Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy

MARIOS CHIRAS, University of Twente, The Netherlands

ABSTRACT

Large Language Models have shown great results in analyzing and processing different types of datasets. In this study, different parameter-size LLMs will be evaluated in a physiotherapist agent application, that process biomechanical running data from wearable devices. The research aims to determine the optimal size of the state-of-the-art LLMs used in this research that balances accuracy and computational resources. The study will explore and evaluate the models in a Retrieval-Augmented Generation (RAG) application.

The different models used in this study were divided into 3 categories Small, Medium, and Large models based on their parameter size. In the experiment, the Larger category was shown to offer the best balance of the highest accuracy when it comes to detecting outliers within the biomechanical data while balancing computational resources.

The study highlights a unique approach to developing physiotherapist applications with the integration of RAG and Text-To-SQL methodologies for processing and retrieving running data. The goal of this application is to provide valuable insights into the trade-offs between the accuracy and computation efficiency of different models. These findings contribute to the development of applications in sports physiotherapy, offering insights into the appropriate selection of LLMs in specific circumstances.

KEYWORDS

“Large Langauge Models”, “Sports Coaching”, “Rehabilitation”, "Biomechanical Data Processing", "AI in Physiotherapy"

1 INTRODUCTION

Wearable devices have changed the field of health and fitness by providing real-time medical data [25]. These advancements with the use of different sensors have allowed continuous monitoring of various data points, including heart rate, muscle activity, and joint movements, which are important for specific rehabilitation purposes [5]. Such wearable devices can capture biomechanical and biomechanic data that can be valuable information for sports physiotherapy [10]. In particular, by collecting data such as joint angles, and muscle efforts, these devices can help runners improve their performance and reduce the risk of injuries with the help of physiotherapists [24]. In this research, different state-of-the-art Large Language Models (LLMs) will be evaluated in analyzing and processing biomechanical and biomechanic data using Retrieval-Augmented Generation (RAG), to create a physiotherapist agent and find the optimal size.

Large Language Models (LLMs) have shown their ability to process complex datasets and provide useful insights into many different fields making them have high natural processing capabilities. This capability can be especially beneficial in healthcare, where LLMs can help health practitioners diagnose and find medical solutions.

However, effectively processing biomechanical, and biomechanical data and deploying LLMs comes with a significant challenge. The main problem is balancing the accuracy of data analysis with the computational resources required. Large parameter sizes usually offer higher accuracy but require substantial hardware, which might not be affordable to many people and small businesses to build such applications. On the other hand, Smaller LLMs are less resource-demanding but may lack the ability to process data accurately and consistently.

This study addresses the optimal size of an LLM that balances accuracy and resource usage in physiotherapist applications. This will be done by evaluating different parameter-size LLMs. The research aims to determine models that can process biomechanical and biomechanics running data from wearable devices effectively while balancing computational requirements. The study proposes a RAG physiotherapist system with a Text-to-SQL agent to retrieve the running data from the database, to evaluate the ideal size.

Current research has yet to address what is the optimal size for physiotherapist's application for LLMs. The study goal is to fill this gap by determining the optimal size of state-of-the-art open-source LLMs for processing running data.

This paper is structured as follows: Section 2 reviews related work on wearable devices and LLMs in health applications. Section 3 details the methodology, including data collection, model selection, and system development. Section 4 presents the experiment procedure, Section 5 provides the evaluation of the models based on specified metrics. Finally, Section 6 discusses the challenges, and answers the research questions and future work.

Research Questions:

1) How can we determine the appropriate size of existing open-source LLMs for processing biomechanical data, while maintaining high performance and minimizing computational requirements?

2) What are the trade-offs between accuracy and resource usage?

2 RELATED WORK

Local Large Language Models (LLMs) provide several advantages over centralized models like ChatGPT or OpenAI. The GPT family is the most well-known and most used model so far, with almost 100 million weekly active users [16]. Local models such as the Llama

TScIT 41, July 5, 2024, Enschede, The Netherlands

family from Meta, ensure data privacy, customization, and performance, which the GPT family does not offer [14]. Platforms like Ollama, allow users to be able to deploy LLMs locally, minimizing the risk of personal data being used to train models, data being sold to third parties, and exposed data breaches. This is especially great when it comes to creating an agent that is dealing with real user data, ensuring it follows privacy regulations [23].

2.1 Problems with LLMs

Large Language Models have shown high capabilities in natural language processing, but they also present a significant challenge. The major problem lies with the substantial hardware required to run these models. Deploying models like Llama3 requires expensive GPU and memory resources, making it difficult for a lot of users to be able to afford it [8]. Despite that, the initial cost of the hardware is not the only issue. These models consume a lot of energy during inference, making it difficult to maintain such a system. A great example of this is chatGPT, which is estimated to consume 100k dollars worth of energy every day [3].

2.2 LLMs in healthcare

Large Language Models have huge potential in many different fields. In health, LLMs have been incredibly useful by having improved the clinical process from recordkeeping to diagnosis prediction [4]. By efficiently analyzing large complex datasets, LLMs can have many different contributions to the healthcare field. LLMs have the capability to improve drug discovery through the capacity to scrutinize intricate molecular structures, radiology, and imaging with their multimodal abilities and help with clinical decision support for doctors [17]. However, despite their incredible potential in this field, there are significant security and privacy issues that must be addressed to ensure safety, when it comes to developing healthcare applications [7, 18].

2.3 Retrieval-Augumented Generation

Retrieval-augmented generation (RAG) is a strong approach for Natural Language Models to achieve high performance on specific tasks [12]. It allows LLMs to retrieve information from existing external knowledge that helps the models to guide toward an answer for a specific question based on a reliable source [11]. Research has been done that evaluated RAG applications for LLMs in specific metrics. It showed promising results by improving the accuracy and relevance in the responses of the models [6]. Regardless of the benefits of RAG in LLMs, implementation contains challenges such as ensuring the relevance of retrieved information, especially when dealing with unstructured sources as the knowledge base, which can often be difficult to process effectively [9].

