Journal of Artificial Intelligence Practice (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2024.070306
ISSN 2371-8412 Vol. 7 Num. 3

The Training Process and Methods for LLMs Using an Own Knowledge Base

Sheng Zhiyuan

Wuhan Foreign Languages School, Wuhan, Hubei, China
cfsheng@126.com

Keywords: Large Language Models, LLMs, Own Data, Own Knowledge Base, Pre-Training, Fine-Tuning, Performance Evaluation, NLP, CV

Abstract: This paper explores the development of frameworks and training methods for large language models (LLMs), focusing on the importance of self-built data (Own data or Own Knowledge Base), specific processes of model pre-training and fine-tuning, and model performance evaluation and deployment effects. By introducing and analysing the advantages and disadvantages of mainstream large language models (such as GPT-4, BERT, LLaMA, and Mistral), we illustrate the strengths and limitations of large language models in natural language processing tasks. This paper particularly emphasises the critical role of self-built data in enhancing the model's professionalism and accuracy, discussing data collection and processing methods. We detail the steps of model pre-training and their impact on model performance, explore the necessity and implementation of model fine-tuning, and validate the effectiveness of the proposed framework training method through performance evaluation metrics and actual deployment effects.

1. Introduction

1.1 Research Background

Large Language Models (LLMs) have made significant progress in the fields of Natural Language Processing (NLP) and Computer Vision (CV) in recent years and have demonstrated strong capabilities in various applications. These models, such as GPT-4 and BERT, typically undergo pre-training on massive text data to master language's complex structures and semantic relationships[1]. This pre-training process enables the models to generate high-quality text, answer questions, translate languages, summarise text, and analyse sentiments. As Llama3 and Mistral 8x22B provide capabilities to solve production-level tasks, the framework or products for LLM pre-training, fine-tuning, evaluation, and deployment based on their knowledge base are rapidly emerging. Personalised AI applications are on the horizon, as anyone can fine-tune models with their application data.

1.2 Research Purpose

This research aims to propose and explore a training process and method for large language models based on its knowledge base to address the various challenges in applying large language models in

specific domains. Firstly, by designing and implementing an efficient data collection and processing workflow, we ensure the quality and relevance of our knowledge base, enhancing the model's performance in specific tasks. Secondly, we investigate how to optimise the model training process, including hyperparameter tuning and training architecture design, to maximise model performance under limited resources. Finally, we explore the training process that is suitable for large language models. We utilise advanced inference optimisation techniques (such as model quantisation and knowledge distillation) to improve the model's inference speed and resource utilisation efficiency.

2. Literature Review

2.1 Popular Large Language Models

The popular LLMs, such as GPT-4, BERT, LLaMA, and Mistral, represent cutting-edge technology in NLP and CV. These models, pre-trained on vast amounts of text data, have mastered the language's complex structures and semantic relationships[2], demonstrating excellent language understanding and generation capabilities.

- GPT-4 (Generative Pre-trained Transformer 4th edition), developed by OpenAI, is the latest generation of language models based on the transformer architecture, learning from massive internet text through unsupervised learning to generate high-quality, coherent natural language text. GPT-4 excels in various application scenarios, including text generation, translation, Q&A systems, and chatbots[3].
- In contrast, BERT (Bidirectional Encoder Representations from Transformers), proposed by Google, innovates its bidirectional encoder architecture, considering contextual information during pre-training to better understand the relationships between words and phrases in sentences[4].
- BERT has achieved leading performance in various NLP tasks, such as Q&A, sentiment analysis, and text classification. LLaMA, developed by Meta AI (formerly Facebook AI), is a series of large-scale pre-trained language models designed for text generation and understanding, available in multiple sizes (e.g., LLaMA-7B, LLaMA-13B) to suit different resource and task requirements.
- Mistral, an open-source large language model, focuses on efficient text generation and understanding, offering improved performance and efficiency through optimised training processes and architecture design, supporting various tasks, including text generation, Q&A, and translation.

