Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

# Towards Efficient LLMs: Analyzing Computational Bottlenecks and Optimization Strategies

Taowen Qian

Faculty of Arts and Science, University of Toronto, Toronto, Canada

taowen.qian@mail.utoronto.ca

**Abstract.** The study major focuses on the efficiency of the current Large Language Model (LLMs). By researching several papers that focus on it, the limitations of the current efficiency in LLM are significant problems that need to be considered by academia. Then, the study will provide some research on the progress of solving issues and explain each solution clearly. Finally, the study will focus on the further needs for developing each solution. This study is conducted on the USER-LLM, OPTIMA, and Infinite-LLM systems that can solve the efficiency problems in LLM and find some benefits in improving LLM efficiency limitations. Experimental results show that some issues in each system need to be solved in further research. This study can explain the main efficiency problems in current LLMs and provide direction for further research. With more research on the efficiency problem, computational costs and response times will decrease, enabling real-time decision-making improvement.

**Keywords:** Large Language Model (LLMs); efficiency; USER-LLM; OPTIMA; Infinite-LLM.

## 1. Introduction

With the rapid advancement of technology, the high techniques in science movies and science fiction are one after another. The most significant are intelligent personal assistants (IPAs), which can deal with complex assignments and have emotions to communicate. Those IPAs are the imagination of artificial intelligence (AI) from human thinking. Along with mobile phones and other portable technological innovations, IPAs such as Siri, Google Assistant, and Alexa are making tremendous progress [1-3]. Those IPAs can help people control their devices more efficiently and handle simple communication tasks by swiftly executing voice commands and enhancing overall productivity. However, the IPAs in the current system are suffering from limitations in several areas, such as understanding of the assignment, efficiency, and correction.

However, with the Large Language Model (LLMs) occurring in the current IPAs' development, the IPAs seem not to be as fantasized as before. The primary example is Chat Generative Pre-Trained Transformer (ChatGPT), a language model that generates human-like text responses based on user prompts. It is primarily used to answer questions, draft content, and assist in various conversational tasks. It also supports multiple applications such as customer service, creative writing, and educational tutoring by facilitating efficient and engaging interactions. AI development is already being mentioned significantly in human life. Many people are researching the accuracy of the AI, such as focusing on the self-correcting LLM-controlled diffusion models, training a free LLM-based approach to general Chinese character error correction, and automatically correcting LLMs [4-6].

However, LLM today not only needs to care about the correction but also needs to pay more attention to the code efficiency. Optimizing code is essential in practical systems, as it enhances processing speed, minimizes delays in algorithm execution, and lowers overall energy usage. Although this is important to focus on, few essays mention this point. Until recently, a high-standard benchmark has been proposed to evaluate the efficiency of LLM-generated code [7]. Furthermore, some different ways are already being used to enhance LLM-generated efficiency. Using user embeddings to efficient LLM contextualization, optimizing effectiveness and efficiency for LLM-Based multi-agent system, and efficient LLM service for long context with Decoupling the attention layers (DistAttention) and distributed key-value cache (KVCache) [8-10]. According to several essays, there

are many limitations in the effectiveness of LLM, and the focus of this essay is to talk about the efficacy of the current LLM and try to find a solution to deal with such limitations.

## 2. Methodology

The most important part of this essay is the effectiveness of the LLM. This essay will first consider the main limitations of effectiveness and explain these limitations and their reasons in the current research. Then, this article will propose some practical solutions based on the current study and try to give the advantages and disadvantages of these solutions. After that, this article will give a simple conclusion on the effectiveness of the LLM in the future and provide the focus points for the development of future effectiveness.

The main limitation of the current LLM model is the connection between input and output. It can be divided into three main parts, which this essay will focus on. First, traditional LLM uses the static model that cannot flexibly adapt to different context lengths, which means that the resource cannot be fully utilized when a short context occurs. For the long context, the memory demand for the Graphics Processing Unit (GPU) is too large to handle. The second part is that the long user history will lead to unnecessary computation during inference. The final part is that when multiple language model agents interact in multi-agent systems, the inference slows down a lot, making the message redundant. In the current research, the major solution is to develop new improvement algorithms focusing on each part to enhance their effectiveness.

