WEP Warwick Evans Publishing

International Journal of Public Health and Medical Research

ISSN: 3006-7707 (Print), ISSN: 3006-3752 (Online) | Volume 6, Number 5, Year 2026
DOI: https://doi.org/10.62051/ijphmr.v6n5.01
Journal homepage: https://wepub.org/index.php/IJPHMR/index

From Reactive to Proactive: Empowering Sports Injury Prevention with Large Language Models

Jianlong Wang [1, #], Tianao Guo [1, #], Zihan Xu [2, *]

[1] Research Center for Sports Psychology and Biomechanics, China Institute of Sport Science, Beijing 100061, China

[2] Institute of Physical Education, Jiangsu Normal University, Xuzhou, 221116, China

*Corresponding Author: 18705253069@163.com

#These authors contributed equally to this work

ABSTRACT

Sports injury management is undergoing a paradigm shift from 'reactive post-event intervention' to 'proactive pre-event prevention'. Although traditional machine learning (ML) and deep learning (DL) have demonstrated some efficacy in identifying injury risk factors, they still face analytical bottlenecks when dealing with complex, heterogeneous, and multimodal sports data. This paper aims to systematically explore the current applications of large language models (LLMs) in competitive sports injury prevention, the optimisation pathways for core technologies, and the empirical challenges they face in complex clinical decision-making. It examines how Retrieval-Augmented Generation (RAG), Parameter-Efficient Fine-Tuning (PEFT), and Multimodal Large Language Models (MLLMs) can bridge the general language barrier to adapt to specialised knowledge in sports medicine. Although LLMs possess powerful natural language processing and multimodal data integration capabilities, issues such as data privacy, model hallucinations, out-of-distribution (OOD) robustness, and insufficient explainability limit their independent application in high-risk medical decision-making. Future research needs to be deepened in the directions of edge computing deployment and explainable artificial intelligence (XAI).

KEYWORDS

Large Language Models; Sports Injury Prevention; Multimodal Data

1. INTRODUCTION

High-level competitive sport places extremely high demands on athletes' physiology, psychology, and biomechanics, which inevitably increases the risk of musculoskeletal injuries. Traditional sports injury management has long relied on a 'reactive' model, in which clinical diagnosis and rehabilitation interventions are initiated only after structural damage has occurred. However, modern sports science is actively driving a paradigm shift towards 'proactive protection'. The core of this shift lies in the continuous monitoring of multi-dimensional data—including training load, kinematic characteristics, and fatigue biomarkers—to issue early warnings before minor injuries accumulate into major trauma. With the deep integration of digital technology and the sports industry, wearable sensors and high-speed camera systems (such as markerless motion capture systems like OpenCap) have been widely adopted in frontline training, accumulating vast amounts of multimodal data. Previously, traditional machine learning algorithms such as random forests, support vector machines, and convolutional neural networks were widely used to identify injury risk factors [1, 2]. However, these models face significant limitations when processing unstructured medical records, coaching

notes, and complex cross-modal semantics. LLMs, such as GPT-4 and LLaMA, offer a revolutionary technical pathway for data integration in sports medicine, thanks to their vast knowledge bases—derived from their billions of parameters—and powerful zero-shot/few-shot reasoning capabilities [3]. LLMs can integrate medical records, injury reports, and research literature to provide analytical support for rehabilitation decisions and injury prevention. While recent reviews have covered the application of traditional ML in sports, comprehensive syntheses evaluating the specific architectural adaptations of LLMs (such as RAG and MLLMs) for sports medicine remain scarce [4]. This study outlines the core technical pathways and typical application scenarios of LLMs in sports injury prevention, whilst providing an in-depth analysis of the empirical challenges they face and future directions for development.

2. CORE TECHNOLOGICAL PATHWAYS FOR DOMAIN ADAPTATION

2.1. Retrieval-Augmented Generation (RAG) in Clinical Decision Support

In high-stakes sports medicine decision-making, model hallucinations can lead to catastrophic consequences (such as providing incorrect load management recommendations). RAG technology effectively mitigates this issue by retrieving relevant document fragments from external authoritative knowledge bases (such as IOC consensus statements and the latest sports medicine guidelines) before the model generates an answer [5]. Furthermore, GraphRAG technology organises structured data into knowledge graphs, capturing the complex network relationships between athletes, past injuries, and training loads, thereby significantly improving answer accuracy when addressing multi-hop reasoning problems [6].

