Journal of Applied Mathematics and Physics, 2026, 14(5), 1775-1788
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Advanced Applications of Fundamental Mathematics in Large Language Processing Models

Meng Guo

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
Email: qlguo@sina.com

How to cite this paper: Guo, M. (2026) Advanced Applications of Fundamental Mathematics in Large Language Processing Models. *Journal of Applied Mathematics and Physics*, **14**, 1775-1788.
https://doi.org/10.4236/jamp.2026.145086

Received: April 8, 2026
Accepted: May 12, 2026
Published: May 15, 2026

Abstract

Large Language Models (LLMs), as a core technology in the field of artificial intelligence, rely heavily on fundamental mathematical theories for their underlying architecture, training optimization, inference mechanisms, and performance improvement. This article systematically summarizes the advanced applications of basic mathematical branches such as linear algebra, probability theory and mathematical statistics, functional analysis, optimization theory, information theory, topology, etc. in large language models. It deeply analyzes the implementation principles of mathematical theory in key links such as word embedding, attention mechanism, model training, inference optimization, and interpretability analysis. Combined with the international cutting-edge research results in the past five years, it elaborates on the core role of basic mathematics in breaking through the performance of large language models, improving theories, and expanding scenarios. At the same time, summarizing the bottlenecks currently faced by the application of mathematical theories, and looking forward to the research direction of deep integration between basic mathematics and large language models in the future, providing mathematical theoretical references and technical ideas for the theoretical research and engineering practice of large language models.

Keywords

Basic Mathematics, Large Language Model (LLM), Linear Algebra, Optimization Theory, Attention Mechanism, Interpretability

1. Introduction

1.1. Research Background and Significance

In recent years, the large language processing model with Transformer architec-

ture as the core has achieved leapfrog development [1]. From GPT series and LLaMA series to domestic Tongyi Qianwen, ERNIE Bot and other models, it has demonstrated strong capabilities in natural language understanding, text generation, machine translation, code writing, mathematical reasoning and other fields, becoming the core carrier of AI industrial applications. The essence of the large language model is a statistical learning model based on massive text data. The entire process from data representation and feature extraction to model optimization and inference output is the engineering implementation of basic mathematical theory: the construction of word embedding space relies on linear algebra, probability distribution modeling relies on probability theory and mathematical statistics, model training optimization is based on convex optimization and non convex optimization theory [2], semantic representation and feature abstraction cannot be separated from functional analysis, information transmission and efficiency optimization rely on information theory, and model interpretability and inference logic involve mathematical branches such as topology [3].

As the theoretical foundation of the large language model, the theoretical depth and application accuracy of basic mathematics directly determine the performance limit of the model. The current large language models face many problems such as high training costs, insufficient inference efficiency, poor interpretability, weak mathematical reasoning ability, and limited generalization performance, which are rooted in the fact that the application of underlying mathematical theories has not yet achieved refinement and depth [4]. Therefore, systematic research on the advanced application of basic mathematics in large language models, exploring the inherent relationship between mathematical theory and model architecture, training, and reasoning, can not only improve the theoretical system of large language models, but also provide new ideas for solving existing technical bottlenecks in models, and promote the development of large language models towards more efficient, controllable, and universal directions, which has important theoretical research value and engineering practical significance [5].

1.2. Current Research Status at Home and Abroad

In recent years, the international academic and industrial communities have attached great importance to the interdisciplinary research of basic mathematics and large language models, and have achieved a series of breakthrough results. In terms of linear algebra applications, researchers use random matrix theory to analyze the distribution characteristics of Transformer model weight matrices [6], reveal the evolution law of parameters during model training, and achieve efficient compression and initialization optimization of model parameters; Using tensor decomposition techniques to optimize the model attention matrix, reducing the computational complexity of long sequence text processing, and improving the ability to understand long texts [7]. In the field of probability theory and mathematical statistics, based on Bayesian theory, reconstruct the attention mechanism of large language models, construct a probabilistic semantic representation

framework, and improve the generalization performance of models in small sample scenarios; By using Hidden Markov Models and Conditional Random Field Optimization Models for sequence annotation and text generation logic, the logic and coherence of text output are enhanced [8].