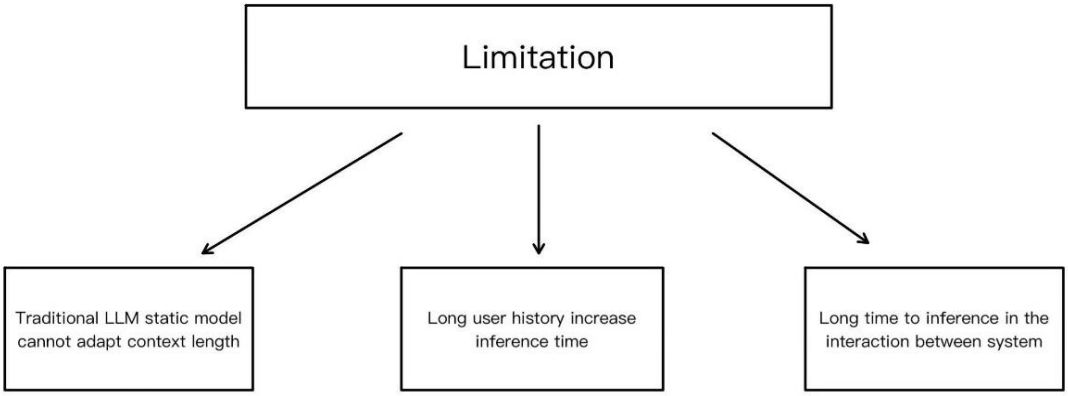


**Fig. 1** The Main Limitations (Picture credit: Original).

### 2.1. Introduction of LLMs System

First, consider the different context length problems. Because in the traditional system, the models are statically allocated the GPUs per instance, short context will lead to underutilization of compute resources. In contrast, for the long input, a single instance may run out of memory since the KVCache of the attention layer grows with the sequence length. The other research provides the Infinite-LLM system (shown in Fig. 1) to deal with these problems [10]. DistAttention is the main thought in the infinite-LLM system. The primary reason for increasing memory usage is attention layers. However, DistAttention only needs static memory usage. In this way, the model can divide the attention of calculations used to travel the whole KVCache into lots of small parts, and these small parts can be calculated in parallel on multiple GPUs, which significantly decreases the transfer of large amounts of data and becomes more flexible for longer inputs. And the detail change is that when the KVCache is divided into two parts, the debtors, which handle long context requests and need more memory usage, handle short context requests and have some remaining memory usage. Then, the system can dynamically allocate these small parts to the GPUs appropriately using the greedy algorithm. Specifically, this solution provides the cluster-level scheduling algorithm, which introduces a gManager to maintain the status of each small part and an rManager to report on the local memory usage. Finally, by the gManager and rManager and greedy algorithm, the model can dynamically allocate the debtors and creditors to different GPUs and enhance the efficiency of the whole model.