2.2. Knowledge-Aided Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning is an indispensable step in enabling models to gain a deep understanding of the biomechanical mechanisms of movement. Given the extremely high computational cost of full-parameter fine-tuning, parameter-efficient fine-tuning techniques such as Low-Rank Adaptive (LoRA) have become the mainstream approach [7]. By freezing pre-trained weights and introducing only a minimal trainable matrix within the attention matrix, LoRA reduces computational requirements by a factor of several thousand. Furthermore, the Knowledge Graph Fine-Tuning (KG-FIT) technique, which incorporates structured triplet information during the fine-tuning process, enables the model to accurately grasp the mechanical and physiological relationships between entities (such as the association between ‘genu valgum’ and ‘anterior cruciate ligament rupture’), resulting in significant performance improvements in knowledge-intensive reasoning tasks [8].

2.3. Architectural Integration of Multimodal Large Language Models (MLLMs)

An athlete’s injury risk profile is constructed from a combination of visual data (motion video), temporal sensor data (heart rate, GPS load), and textual data (subjective fatigue questionnaires). By integrating a visual encoder with a language model, multimodal large language models achieve the alignment and fusion of cross-modal features within a unified semantic space [9]. This architecture is capable of processing dynamic inputs from real-world sporting environments and represents the most advanced form of intelligent sports injury assessment currently available.

3. KEY APPLICATION SCENARIOS IN INJURY PREVENTION

3.1. Multimodal Risk Identification and Biomechanical Analysis

The occurrence of sports injuries is the result of the combined effects of biomechanical characteristics, training load, and fatigue. Kinematic analysis based on untagged motion capture systems such as

OpenCap has been shown to offer good reliability when evaluating movements such as landing from a jump. Deep neural networks are responsible for detecting abnormal movement patterns at the foundational level (such as excessive knee valgus angle upon landing), whilst multimodal LLMs integrate the athlete's historical training loads and physiological indicators on this basis, converting complex numerical outputs into natural language warnings that are easily understood by coaches. For example, the model might output: "This athlete's landing cushioning time has significantly decreased. Combined with a 20% overload in high-speed running distance this week, the risk of lower limb soft tissue injury is extremely high. It is recommended that high-intensity training be ceased immediately."

The occurrence of sports injuries is multifactorial and results from a complex interplay between acute and chronic training loads, neuromuscular fatigue, and abnormal biomechanical characteristics. In the past, precise biomechanical assessment was confined to controlled laboratory settings and relied on traditional optical marker-based systems (such as Qualisys or Vicon). However, the advent of advanced computer vision and markerless motion capture systems has revolutionised kinematic analysis, enabling the direct and continuous capture of high-speed, multi-plane movements (such as complex aerial manoeuvres, gymnastic jumps and rapid changes of direction) within ecologically valid training environments [10]. The aetiology of non-contact lower limb injuries, particularly anterior cruciate ligament (ACL) ruptures, depends largely on subtle kinematic deviations, including dynamic knee valgus, excessive tibial anterior shear forces, and insufficient shock absorption upon landing from jumps [11]. Deep neural networks and pose estimation algorithms excel at extracting these spatio-temporal coordinates. However, interpreting these high-dimensional time-series data and validating them against optical ground truth through parameters such as root mean square error (RMSE) requires advanced semantic processing. Here, multimodal LLMs serve as a crucial analytical bridge. By integrating kinematic data with athletes' historical training loads, LLMs can autonomously construct robust early-warning workflows. To mitigate the curse of dimensionality and prevent data leakage when processing complex kinematic curves, these integrated AI systems typically employ dimension reduction techniques, such as functional principal component analysis (fPCA). Advanced LLM agents can monitor the analytical workflow to ensure methodological rigour, verifying that subject-level cross-validation strategies are strictly enforced and that the fPCA transformation matrix is applied to the test set only after being fitted to the training set. Ultimately, the LLM not only plots quantitative joint angles but also generates precise diagnostic reports: "Compared to baseline, this athlete exhibits a 4-degree increase in the valgus angle of the right knee during the absorption phase of the vault landing. This deteriorated movement strategy, combined with a 15% increase in acute training load, significantly elevates the risk of non-contact ACL injury. Immediate biomechanical correction and a targeted reduction in high-intensity training volume are recommended."