These large language models have several advantages. First, they capture deep linguistic features through large-scale pre-training, achieving superior performance in various downstream tasks. GPT-4's generation capability is powerful, generating coherent and contextually relevant long texts, and it is widely used in automatic writing, content creation, and virtual assistants. BERT excels in understanding and processing complex relationships between sentences, with its bidirectional encoder architecture making it advantageous in tasks requiring precise context understanding [4]. LLaMA offers models of different scales, allowing users to choose the appropriate model based on computing resources. Mistral shows significant improvements in performance and efficiency due to optimised training and architecture design.

However, these large language models also have notable drawbacks and challenges. The first is resource consumption. Training and running these models require substantial computing resources, especially GPUs or TPUs, resulting in high costs and environmental impacts. Additionally, larger models take longer to train and infer, potentially failing to meet real-time application needs. Although these models learn linguistic features from extensive data during pre-training, they need to fully understand the background and logic of this data, sometimes generating inaccurate or inappropriate content. For instance, GPT-4 may produce factually incorrect content or language that includes bias and discrimination, posing risks and issues in practical applications.

Another significant challenge is data privacy and security. Large language models may memorise sensitive information from the training data, leading to potential data leakage risks. Moreover, without proper security measures during application, models may be used to generate malicious content or conduct social engineering attacks. Developers must adopt stringent privacy protection and security measures during model training and deployment to prevent data misuse and leakage.

Overall, mainstream large language models like GPT-4, BERT, LLaMA, and Mistral demonstrate solid capabilities and broad application prospects in NLP and CV fields but also face challenges related to resource consumption, content accuracy, and data security. Addressing these issues while leveraging the models' strengths is crucial for current and future research.

2.2 The Significance of Own Knowledge Base

Built Own Knowledge Base is profoundly significant in the training and applying of large language models, crucial for enhancing model performance and application value. First, self-built data can be highly customised for specific fields or tasks, ensuring the relevance and high quality of the data. For example, in the medical field, self-built data can include much medical literature, case reports, and clinical notes, enabling the model to understand better and generate medical-related content, significantly improving the model's performance in medical consultations, diagnostic assistance and other applications. Second, self-built data can effectively address the biases and imbalances present in public datasets. Purposeful data collection and annotation can reduce the model's bias when dealing with specific populations or topics, enhancing the model's fairness and reliability. For instance, collecting and organising market data in the financial industry allows the model to perform financial analysis and predictions accurately. Furthermore, self-built data has significant advantages in data privacy and security. Many industries (such as law, insurance, etc.) have strict requirements for data privacy and security, and self-built data can be collected and used under privacy protection regulations, reducing data leakage and legal risks. Data security and compliance can be ensured by controlling the data sources and usage processes.

In summary, self-built data enhances the model's performance in specific fields and ensures data usage security and compliance, providing a solid foundation for the practical application of large language models.

3. Training Process and Methods

3.1 Own Knowledge Base Collection and Processing

Collecting and processing the knowledge base are critical steps to ensure the efficient and accurate performance of large language models in specific fields. First, the data collection stage requires transparent data sources and screening criteria to ensure the quality and relevance of the collected data. For example, relevant information can be collected from medical literature, case reports, doctors' notes, and clinical trial data in the medical field. This data must be rigorously screened to exclude irrelevant or low-quality content, ensuring the model can learn helpful information during training. Second, data processing is complex and crucial, including cleaning, annotation, and enhancement. Data cleaning removes noise, fills in missing values, and unifies data formats, making the data more consistent and reliable. Data annotation involves adding clear labels to the data through manual or semi-automated methods, enabling the model to understand and learn specific features and patterns. Data enhancement involves generating additional training samples, such as data expansion, synthesis, or translation, to increase the diversity and coverage of the data, thereby improving the model's generalisation ability. Moreover, encryption and anonymisation techniques should be employed during data processing to protect sensitive information and ensure data privacy and security. Through

a well-designed data collection and processing workflow, high-quality, knowledge-based professional datasets can be constructed, significantly enhancing the performance and reliability of large language models in specific application scenarios.