In the application of optimization theory, researchers propose an adaptive learning rate optimization algorithm for non convex optimization problems in large language models, combined with gradient descent theory to improve training strategies and accelerate model convergence speed; Based on the dual theory of convex optimization, a loss function optimization framework for model training is constructed to alleviate the problems of model overfitting and gradient vanishing [9]. In the field of functional analysis and topology, the semantic space of word embeddings is characterized by Hilbert space, and the distribution of semantic features of the model is analyzed using topological manifold theory, providing mathematical basis for the interpretability of the model; Construct an evaluation system for model inference ability based on topological invariants, and quantitatively analyze the logical inference level of the model.

Domestic research focuses on the application of mathematical theory in the engineering of domestic large language models, optimizing the information transmission efficiency of models and reducing training energy consumption under the guidance of information theory; Combining combinatorial optimization theory to optimize model decoding algorithms and improve text generation speed [10]. However, overall, domestic research tends to focus on engineering applications, with insufficient in-depth exploration and theoretical innovation of basic mathematical theories, and a certain gap compared to top international research.

1.3. Research Content and Framework

This article focuses on the core theories of various branches of basic mathematics and conducts research around the entire process of the large language model. Firstly, it outlines the core architecture and mathematical foundation of the large language model. Secondly, it elaborates on the advanced applications of linear algebra, probability theory and mathematical statistics, functional analysis, optimization theory, information theory, and topology in the model in chapters. Combining mathematical formulas and cutting-edge case analysis, the application principles are analyzed. Finally, the research bottlenecks are summarized and future research directions are discussed [11]. The full text is divided into seven chapters, following the logical framework of "theoretical basis application analysis problem summary prospect", systematically presenting a deep integration path of basic mathematics and large language models.

1.4. Overview of Search and Literature Grouping Methods

This paper employs a systematic review method for research, limiting the scope of literature retrieval to high-level papers, preprints, and monographs published from 2020 to 2026. The search databases include arXiv, IEEE Xplore, ACM Digital

Library, Web of Science, CNKI, and top conference proceedings (NeurIPS, ICML, ICLR, ACL, etc.). The core search terms include: large language models, Transformers, attention mechanisms, probability theory, Bayesian inference, functional analysis, optimization theory, information theory, topology, interpretability, etc. The papers are grouped according to six major branches of mathematics: linear algebra, probability theory and mathematical statistics, functional analysis, optimization theory, information theory, and topology, corresponding to modules of model representation, modeling, training, optimization, efficiency, and interpretability, respectively. This approach ensures reproducibility and ease of evaluation.

2. Overview of the Core Architecture and Mathematical Foundations of Large Language Processing Models

2.1. Transformer Core Architecture

The current mainstream language models are built on the Transformer architecture, which consists of an encoder and a decoder. The core modules include word embedding layer, multi head attention layer, feedforward neural network layer, layer normalization, and residual connection module. Transformer abandons the traditional recursive structure of recurrent neural networks and achieves parallel extraction of global features of text sequences through self attention mechanism, greatly improving the efficiency of model training and inference. Each module of this architecture relies on basic mathematical theory for implementation.

2.2. Core Mathematical Fundamentals Analysis

The basic mathematical theories involved in the large language model cover multiple branches, and the core theories and corresponding application scenarios are as follows:

Linear Algebra: Vector Space, Matrix Operations, Tensor Decomposition, Eigenvalue Decomposition, Random Matrix Theory, Applied to Word Embedding, Attention Mechanisms, Parameter Matrix Operations;

Probability theory and mathematical statistics: probability distribution, Bayes' theorem, law of large numbers, central limit theorem, hidden Markov model, applied to text probability modeling, model generalization, uncertainty reasoning;

Functional analysis: Hilbert space, Banach space, linear operators, applied to semantic space representation and feature abstraction;

Optimization theory: convex optimization, non convex optimization, gradient descent, Lagrange multiplier method, applied to model training and loss function minimization;

Information theory: Information entropy, cross entropy, mutual information, KL divergence, applied to the construction of loss functions and optimization of information transmission efficiency;

Topology: Topological manifolds, topological invariants, connectivity, applied to model interpretability and logical reasoning analysis.

These mathematical theories intersect with each other and jointly support the

operation and optimization of large language models, which is the core foundation for achieving high model performance.