3.2. Real-time Physiological Monitoring via Wearable Devices

The integration of wearable biosensors with microtechnology has laid the foundation for data-driven, continuous monitoring of athletes' internal and external loads. LLM agents are uniquely capable of processing asynchronous, high-speed data streams from smart clothing, GPS trackers, and heart rate monitors [12]. The relationship between training load and overuse injuries is not linear; it is closely linked to the acute-to-chronic work ratio (ACWR) and the athlete's physiological capacity to withstand mechanical stress [13]. Whilst traditional machine learning models excel at flagging breaches of quantitative thresholds, the advantage of LLMs lies in their ability to contextualise these physiological indicators—such as reduced heart rate variability (HRV), altered sleep architecture, and elevated resting heart rate—through natural language interaction [14]. For example, an LLM-driven conversational agent can proactively query athletes reporting high levels of subjective fatigue (perceived exertion ratings) to distinguish between normal delayed-onset muscle soreness (DOMS) and early systemic manifestations of tissue degeneration. By seamlessly integrating objective sensor telemetry data with subjective natural language processing (NLP) inputs, the LLM constructs a comprehensive real-time physiological profile. If the system detects maladaptive responses (such as

a sustained decline in parasympathetic HRV accompanied by a persistently high external running load), it dynamically generates automated intervention protocols. This might include immediately replacing a scheduled high-intensity interval training (HIIT) session with an active recovery regimen, thereby preventing structural damage before fatigue accumulation leads to it.

3.3. Generation of Personalised Rehabilitation Plans and Return-to-Play (RTP) Decisions

Following severe musculoskeletal trauma, the rehabilitation process is extremely complex and fraught with risks of graft failure, muscle atrophy and secondary injuries. The decision to RTP is arguably the most critical and challenging clinical judgement in sports medicine, requiring a detailed, evidence-based, comprehensive assessment of tissue healing, functional biomechanics and psychological readiness [15]. LLMs have demonstrated unprecedented utility in the design and iteration of personalised rehabilitation programmes by effectively processing large volumes of unstructured clinical data alongside structured physical metrics [16]. By analysing the specific details of surgical interventions (e.g., autograft versus allograft anterior cruciate ligament reconstruction), daily physiotherapist records, and progressive functional movement screenings, LLMs can devise highly personalised, multi-stage rehabilitation plans. They can dynamically adjust these long-term project management frameworks based on real-time clinical milestones—functioning almost like an automated Gantt chart for the athlete's rehabilitation cycle. During the critical RTP phase, athletes undergo rigorous assessments, including isokinetic strength testing to evaluate strength symmetry in the quadriceps and hamstrings, as well as multi-directional jump tests. Long-term rehabilitation managers combine these structured datasets with qualitative clinical evaluations to calculate a comprehensive injury risk phenotype. Furthermore, an athlete's successful return to play depends largely on their psychological state. Sports phobia (the fear of re-injury) is one of the primary reasons athletes fail to return to their pre-injury competitive level. Advanced LLMs can assess an athlete's psychological readiness by analysing emotional and semantic structures in self-reported confidence logs or clinical interviews—typically using tools such as the Anterior Cruciate Ligament Return to Sport Index (ACL-RSI) for measurement [20]. Ultimately, the LLM acts as a central hub, providing the multidisciplinary medical team with a comprehensive and objective risk-benefit analysis. This ensures that athletes are not only in good physical condition and biomechanically stable, but also psychologically fortified, enabling them to cope with the gruelling challenges of high-level competitive sport.

