 2. Comprehensive configuration management using hierarchical configuration files. 3. Automated experiment tracking with metadata about hyperparameters and results. 4. Version control for code, configurations, and key results.

This methodological framework enables us to systematically investigate the relationship between transformer hidden states and token embeddings, develop and evaluate alternative formulations of language modeling, and explore the geometric properties of transformer representations in embedding space.

Chapter 6
Numerical experiments

In this section the results of numerical experiments should be placed. They can confirm or contradict the expectation from the previous sections. The forms of result presentation are plots, tables, histograms, pictures, etc.

1. Enumerating languages and programs used in the research.
2. Describing in detail data sets you used.
3. Presenting the metrics used.
4. Presenting the results achieved in the experiment.
5. Comparing the results of the experiment with those obtained using other methods.
6. Outlining the benefits of the used method.

Table 6.1: The description of the table. Highlighting the main point following from this table.

	Method 1	Method 2	Method 3
Metric 1	**10**	20	30
Metric 2	0.9	0.4	**0.1**

Main requirements to plots:

- sufficiently large font size for legend, axis labels, axis ticks;
- proper scale of y-axis: linear or logarithmic;
- clear difference in the rendering of lines corresponding to different methods.

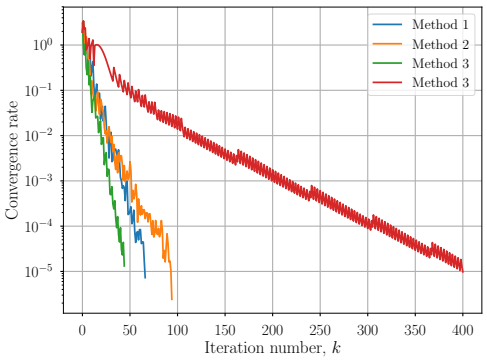

Figure 6.1: Comparison of the considered methods

Chapter 7
Discussion and conclusion

Summarize your work in this section.

1. Summary of the main results of the work that is consistent with the Aim and Objectives.
2. Overall position on the global research landscape.
3. Comparative critical analysis: what you have deduced from the findings and how these results relate to previous research or other studies.
4. Research limitations.

Acknowledgements

Write here acknowledgments of financial assistance for the conduct of research and to specific individuals who contributed to the science. Dedications are not recommended and must reference scientific contributions.

Innovations

Innovation component of your research project (if any), i.e.:

- Start-up potential
- Industrial application of research results

Bibliography

[1] Aken, B. v., Winter, B., Löser, A., and Gers, F. A. Visbert: Hidden-state visualizations for transformers. In *Companion Proceedings of the Web Conference 2020* (2020), pp. 207–211.

[2] Chowdhury, S., Solomou, A., and Dubey, A. On learning the transformer kernel. *arXiv preprint arXiv:2110.05312* (2021).

[3] Dar, G., Geva, M., Gupta, A., and Berant, J. Analyzing transformers in embedding space. *arXiv preprint arXiv:2209.02535* (2022).

[4] Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913* (2022).

[5] Haris, T. k-nn attention demystified: A theoretical exploration for scalable transformers. *arXiv preprint arXiv:2411.04013* (2024).

[6] Simpson, F., Davies, I., and Lalchand, V. Kernel identification through transformers. In *Advances in Neural Information Processing Systems* (2021).

[7] Singh, S. S. Analyzing transformer dynamics as movement through embedding space. *arXiv preprint arXiv:2308.10874* (2023).

[8] Song, J., and Zhong, Y. Uncovering hidden geometry in transformers via disentangling position and context. *arXiv preprint arXiv:2310.04861* (2023).

[9] Valeriani, L., Doimo, D., and Cuturello, F. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems* (2023).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

Appendix