3. Advanced Applications of Linear Algebra in Large Language Models

Linear algebra is the most fundamental mathematical foundation of large language models, running through the entire process of model data representation, feature extraction, and parameter operations. Its advanced applications break through the limitations of traditional matrix operations and achieve a dual improvement in model efficiency and performance.

3.1. Vector Space Representation of Word Embeddings

Word embedding is the process of converting textual vocabulary into vector form that can be recognized by computers, and is the first step in large language models for processing textual data. Based on the theory of linear algebraic vector space, vocabulary is mapped to a high-dimensional real vector space $\mathbb{R}^d(d)$ as the embedding dimension, so that semantically similar vocabulary presents features of similar distance in the vector space.

The core mathematical expression of word embedding is: given a collection of words $V = w_1, w_2, \cdots, w_n$, the mapping of vocabulary to vectors is achieved through an embedding matrix $W_e \in \mathbb{R}^{d \times |V|}$, that is:

$$\boldsymbol{x}_i = W_e \cdot \boldsymbol{o}_i$$

Among them, $\boldsymbol{o}_i$ is the unique heat encoding vector of the vocabulary, and w_i is the word embedding vector of the vocabulary.

On this basis, **singular value decomposition (SVD)** is used to perform dimensionality reduction optimization on the word embedding matrix, reducing the vector dimension and computational complexity while preserving the core semantic information. The SVD decomposition formula is:

$$W_e = U \Sigma V^{\mathrm{T}}$$

Among them, U is the left singular matrix, Σ is the diagonal matrix of singular values, and V^{T} is the transpose of the right singular matrix. By retaining the matrix components corresponding to the first k largest singular values, a low rank approximation of the embedding matrix is achieved, greatly improving the computational efficiency of the model.

3.2. Matrix Operation Optimization of Self Attention Mechanism

The self attention mechanism is the core module of Transformer, which essentially calculates the correlation between different words in a text sequence through matrix operations. The mathematical expression of the standard self attention mechanism is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V$$

Among them $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ are respectively the query matrix, key matrix, and value matrix, X is the input word embedding matrix, W_Q, W_K and W_V are the parameter matrices, and d_k are the key vector dimensions.

A linear attention mechanism is proposed based on tensor decomposition theory to address the computational complexity $O(n^2)$ (n is the sequence length) of self attention mechanism. The attention matrix is decomposed into the product of low rank tensors, reducing the complexity to $O(n)$. Using kernel functions $\phi(\cdot)$ to map Q and K, the optimized attention formula is:

$$\text{LinearAttention}(Q,K,V) = \frac{\phi(Q)\phi(K)^{\mathrm{T}}V}{\phi(Q)\phi(K)^{\mathrm{T}}\mathbf{1}}$$

This method solves the computational bottleneck of long sequence text processing and achieves efficient modeling of ultra long texts through the theory of linear algebra low rank approximation.

3.3. Random Matrix Theory and Model Parameter Optimization

Random matrix theory is an advanced application of linear algebra in large language models, used to analyze the spectral distribution characteristics of weight matrices during model training, optimize model parameter initialization and training processes. Research has shown that the weight matrix spectral distribution of pre trained large language models follows the Markov Pastur distribution, with a probability density function of:

$$\rho_{MP}(x) = \frac{1}{2\pi\sigma^2 x}\sqrt{(b_+ - x)(x - b_-)}$$

Among them $b_\pm = \sigma^2\left(1 \pm \sqrt{\frac{N}{M}}\right)^2$, M and N are the number of rows and columns of the matrix, and σ is the variance.

Based on this theory, the initialization distribution of model parameters can be precisely controlled to avoid gradient explosion or vanishing in the early stages of training; At the same time, by analyzing the offset law of the matrix spectral distribution, the key parameter layers in the model training are located, and the fine tuning of the model parameters is achieved, greatly improving the training efficiency and model convergence performance.

3.4. Linear Algebraic Principles of Residual Connections and Layer Normalization

The residual connection and layer normalization module in Transformer is implemented based on linear algebraic linear operators and norm theory. Residual connections solve the gradient vanishing problem in deep models through identity mapping, expressed mathematically as:

$$\boldsymbol{y} = \text{LayerNorm}(X + \text{FFN}(X))$$

Layer normalization ensures the stability of model training by normalizing the feature vectors, and its formula is:

$$\text{LayerNorm}(x) = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta$$

Among them μ and σ^2 are respectively the mean and variance of the feature vectors, γ and β are learnable parameters, and ε are the minimum values to prevent the denominator from being zero. The linear superposition of residual connections and the vector normalization of layer normalization jointly ensure the stable training of deep large language models.