4. CURRENT CHALLENGES AND LIMITATIONS

4.1. Data Quality, Scarcity, and Multimodal Alignment

High-quality, medical-grade sports data is extremely costly to acquire. Compared to general-purpose datasets comprising millions of entries, high-quality question-and-answer or multimodal datasets specific to particular sports injuries remain very limited in scale. Even more challenging is the issue of spatiotemporal alignment in multimodal data: video cameras may operate at 60 frames per second, whilst position tracking systems have a sampling rate of 10 hertz, and commentary is non-continuous. This inconsistency in sampling frequencies and timestamps leads to a high risk of misalignment in semantic relationships between different modalities, severely limiting the accuracy of model fusion reasoning [1, 8].

4.2. Model Reliability, Hallucinations, and Out-of-Distribution (OOD) Robustness

General-purpose large language models often produce severe 'hallucinations' when answering specialised sports-related questions, particularly those involving specific numerical values or medical interventions. Furthermore, the competitive sports environment is highly unpredictable, frequently

presenting rare injury mechanisms or extreme load scenarios not covered in the training data (i.e., out-of-distribution situations, OOD). Research indicates that existing large language models are highly prone to producing unreliable and erroneous outputs when faced with OOD inputs, and their performance is highly susceptible to fluctuations caused by even subtle changes in prompts [18].

4.3. Data Privacy and Computational Architecture Conflicts

Athletes' physiological data and medical records are not only personal privacy matters but also constitute core confidential information for national teams and professional clubs. The inference processes of large language models typically require uploading these distributed, heterogeneous data to cloud servers, which poses a serious risk of data leakage. Although federated learning (FL) and differential privacy (DP) offer theoretical solutions, the computational capacity and power supply of edge nodes in sports venues are extremely limited, making it difficult to bear the heavy communication burden [19].

4.4. Black-Box Nature and Lack of Interpretability

In medical decision-making that affects athletes' careers, medical teams and coaches must understand the logic behind model recommendations. However, the inherent 'black-box' nature of LLMs results in a lack of transparency in their decision-making processes. If a model suggests adjusting training, experts need to know whether this is based on accumulated fatigue or biomechanical compensation. Current models are unable to provide such high-quality explanations with causal logic, limiting their clinical trustworthiness [20].

5. FUTURE DIRECTIONS

5.1. Privacy Protection and Edge Computing Deployment

Future system architectures must shift towards edge computing. By introducing technologies such as Software-Defined Networking (SDN), local data processing can be achieved in resource-constrained environments, thereby reducing latency and ensuring absolute data security 错误!未找到引用源。. Researchers need to further explore model compression and low-bit quantisation techniques, and develop lightweight models (1B–7B parameter level) optimised for sports scenarios to enable offline, real-time inference support at training venues [19, 21].

5.2. Building a Highly Transparent Explainable Artificial Intelligence (XAI) Ecosystem

Breaking down the 'black box' is a prerequisite for clinical implementation. Future LLMs must deeply integrate XAI technical frameworks. The practical paradigm of the FST.ai 2.0 system demonstrates that by combining cognitive uncertainty modelling via graph convolutional networks (GCNs) with a visualisation and explanation overlay layer, it is possible to provide users with transparent decision support, raising trust in AI-assisted decision-making to 93% [22]. Future models must incorporate mechanisms for quantifying uncertainty, enabling them to proactively 'refuse to answer' or provide confidence prompts when faced with medical issues beyond their cognitive boundaries.

6. CONCLUSION

Multimodal large language models represent the pinnacle of next-generation AI in the field of sports science and are profoundly reshaping the technological landscape of injury prevention in competitive sports. By integrating advanced architectures such as RAG and LoRA, LLMs have acquired the

capability to transform complex biomechanical parameters, wearable sensor data, and clinical text into proactive injury warnings and personalised rehabilitation interventions. However, challenges such as data alignment difficulties, model hallucinations, insufficient out-of-distribution robustness, and data privacy necessitate a rigorous and cautious approach to their clinical deployment. Future breakthroughs lie in the implementation of edge computing frameworks and the refinement of explainable AI systems. Only by adhering to the ethical principle that 'AI assists rather than replaces expert decision-making', and by deeply fostering interdisciplinary collaboration between computer science and sports medicine, can large language models truly become a powerful intelligent engine safeguarding the health of elite athletes and supporting the high-quality development of competitive sports.