3.2 Pre-training

Model pre-training is crucial in developing large language models (LLMs). It involves the initial training of the model on large-scale text data to capture broad linguistic features and semantic structures. This process is typically conducted in an unsupervised learning framework, where the model learns patterns and relationships in the data through self-supervision without requiring manual annotations[5]. The primary goal of pre-training is to equip the model with foundational language knowledge, enabling it to be more effectively fine-tuned for specific tasks in subsequent stages.

Two main training strategies are commonly employed during pre-training: autoregressive models (such as the GPT series) and autoencoding models (such as BERT). Autoregressive models are trained by predicting the next word in a sequence, given the preceding words. For example, in GPT-4's training, the model processes large amounts of internet text, learning to generate coherent and contextually relevant natural language text. Autoencoding models, on the other hand, use a masked language model approach during pre-training. For instance, BERT randomly masks portions of the sentence and trains the model to predict the masked words. This method allows the model to consider both left and right context simultaneously, thereby better understanding the relationships between words and phrases in a sentence.

The datasets used for pre-training are typically massive, comprising hundreds of millions to billions of words[5], sourced from diverse origins such as news articles, books, social media posts, and forum discussions. This dataset's diversity and extensive coverage enable the model to learn language features across domains and contexts, thereby equipping it with strong generalisation capabilities. However, this process demands substantial computational resources and time. Training a large-scale model like GPT-4 or BERT usually requires hundreds to thousands of high-performance GPUs or TPUs and can take weeks to months to complete. This results in high costs and significant energy consumption.

The outcome of pre-training is a model capable of generating rich semantic representations that capture linguistic features at word, phrase, and sentence levels. During the fine-tuning stage, these representations enable the pre-trained model to quickly adapt to various downstream tasks, such as text classification, Q&A systems, sentiment analysis, and machine translation. In fine-tuning, the model is further trained on task-specific datasets, typically requiring fewer training data since the model has already acquired extensive language knowledge during pre-training.

Despite the impressive performance and wide application range of pre-trained models, the process faces challenges. Resource consumption is a primary concern, as pre-training requires extensive computational resources and electricity, leading to high costs and environmental impacts. Moreover, pre-trained models may learn biases and inaccuracies in the training data, resulting in biased and misleading application outputs. Protecting data privacy and security during pre-training is also a critical issue that needs attention.

In summary, model pre-training is a core component of developing large language models. By training on extensive text data, models can learn broad linguistic features and semantic structures, providing a solid foundation for specific tasks. However, future research and practice must address resource consumption, data biases, and privacy protection challenges.

3.3 Model Fine-Tunning

Model fine-tuning is also crucial in developing large language models, allowing the pre-trained

model to adapt to specific tasks or domains. While pre-trained models have broad language features and semantic understanding, fine-tuning aligns the model's capabilities with the requirements of particular applications. This process typically follows a supervised learning framework, using labelled task-specific datasets for training. For instance, if a pre-trained model is applied to a text classification task, the fine-tuning dataset will contain annotated texts categorised into different classes, enabling the model to learn to distinguish these categories.

The primary method of fine-tuning involves using the pre-trained model's parameters as initial values and continuing training on task-specific datasets. This process requires less data and computational resources than pre-training, as the model already possesses fundamental language understanding. Fine-tuning adjusts the model's parameters to suit new task requirements better. For example, fine-tuning a pre-trained BERT model for sentiment analysis can enable it to accurately identify sentiment trends in text, with the fine-tuned model optimising its weights to improve the prediction of positive and negative sentiments.

The flexibility and efficiency of fine-tuning allow large language models to be widely applied to various NLP tasks, including Q&A systems, text generation, machine translation, and named entity recognition[6]. During fine-tuning, model architecture and training strategies can be adjusted according to task needs, such as modifying the loss function, adjusting learning rates, and introducing regularisation techniques to prevent overfitting.