4. Advanced Applications of Probability Theory and Mathematical Statistics in Large Language Models

Probability theory and mathematical statistics provide theoretical support for uncertainty modeling, generalization ability improvement, and text generation optimization of large language models. Its core is to characterize the statistical rules of text data through probability distribution, and achieve accurate understanding and generation of language by the model.

4.1. Probability Modeling of Text Sequences

The core task of the large language model is to predict the probability distribution of the next vocabulary. Based on Markov chain and joint probability distribution theory, the probability of a text sequence can be decomposed into the product of conditional probabilities:

$$P(w_1, w_2, \cdots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \cdots, w_{i-1})$$

The model achieves coherent text generation by learning the conditional probability distribution of massive text data. On this basis, the probability features of vocabulary output in polynomial distribution modeling are introduced, combined with Laplace smoothing to solve the problem of data sparsity. The smoothing formula is:

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i) + \alpha}{C(w_{i-1}) + \alpha |V|}$$

Among them, $C(\cdot)$ is the co-occurrence frequency of vocabulary, α is the smoothing coefficient, and $|V|$ is the size of the vocabulary list, effectively avoiding the zero probability problem and improving the model's generalization ability.

The native training of Transformer-LLM does not utilize HMM-style sequence modeling and Laplace smoothing. This section merely provides a traceability and comparative explanation of probability theory.

4.2. Integration of Bayesian Theory and Attention Mechanism

The traditional attention mechanism is deterministic computation and cannot model the uncertainty of semantic associations. Based on Bayes' theorem, a Bayes-

ian attention model is constructed, which values attention weights as random variables and introduces prior and posterior distributions to optimize weight calculation.

The core formula of Bayesian attention is:

$$P(A \mid Q,K,V) = \frac{P(Q,K,V \mid A)P(A)}{P(Q,K,V)}$$

Among them, A is the attention weight matrix, $P(A)$ is the prior distribution of weights, and $P(Q,K,V \mid A)$ is the likelihood function. By using variational inference to approximate the posterior distribution, probabilistic modeling of attention weights is achieved to improve the robustness of the model in small sample and noisy data scenarios.

4.3. Hidden Markov Models and Sequence Optimization

Hidden Markov Model (HMM) is used to optimize the generation and error correction of text sequences in large language models. Based on the transition probability and observation probability of hidden states, a hidden state representation of text sequences is constructed. The three core formulas of HMM are:

Initial state probability: $\pi_i = P(q_1 = i)$

State transition probability: $a_{ij} = P(q_{t+1} = j \mid q_t = i)$

Observation probability: $b_j(k) = P(o_t = k \mid q_t = j)$

By using the Viterbi algorithm to solve the optimal hidden state sequence, we can achieve syntax correction and logic optimization in text generation, thereby improving the fluency and accuracy of the output text.

Mainstream LLMs do not directly employ HMMs for sequence modeling, and they are only introduced as comparison or enhancement modules in some lightweight and interpretability improvement studies.

4.4. Law of Large Numbers and Analysis of Model Generalization Error

Based on the law of large numbers and empirical risk minimization theory, analyze the generalization error of large language models and construct a generalization performance evaluation framework. The model generalization error E satisfies:

$$E = \left| R(f) - R_{emp}(f) \right|$$

where $R(f)$ represents expected risk and $R_{emp}(f)$ empirical risk. According to the Hofdin inequality, the upper bound of the generalization error can be obtained:

$$P\left(\left| R(f) - R_{emp}(f) \right| \geq \epsilon \right) \leq 2\exp\left(-2n\epsilon^2\right)$$

where n is the number of training samples. This theory provides a mathematical basis for selecting the sample size for model training and optimizing generalization performance, guiding the model to achieve a balance between data volume

and performance.