REFERENCES

[1] Musat, C. L., Mereuta, C., Nechita, A., Tutunaru, D., Voipan, A. E., Voipan, D., ... & Nechita, L. C. (2024). Diagnostic applications of AI in sports: A comprehensive review of injury risk prediction methods. Diagnostics, 14(22), 2516.

[2] Claudino, J. G., Capanema, D. D. O., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., & Nassis, G. P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. Sports Medicine - Open, 5(1), 28.

[3] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv Preprint arXiv:2303.18223, 1(2), 1-124.

[4] Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. Journal of Experimental Orthopaedics, 8(1), 27.

[5] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv Preprint arXiv:2312.10997, 2(1), 32.

[6] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv Preprint arXiv:2404.16130.

[7] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. ICLR, 1(2), 3.

[8] Jiang, P., Cao, L., Xiao, C., Bhatia, P., Sun, J., & Han, J. (2024). KG-FIT: Knowledge graph fine-tuning upon open-world knowledge. Advances in Neural Information Processing Systems, 37, 136220-136258.

[9] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. National Science Review, 11(12), nwae403.

[10] Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., & Delp, S. L. (2018). Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. Journal of Biomechanics, 81, 1-11.

[11] Turner, J. A., Kiefer, A. W., Bullock, G. S., Kucera, K. L., Cameron, K. L., Boling, M. C., ... & Padua, D. A. (2025). Reliability and predictive validity of trunk and lower extremity kinematics during a jump-landing task using OpenCap markerless motion capture system. Journal of Biomechanics, 113026.

[12] Ferrara, E. (2024). Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and challenges. Sensors, 24(15), 5045.

[13] Gabbett, T. J. (2016). The training—injury prevention paradox: Should athletes be training smarter and harder? British Journal of Sports Medicine, 50(5), 273-280.

[14] Merrill, M. A., Paruchuri, A., Rezaei, N., Kovacs, G., Perez, J., Liu, Y., ... & Liu, X. (2026). Transforming wearable data into personal health insights using large language model agents. Nature Communications.

[15] Ardern, C. L., Glasgow, P., Schneiders, A., Witvrouw, E., Clarsen, B., Cools, A., ... & Bizzini, M. (2016). 2016 Consensus statement on return to sport from the First World Congress in Sports Physical Therapy, Bern. British Journal of Sports Medicine, 50(14), 853-864.

[16] Desai, V. (2024, April). The future of artificial intelligence in sports medicine and return to play. In Seminars in Musculoskeletal Radiology (Vol. 28, No. 2, pp. 203-212). Thieme Medical Publishers, Inc.

[17] Webster, K. E., Nagelli, C. V., Hewett, T. E., & Feller, J. A. (2018). Factors associated with psychological readiness to return to sport after anterior cruciate ligament reconstruction surgery. The American Journal of Sports Medicine, 46(7), 1545-1550.

[18] Yuan, L., Chen, Y., Cui, G., Gao, H., Zou, F., Cheng, X., ... & Sun, M. (2023). Revisiting out-of-distribution robustness in NLP: Benchmarks, analysis, and LLMs evaluations. Advances in Neural Information Processing Systems, 36, 58478-58507.

[19] Piccialli, F., Chiaro, D., Qi, P., Bellandi, V., & Damiani, E. (2025). Federated and edge learning for large language models. Information Fusion, 117, 102840.

[20] Kranzinger, S., Halmich, C., Hofer, D., & Kranzinger, C. (2025). A scoping review of explainable artificial intelligence in sports science. Discover Artificial Intelligence.

[21] Yang, M., Gao, C., & Han, J. (2022). Edge computing deployment algorithm and sports training data mining based on software defined network. Computational Intelligence and Neuroscience, 2022, 8056360.

[22] Shariatmadar, K., Osman, A., Ray, R., & Kim, K. (2025). FST.ai 2.0: An explainable AI ecosystem for fair, fast, and inclusive decision-making in Olympic and Paralympic Taekwondo. arXiv Preprint arXiv:2510.18193.