While fine-tuning significantly enhances model performance in specific tasks, it also faces challenges. The quality and scale of the fine-tuning dataset directly affect model performance. High-quality, diverse annotated datasets can better exploit the potential of pre-trained models. Moreover, biases in fine-tuning data can still impact model outputs, requiring data selection and processing to mitigate bias effects on the model.

In summary, model fine-tuning is critical in applying pre-trained models to specific tasks. By training on task-specific datasets, models transition from general language capabilities to targeted task abilities, achieving efficient and accurate performance in practical applications.

3.4 Performance Evaluation Methods

Performance evaluation is vital to developing and applying large language models to measure the model's effectiveness and reliability in specific tasks. Evaluating model performance requires various metrics and methods to reflect the model's capabilities and limitations comprehensively. Standard performance evaluation metrics include accuracy, precision, recall, F1 score, and AUC-ROC curve. These metrics quantitatively analyse the model's performance in classification tasks. For instance, accuracy measures the proportion of correct predictions, precision evaluates the proportion of true positives among all optimistic predictions, recall reflects the proportion of true positives among all actual positives, the F1 score combines precision and recall, and the AUC-ROC curve shows the model's classification ability at different thresholds.

In NLP tasks, specific performance evaluation methods vary by task type. For text generation tasks, standard evaluation metrics include BLEU, ROUGE, and METEOR, which measure the similarity between generated text and reference text to evaluate the quality of text generation[6]. In Q&A systems, evaluation metrics may include accuracy, average response time, and user satisfaction, assessing the accuracy and efficiency of the model's responses.

Besides quantitative metrics, qualitative evaluation is also crucial in assessing model performance. A manual review of model outputs can identify the model's performance in specific contexts, such as whether it generates inappropriate or biased content. User feedback and expert reviews are also crucial qualitative evaluation methods, providing insights for model optimisation and improvement through real-world application feedback.

In performance evaluation, model generalisation ability, i.e., the model's performance on unseen data, is also crucial. This is typically achieved through cross-validation and hold-out, as well as training and testing the model on different data subsets to ensure it fits the training data and has good generalisation ability.

Overall, performance evaluation is critical to ensuring the reliability and effectiveness of large language models in practical applications. The model's strengths and weaknesses can be fully understood by comprehensively using various quantitative and qualitative evaluation methods, guiding subsequent optimisation and improvement work.

3.5 Deployment Optimization

Deployment optimisation is essential to evaluating large language models' performance and impact in practical applications, covering multiple factors such as response time, throughput, resource utilisation, and user satisfaction. First, response time is a crucial metric for measuring the model's speed in handling requests, especially in real-time applications like online customer service and chatbots. A well-optimized deployment should return results within milliseconds, ensuring smooth and natural user interaction.

Throughput refers to the number of requests a model can process per unit of time, directly reflecting the system's concurrent processing capability. High throughput is essential in high-traffic application scenarios, such as intelligent assistants in large websites and mobile apps[7]. To enhance throughput, deployment may involve load balancing, distributed computing, and caching techniques to optimise resource allocation and improve overall system efficiency.

Resource utilisation is another critical metric involving the model's use of computational resources (such as CPU, GPU, and memory) during operation. Effective resource utilisation means maximising performance while minimising hardware and energy costs. This helps reduce operational expenses and environmental impact, aligning with green computing principles. Techniques like model quantisation, knowledge distillation, and mixed-precision computing can significantly enhance resource utilisation efficiency, achieving high-performance deployment under limited hardware conditions.

User satisfaction is a critical dimension for evaluating deployment effects. User feedback and behaviour data provide direct insights into the model's performance in real-world use. High user satisfaction usually indicates that the model meets expectations in task completion, interaction smoothness, and result accuracy. Regular collection and analysis of user feedback help identify and address issues in practical applications, leading to timely optimisation and improvement.

Additionally, deployment optimisation includes the model's stability and reliability, ensuring consistent high-quality service under various usage scenarios and loads. This requires rigorous testing and monitoring during deployment to ensure the system has good fault tolerance and scalability.

