International Journal of Mobile Computing and Multimedia Communications
Volume 16 • Issue 1 • January-December 2025

Full-Parameter Fine-Tuning Method of LLMs for Sports Injury Prevention and Treatment

Xinli Zhu
https://orcid.org/0009-0002-6372-8886
North Automatic Control Technology Institute, China

Zhiqiang Gao
https://orcid.org/0000-0003-1676-2950
Engineering University of PAP, China

Xu An Wang
https://orcid.org/0000-0003-2070-4913
Engineering University of PAP, China

ABSTRACT

Fine-tuning large language models (LLMs) for sports injury prevention and treatment in resource-constrained environments poses significant challenges due to memory demands and growing size of data. This paper proposes an efficient full-parameter fine-tuning approach based on Gradient Low-Rank Projection (GaLore) to reduce memory usage. Further, a data augmentation strategy for sports injury prevention and treatment is utilized to finetune a question-and-answer (Q&A) model with 0.5B parameter on consumer GPUs with 24GB memory. Experiment results show that the proposed method enhanced by GaLore is superior to SOTA methods such as low-rank adaptation (LoRA) in terms of convergence accuracy, training time, memory consumption, and indicators of BLEU-4 and ROUGE-2. Meanwhile, the empirical effect of injury prevention Q&A cases indicate that Qwen2-0.5B-Instruct trained by the proposed method have obvious advantages in professional knowledge understanding and overcoming hallucinations.

KEYWORDS

Vertical Large Language Models, Gradient Low-Rank Projection, Question-and-Answer, Low-Rank Adaptation, Qwen2-0.5B-Instruct, System Framework, Hallucinations

INTRODUCTION

Large language models (LLMs) have achieved remarkable progress across multiple disciplines, including conversational artificial intelligence (AI) and language translation, owing to their powerful language understanding and generation capabilities. From the perspective of applications, LLMs have also demonstrated impressive capabilities, but the bar for clinical applications is high. Yet today's AI models for applications in healthcare have largely failed to fully utilize language. For instance, sports injury is a common problem in sports; these injuries have a negative impact on athletes' physical health and sports performance. On one hand, LLMs can be used to analyze the cases of sports injuries as well as provide diagnosis and treatment suggestions for doctors and rehabilitation practitioners. On the other hand, LLMs can summarize the prevention methods and precautions for different types of sports injuries and provide personalized sports injury prevention programs for athletes and sports

DOI: 10.4018/IJMCMC.376486

enthusiasts. However, LLMs in the field of sports injury prevention and treatment face a series of challenges:

1. Multimodality: How to effectively integrate and interpret these different types of data is a key issue (Cui et al., 2021).
2. Accuracy: In diagnosis and treatment, the output of the model must be highly precise, as any minor error can lead to serious consequences.
3. Customization: Treatment methods need to be personalized according to the specific conditions of patients, handle individual differences, and provide personalized treatment recommendations (Silvestri et al., 2024). Additionally, a large amount of labeled data is usually unavailable to train efficient LLMs in low-resource scenarios.

Therefore, in response to the intelligent Q and A needs for sports injury prevention and treatment, it is of great significance to leverage the advantages of LLMs for sports injury prevention, to safeguard the health of fitness enthusiasts, and to create a more professional and effective Q and A assistant in the field of sports injury prevention and treatment.

This paper is focused on fine-tuning LLMs for sports injury prevention and treatment in resource-constrained environments. The authors are the first to utilize GaLore for full-parameter fine-tuning LLMs and data augmentation strategies for sports injury on consumer-level graphics processing units (GPUs). Overall, the contributions of this work can be summarized as follows:

1. Due to memory demands and the growing size of data in resource-constrained environments, an efficient full-parameter fine-tuning approach based on GaLore is proposed for sports injury prevention and treatment.
2. To improve data quality and reliability, a human-in-the-loop data augmentation strategy for sports injury prevention and treatment is utilized. Chinese medical dataset cMedQA2 and high-quality injury Q and A pairs are integrated, thereby ensuring the high relevance and practical quality of the dataset.
3. To assess LLMs using GaLore for sports injury prevention and treatment, we trained Qwen2-0.5B-instruct on consumer GPUs with 24GB of memory. Our method is superior to state of the art in terms of convergence accuracy, training time, memory consumption, and indicators of bilingual evaluation understudy (BLEU)-4 and recall-oriented understudy for Gisting evaluation (ROUGE)-2.

The rest of the paper is organized as follows. The second section makes a systematic summary of related works. In the third section, we propose an efficient full-parameter fine-tuning approach based on GaLore. In addition, a human-in-the-loop data augmentation strategy for sports injury prevention and treatment is illustrated. The fourth section evaluates the performance of our method compared with the state-of-the-art approaches. The final section summarizes the paper and points out the future research directions.

RELATED WORK

Medical LLMs

Inspired by the great success of general LLMs, the development and application of medical LLMs have received increased attention. However, LLMs still face difficulties in knowledge updating and the problem of hallucinations, especially in the complex and professional medical field. Currently, there are mainly two kinds of strategies for developing medical LLMs: to train a medical LLM on a large-scale corpus containing medical data and to fine-tune a general LLM using medical data. A