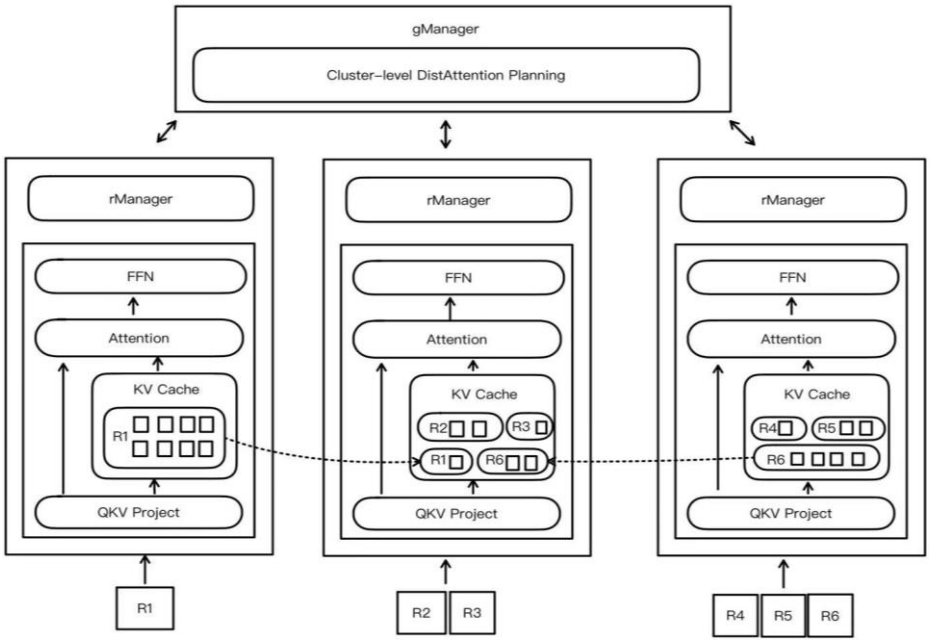


**Fig. 2** Infinite-LLM System Overview [10].

The second problem is low efficiency due to the high noise of the user history. Because the traditional model needs to remember the whole user history to text prompt, it will require considerable calculation resources and may lose detail in the calculation. The research provides the USER-LLM system (shown in Fig. 2) to handle the redundant user history [8]. The first step in accomplishing this system is to represent the user data. Each user history can comprise several features (name, score, etc.), which can be mapped to the different integer ID numbers and vectorized to be embedded via dedicated embedding layers. Then, the sequence of fused embeddings is fed into an Autoregressive Transformer to capture temporal dynamics and contextual relationships in the user history. And the system uses a projection layer to reduce the dimension of the output, a sequence of dense user embeddings, which can reduce the complexity of the model again. Without the redundant text prompt, the USER-LLM system can let the model dynamically adjust output using a cross-attention mechanism combined with the history embedding. Finally, USER-LLM can significantly improve the inference speed by converting user timelines into compact, informative embeddings and integrating them via cross-attention.

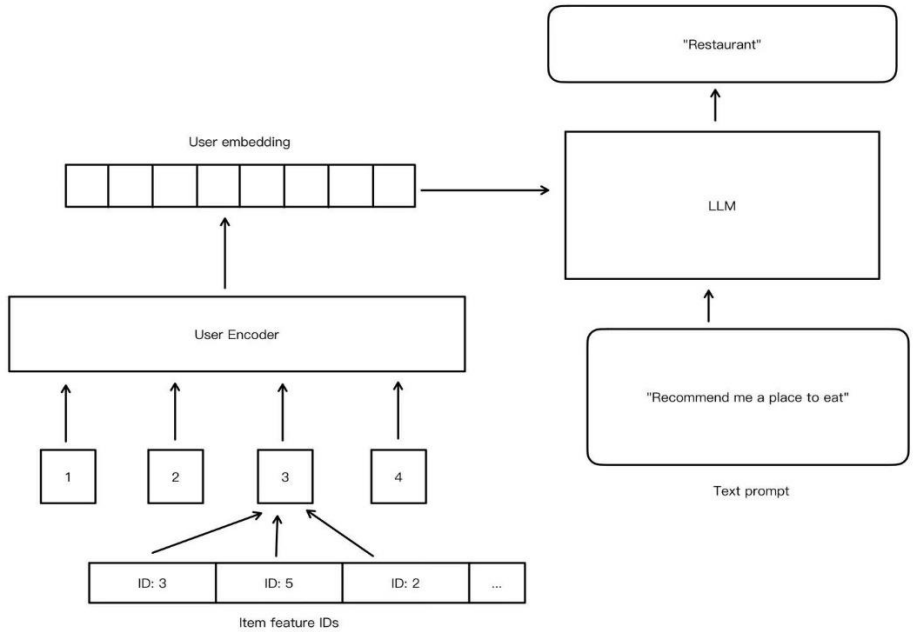


**Fig. 3** USER-LLM System Overview [8].

The final problem is the inefficiency of the inter-agent communication of LLM-based multi-agent systems (MAS), because of the duplicate and redundant communication (shown in Fig. 3). The research provides the OPTIMA structure to solve this problem [9], and the primary thought is to increase communication and output by iterative training. Because the whole OPTIMA structure is too complex to explain, this essay only focuses on some essential parts. Because the main thought is iteration training, the structure has a reward formula based on the communication quality, the token used, and the language model loss. By this formula, the structure can ensure the model obtains a high reward when the communication becomes efficient, concise, and natural. Then, the OPTIMA structure only maintains the best communication strategy for the training part. After training the strategy, the model can reduce the unnecessary tokens used and enhance communication efficiency, and the model inference speed in the MAS will also increase significantly.

### 2.2. Applications Analysis

The different strategies can increase the efficiency of LLM from various perspectives. The OPTIMA system can be used in customer service and business decision support, because multi-agent systems can collaborate to handle complex customer queries and provide coordinated decision support in call centers or enterprise help desks. In addition, the Infinite-LLM can be used for financial analysis and academic research, since this system can deal with redundant financial reports and a massive research essay with high efficiency. Also, the USER-LLM system can improve personalized recommendations and customized digital assistants, because the system has high personalization by embedding features from historical data.