5. Advanced Applications of Functional Analysis and Optimization Theory in Large Language Models

5.1. Functional Analysis: Hilbert Space Representation of Semantic Space

The Hilbert space in functional analysis provides a comprehensive mathematical framework for abstract representation of language semantics, constructing word embedding vector spaces as complete inner product spaces $\left(\mathcal{H},\langle\cdot,\cdot\rangle\right)$, and quantifying semantic similarity through inner product operations.

$$\mathrm{sim}\left(\boldsymbol{x},\boldsymbol{y}\right)=\frac{\langle\boldsymbol{x},\boldsymbol{y}\rangle}{|\boldsymbol{x}||\boldsymbol{y}|}$$

Among them, $\langle\boldsymbol{x},\boldsymbol{y}\rangle$ is the vector inner product and $|\cdot|$ is the vector norm.

Based on the theory of linear operators, the attention layer and feedforward layer of Transformer are regarded as linear transformation operators in Hilbert space $T:\mathcal{H}\to\mathcal{H}$, and the forward propagation process of the model is the iterative abstraction of semantic features by linear operators, achieving the extraction from shallow vocabulary features to deep semantic features.

5.2. Convex Optimization and Non Convex Optimization: The Core Theory of Model Training

The essence of training large language models is to solve the problem of minimizing the loss function, which belongs to the core application scenario of optimization theory. The objective function for model training is:

$$\min_{\theta}\mathcal{L}\left(\theta\right)=\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(f_{\theta}\left(x\right),y\right)\right]$$

Among them θ are model parameters, $\mathcal{D}$ data distribution, $\ell\left(\cdot\right)$ loss function, and $f_{\theta}\left(x\right)$ model prediction output.

5.2.1. Gradient Descent Optimization Algorithm

Based on gradient descent theory, the model parameters are iteratively updated to approximate the optimal solution, and the parameter update formula is:

$$\theta_{t+1}=\theta_{t}-\eta\cdot\nabla\mathcal{L}\left(\theta_{t}\right)$$

where η is the learning rate and $\nabla\mathcal{L}\left(\theta_{t}\right)$ is the gradient of the loss function. Aiming at the local optimal problem of non convex optimization in large language models, momentum based stochastic gradient descent (SGD-M), Adam and other adaptive optimization algorithms are proposed to improve convergence speed by accumulating gradient directions and avoid falling into local optima.

5.2.2. Lagrange Multiplier Method and Constrained Optimization

To address the issues of parameter constraints and computational resource constraints in model training, a constraint optimization model is constructed based on the Lagrange multiplier method, and the Lagrange function is introduced.

$$\mathcal{L}(\theta,\lambda)=\mathcal{L}_{orig}(\theta)+\lambda\cdot g(\theta)$$

Among them, $\mathcal{L}_{orig}(\theta)$ is the original loss function, $g(\theta)$ is the constraint condition, and λ is the Lagrange multiplier. By solving the extremum of the Lagrange function, a balanced optimization of model performance and resource consumption can be achieved.

5.3. Dual Theory and Model Lightweighting

Based on the dual theory of convex optimization, the original optimization problem of the large language model is transformed into a dual problem, achieving sparsity and lightweighting of model parameters. By using KKT (Karush Kuhn Tucker) conditions to solve the optimal solution, key parameters in the model are screened, redundant parameters are removed, and the number of parameters and computation is significantly reduced while ensuring model performance, making it suitable for low resource scenarios such as mobile and edge devices.

6. Advanced Applications of Information Theory and Topology in Large Language Models

6.1. Information Theory: Model Information Transmission and Efficiency Optimization

6.1.1. Cross Entropy Loss Function

The cross entropy in information theory is the core loss function of large language models, used to measure the difference between the predicted distribution and the true distribution of the model. The formula is:

$$\mathcal{H}(p,q)=-\sum_{i=1}^{n}p(x_i)\log q(x_i)$$

where p is the true probability distribution and q is the model predicted distribution. Cross entropy loss can directly guide model parameter updates and accelerate model convergence speed.

6.1.2. KL Divergence and Distribution Alignment

KL divergence (relative entropy) is used to quantify the difference between two probability distributions, aligning the model distribution with the true data distribution. The formula is:

$$D_{KL}(p\,\|\,q)=-\sum_{i=1}^{n}p(x_i)\log\frac{q(x_i)}{p(x_i)}$$

In model distillation and domain adaptation tasks, the transfer of knowledge from large models to small models is achieved by minimizing KL divergence, thereby improving the performance of small models.