Overall, deployment optimisation is a comprehensive standard for evaluating large language models' performance in real-world applications. The model's real-world performance can be fully understood through multi-dimensional evaluation, guiding further optimisation and improvement to ensure it performs at its best in various application scenarios.

4. Conclusion

This study delves into the training process and methods of large language models, focusing on the significance of one's knowledge base, the specific processes of model pre-training and fine-tuning, model performance evaluation, and deployment effects assessment. We have summarised an efficient and flexible training process for large language models through systematic analysis, enhancing their performance in specific tasks and providing solid theoretical and technical support for practical applications.

Firstly, the study concludes that a knowledge base is crucial for improving the professionalism and

accuracy of large language models. High-quality data collection, cleaning, and processing workflows ensure the relevance and reliability of training data, significantly enhancing the model's performance in specific fields. For example, in specialised fields like healthcare and finance, one's knowledge base compensates for the limitations of public datasets, allowing the model to understand better and generate professional language content. Additionally, model pre-training and fine-tuning are key stages ensuring an efficient transition from broad language knowledge to specific tasks. The pre-training stage leverages massive general data for language feature learning. In contrast, the fine-tuning stage further trains the model on task-specific data, achieving excellent performance in practical applications.

Model performance evaluation and deployment effect analysis further verify the proposed framework's effectiveness. Through metrics like accuracy, precision, recall, F1 score, and actual deployment assessments of response time, throughput, and resource utilisation, the study demonstrates that optimising model architecture and training strategies can significantly reduce computational resources and energy consumption while ensuring performance and achieving efficient and economical deployment.

Regarding research significance, this study provides practical guidelines for large language model training and offers important insights for future technological development. Firstly, the study emphasises the importance of an own knowledge base, recommending future large language model development to focus on constructing and managing high-quality datasets to ensure model professionalism and accuracy. Secondly, by proposing resource optimisation and bias handling techniques, the study provides feasible solutions for addressing resource consumption and data bias issues in large language models. These techniques can reduce operational costs and enhance the fairness and reliability of model outputs. Furthermore, the study proposes new approaches to improving the interpretability and controllability of large language models. Enhancing model interpretability allows users to better understand the model's decision-making process, increasing trust in the model. At the same time, controllable generation techniques ensure the quality and compliance of model outputs, meeting the needs of various application scenarios. These improvements collectively provide theoretical and practical support for the widespread application and continuous optimisation of large language models.

In summary, this study provides a systematic training process and methods for the efficient deployment and application of large language models in specific fields, summarising current technological status and challenges and proposing optimisation strategies and research directions for the future. This contributes positively to advancing large language models' practical application and technological progress.

References

[1] Wolf T., Chaumond J., Debut L., Sanh V., & Delangue C. (2020). "Transformers: State-of-the-Art Natural Language Processing." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp 38-45.

[2] Strubell E., Ganesh A., & McCallum A. (2019). "Energy and Policy Considerations for Deep Learning in NLP." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645-3650.

[3] Radford A., Narasimhan K., Salimans T., & Sutskever I. (2018). "Improving Language Understanding by Generative Pre-Training." OpenAI, pp. 1-12.

[4] He P., Liu X., Gao J., & Chen W. (2021). "DeBERTa: Decoding-enhanced BERT with Disentangled Attention." In International Conference on Learning Representations, pp 1-19.

[5] Liu Zhiyuan, Sun Maosong. (2019). "Natural Language Processing: Methods Based on Pre-trained Models." Beijing: Science Press, pp. 15-20.

[6] Wang Liwei, Zhang Min. (2020). "Deep Learning-Based Natural Language Processing." Beijing: People's Posts and Telecommunications Press, pp. 85-92.

[7] Jiang Tianzai, Liu Peng, Yang Jian. (2021). "Deep Learning and Natural Language Processing: Algorithms, Models, and Applications." Beijing: Electronic Industry Press, pp. 110-120.