As the world develops, the LLM model needs to be improved not only in the correction part but also in the efficiency part. Only efficient models can be used in the real world and combined with other systems requiring real-time response (such as medical or fire protection systems, etc.). Therefore, even with the several models mentioned above, the progress of LLM in efficiency is insufficient and should be paid more attention to.

## 3. Results and Discussion

Then this paper will consider the benefits and disadvantages of each model and give some direction for further development. The OPTIMA system can significantly increase the efficiency of each model, and this system can divide problems into multiple parts to enhance the inference speed. However, this system needs to use a complex, sophisticated communication protocol and reward function, making the system more complicated and leading to more challenges in the improvement process. In addition, this system also needs to use more calculation resources. If the system needs to be improved, researchers should focus on the following parts. First, to improve the communication protocols, researchers need to refine inter-agent coordination to reduce redundancy in the communication further. Then, the hybrid training method may also improve the system, so the researchers should explore more hybrid training methods, such as integrating Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) to balance quality and efficiency. The Infinite-LLM system can solve the extremely long document mentioned in this paper, and the system can improve throughput by high-efficiency resource utilization. For the disadvantage, the system only has limited benefits on long documents and cannot benefit tasks involving shorter text. Also, the system relies on dynamic scheduling and is very complex for resource management, requiring advanced infrastructure to exploit efficiency fully.

This paper will provide some advice for each system limitation. First, researchers can develop and optimize adaptive scheduling algorithms to better allocate GPU resources with different context sizes. In addition, the enhancement in compatibility with various cloud infrastructures can also deal with complex management problems. The last system, USER-LLM, can use cross-attention to embed user information and allow the system to provide personalized recommendations and responses. The system can also decrease the inference time by embedding user histories. Nevertheless, this system

highly depends on the quality of user data; if the data is sparse or poor, the recommendations may be worse. Also, privacy and security are the main problems for the user. Handing sensitive user data requires strong privacy measures to support, which may cause the system to be complex to implement. The researchers should focus on developing more robust user encoders that can extract high-quality embeddings from noisy user data, and they also need to implement advanced data protection protocols to guarantee sensitive user information complies with privacy regulations.

## 4. Conclusion

This study introduces the efficiency limitations in the current LLM model and provides some current research to deal with those limitations. This paper proposes to analyze three systems, USER-LLM, Infinite-LLM, and OPTIMA, that can increase the efficiency of the LLM model. Extensive experiments are conducted to evaluate the proposed method. Experimental results show that this system can significantly improve the efficiency of LLM in certain areas. Still, they also have some problems that must be solved in the following research. In the future, A will be considered as the research objective for the next stage. The study will focus on refining the inter-agent coordination of the OPTIMA system, adaptation scheduling algorithms of the Infinite-LLM system, and the user encoder of the USER-LLM system.

## References


[1] BEAUDETTE de, OGEEN at. An iPhone application for on-demand access to digital soil survey information. Soil Science Society of America Journal, 2010, 74(5): 1682-1684.

[2] LAZIC Aleksandar, et al. Google Assistant integration in TV application for Android OS. Telecommunications Forum (TELFOR). 2018, 420-425.

[3] MATHEW, S., & VARIA, J. Overview of amazon web services. Amazon Whitepapers, 2014, 105(1), 22.

[4] WU TSUNG Han, et al. Self-correcting llm-controlled diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 6327-6336.

[5] ZHOU Hou Quan, et al. A Training-free LLM-based Approach to General Chinese Character Error Correction, arXiv preprint arXiv:2502.15266, 2025.

[6] PAN Liang Ming, et al. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. Transactions of the Association for Computational Linguistics, 2024, 12: 484-506.

[7] QIU Rui Zhong, et al. How Efficient is LLM-Generated Code? A Rigorous & High-Standard Benchmark, arXiv preprint arXiv:2406.06647, 2024.

[8] NING Lin, et al. User-llm: Efficient llm contextualization with user embeddings, arXiv preprint arXiv:2402.13598, 2024.

[9] LI Xin Yi, et al. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth, 2024, 1(1): 9.

[10] LIN Bin, et al. Infinite-LLM: Efficient LLM Service for Long Context with DistAttention and Distributed KVCache, arXiv preprint arXiv:2401.02669, 2024.