6.1.3. Mutual Information and Feature Filtering

Based on the theory of mutual information, the core semantic features in the text sequence are screened to remove redundant information. The formula for mutual

information is:

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

By maximizing the mutual information between input features and output labels, the accuracy of semantic feature extraction in the model is improved, and the efficiency of model information transmission is optimized.

6.2. Topology: Analysis of Model Interpretability and Reasoning Ability

6.2.1. Topological Manifold and Semantic Space Analysis

The semantic feature space of the large language model can be viewed as a topological manifold, where the semantic representations of vocabulary and sentences are distributed in different regions of the manifold, and semantic associations correspond to the connected paths on the manifold. Based on the theory of topological manifolds, the semantic space structure of the model can be visualized, revealing the inherent logic of the model's understanding of language and providing a new perspective for the interpretability of the model.

6.2.2. Topological Invariants and Inference Ability Evaluation

A quantitative evaluation system for the reasoning ability of large language models based on topological invariants is constructed, which quantifies the logical reasoning, mathematical reasoning, and common sense reasoning levels of the model by calculating topological invariants such as the Betti number and Euler characteristic number in the semantic feature space. Research has shown that the reasoning ability of models is positively correlated with the topological connectivity of semantic space, providing a mathematical basis for optimizing the reasoning ability of models.

7. Research Bottlenecks and Future Prospects

7.1. Current Research Bottlenecks

Insufficient depth of integration between mathematical theories and models: Existing research mostly focuses on the engineering application of mathematical theories, lacking innovation in mathematical theories targeting the characteristics of large language models, and the application of complex mathematical theories such as algebraic topology and differential geometry is not yet mature;

Difficulty in solving high-dimensional mathematical problems: The high-dimensional parameter space and high-dimensional semantic space of large language models result in extremely high computational complexity for mathematical analysis and optimization, making it difficult for existing mathematical tools to achieve fine-grained analysis;

The interpretability mathematical framework is incomplete: interpretability research based on topology and functional analysis is still in its early stages and cannot fully characterize the mathematical reasoning logic of models from input to

output;

The disconnect between theory and engineering implementation: Some cutting-edge mathematical theories are difficult to adapt to the engineering training and deployment of large language models, resulting in high computational costs and inability to achieve large-scale applications.

Meanwhile, mathematical methods face multiple inherent trade-off constraints in practical implementation: Firstly, linearization methods such as linear attention reduce complexity from $O(n^2)$ to $O(n)$ to improve efficiency, but they lose high-order feature interaction information, leading to an inherent contradiction between efficiency and representational accuracy. Secondly, compression techniques such as low-rank decomposition and tensor decomposition reduce the number of parameters and computational complexity, but they can cause loss of semantic information and decrease in model expressive power, forming a direct trade-off between compression ratio and model performance. Thirdly, interpretability methods based on topological analysis, SHAP, and LIME can enhance model transparency, but their computational complexity increases exponentially with the size of the model, posing a difficult trade-off between interpretability strength and inference computation cost. Fourthly, theories such as Bayesian inference and probabilistic modeling can improve model generalization and robustness, but they significantly increase training and inference overhead, further exacerbating the coupled constraints of accuracy, efficiency, and cost. These mathematical trade-offs directly determine the design choices of model architecture, training strategies, and deployment schemes, and are also the core practical bottlenecks that current basic mathematical applications struggle to overcome.

7.2. Future Research Prospects

Building an exclusive mathematical theoretical system: targeting the high-dimensional, non convex, and large-scale characteristics of large language models, innovatively integrating multiple branches of basic mathematical theories, and constructing an exclusive mathematical analysis and optimization framework for large language models;

Lightweight application of complex mathematical theories: promoting the lightweight implementation of complex mathematical theories such as algebraic topology, differential geometry, and stochastic analysis in models, developing low complexity mathematical algorithms, and solving problems in model interpretability and inference optimization;

Mathematics driven model architecture innovation: Based on fundamental mathematical theory, reconstruct the attention mechanism, training paradigm, and decoding algorithm of the model, and break through the performance bottleneck of the existing Transformer architecture from a mathematical perspective;

Interdisciplinary research: Strengthen interdisciplinary cooperation between mathematics, computer science, and linguistics, explore the mathematical laws of language itself, and achieve accurate mathematical modeling of language by mod-

els;

Efficient mathematical optimization algorithm development: Targeting the computational bottleneck of large model training and inference, we develop efficient algorithms based on optimization theory and linear algebra to reduce model training costs and improve inference efficiency.

8. Conclusions

As the theoretical core of large language models, basic mathematics plays an irreplaceable role in key areas such as model word embedding, attention mechanisms, training optimization, text generation, interpretability analysis, and other branches such as linear algebra, probability theory and mathematical statistics, functional analysis, optimization theory, information theory, and topology. Research and practice in the past five years have shown that the deep application of basic mathematical theories is the core driving force for improving the performance and theoretical perfection of large language models.

At present, the integration of basic mathematics and large language models still faces many bottlenecks. In the future, it is necessary to further deepen the innovation and application of mathematical theory, and solve core problems such as training efficiency, reasoning ability, and interpretability of large language models from a mathematical perspective. With the continuous deepening of interdisciplinary research between basic mathematics and artificial intelligence, mathematical theory will continue to provide support for technological breakthroughs in large language models, promote the development of large language models towards more universal, intelligent, and controllable directions, and lay a solid foundation for the comprehensive progress of the field of artificial intelligence.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Pakray, P., Gelbukh, A. and Bandyopadhyay, S. (2025) Natural Language Processing Applications for Low-Resource Languages. *Natural Language Processing*, **31**, 183-197. https://doi.org/10.1017/nlp.2024.33

[2] Nastase, V., Mihalcea, R. and Radev, D.R. (2015) A Survey of Graphs in Natural Language Processing. *Natural Language Engineering*, **21**, 665-698. https://doi.org/10.1017/s1351324915000340

[3] Zampieri, M., Rosenthal, S., Nakov, P., Dmonte, A. and Ranasinghe, T. (2023) Offenseval 2023: Offensive Language Identification in the Age of Large Language Models. *Natural Language Engineering*, **29**, 1416-1435. https://doi.org/10.1017/s1351324923000517

[4] Chen, F., Zhang, Z., Jia, Y., *et al.* (2025) Research on the Similarity Calculation of Short Text in the Terminology Domain Based on Siamese BERT Model. *Scientific Reports*, **15**, Article No. 36954. https://doi.org/10.1038/s41598-025-20908-8

[5] Karttunen, L., Koskenniemi, K. and Van Noord, G. (2003) Finite State Methods in

Natural Language Processing. *Natural Language Engineering*, **9**, 1-3. https://doi.org/10.1017/s1351324903003139

[6] Garcia Quevedo, D. and Kuri, J. (2026) Overview of Artificial Intelligence, Machine Learning, Natural Language Processing, and Large Language Models. In: Quevedo, D.G. and Kuri, J., Eds., *AI for Qualitative Research*, Springer, 7-21. https://doi.org/10.1007/978-3-032-08872-7_2

[7] Xiong, S., Pan, L., Ma, X., *et al.* (2024) Unsupervised Deep Hashing with Multiple Similarity Preservation for Cross-Modal Image-Text Retrieval. *International Journal of Machine Learning and Cybernetics*, **15**, 4423-4434. https://doi.org/10.1007/s13042-024-02154-y

[8] Mysior, M. and Cavallucci, D. (2026) Contradiction Processing Using Large Language Models and Generative Artificial Intelligence. In: Cavallucci, D., Brad, S., Livotov, P. and Houssin, R., Eds., *IFIP Advances in Information and Communication Technology*, Springer, 171-183. https://doi.org/10.1007/978-3-032-08847-5_12

[9] Trummer, I. (2025) Generating Highly Customizable Python Code for Data Processing with Large Language Models. *The VLDB Journal*, **34**, Article No. 21. https://doi.org/10.1007/s00778-025-00900-4

[10] Kaye, N.G. and Gordon, P.C. (2025) Sentence Processing by Humans and Machines: Large Language Models as a Tool to Better Understand Human Reading. *Psychonomic Bulletin & Review*, **32**, 2719-2733. https://doi.org/10.3758/s13423-025-02756-9

[11] Nguyen, D., Nguyen, M., Chu, Q., Luu, S.T., Chu, N., Vo, T., *et al.* (2026) Enhancing Legal Text Processing and Structural Analysis with Large Language Models at COLIEE 2025. *The Review of Socionetwork Strategies*, **20**, 361-383. https://doi.org/10.1007/s12626-026-00211-2