2.4 Text-to-SQL

Text-to-SQL technology allows converting natural language into sending an SQL query to a database to fetch data [13]. The study in [20] showed the capabilities of large language models (LLMs) like GPT-4 that can effectively generate SQL queries and evaluated specific benchmarks for this task. The research highlighted the importance of prompt engineering for LLMs to provide relevant results in the context of text-to-SQL tasks and how text-to-SQL models can retrieve user data. Text-to-SQL can be integrated with RAG systems which helps to enhance data analysis [26]. Nevertheless, Text-to-SQL methodology has limitations regarding when it has to deal with complex database schemas and user demands requiring complex queries that can make the models achieve relatively low accuracy.

3 METHODOLOGY

This section shows each step of the methodology used in this research, focusing on each step involved in conducting this research. The methodology is structured into several phases: data collection, model selection, prompt engineering, development of a Retrieval-Augmented Generation system, and document selection for the system. Below is a roadmap where it shows each step of the methodology.

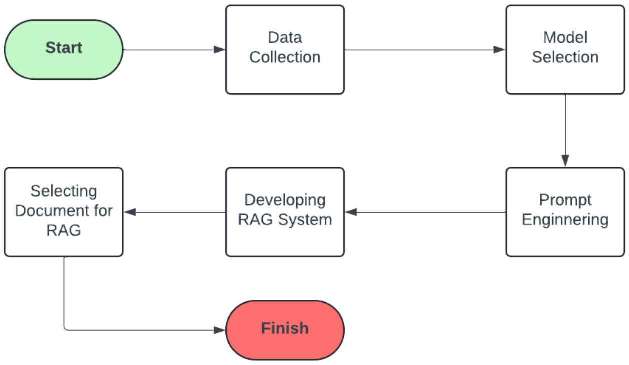

Fig. 1. This diagram provides an overview of the methods used in this research. It includes data collection, model selection, implementing prompt engineering techniques, developing a RAG system, and selecting a document for the system

3.1 Data Collection

This research will make use of generated synthetic data using Synthea Open Source Software to stimulate biomechanical data collected by wearable devices that is relevant for physiotherapy for runners [22]. The data will mimic real-world information collected by wearable devices, which physiotherapists can use to help them diagnose patients, create personal rehabilitation plans for running-related injuries, and understand patterns to improve the performance of athletes. This data will contain different joint movements and muscle activities. The data will be stored in an SQL database for retrieval and to be analyzed by the Retrieval Augmented Generation System. The following class diagram represents the SQL schema and the different data points.

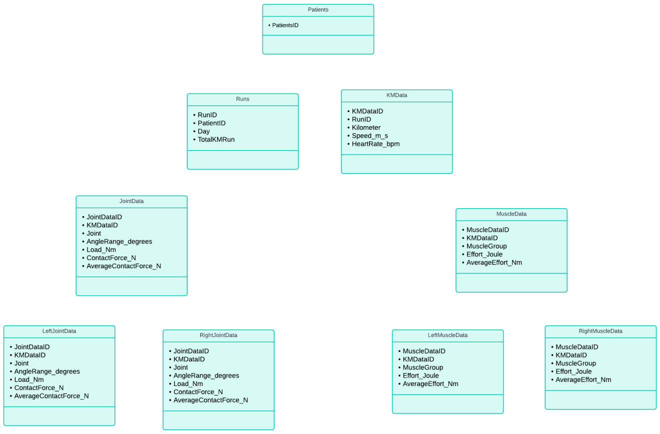

Fig. 2. The figure illustrates the database schema and the different biomechanical data points used in the study to evaluate the models. The data points are distinguished in 6 different ways: left, right, joint, muscle, run, and by patient.

3.2 Model Selection

In this research, several open-source Large Language Models (LLMs) will be selected to determine the optimal size of current state-of-the-art LLMs for processing biomechanical and biomechanic data accurately for physiotherapy applications, while balancing resource usage during inference. These models will be categorized into 3 different categories: small, medium, and large based on their parameter size. Categorizing the model based on the parameter size is essential since parameters influence not only the model's performance but also its computational requirements.

- **Small Models**: Models with only a few billion parameters, fall into the "Small" category. These models usually generate very fast responses during inference and require minimal computational resources, making them a great fit for applications that require fast responses but do not require accurate results.
- **Medium Models**: Models with several billion parameters are categorized in the "Medium" category. These models can generate relatively fast responses and accurate results, however, their accuracy results are not always consistent and it can vary in the use case. Medium models usually require significant computational resources, which may cause some challenges for deployment on standard hardware.
- **Large Models**: Large models with more than 10 billion parameters have state-of-the-art performance and accuracy. However, their accuracy does not come without a cost. They require extensive computational resources that most people cannot afford.

By evaluating these different models, this research aims to determine the optimal size between models for balancing accuracy and resource usage when it comes to processing biomechanical data for running. Below you will find a table with different selected LLMs with the category that falls, parameter size, and their minimum computational requirements.

Model	Category	Parameter Size	Minimum Hardware
Phi3	Small	3.8 Billion	Not available
Llama3	Medium	7 Billion	28GB VRAM
Llama3:70b	Large	70 Billion	140GB VRAM

Table 1. The table above shows the different models selected for this study by showcasing the category that they fall in, their parameter size, and their hardware requirements

3.3 Prompt Engineering

Prompts are very important when it comes to creating AI agents. They serve as a method of communication to guide the LLMs, enabling them to generate relevant and accurate outputs for specific tasks [1]. Prompt engineering allows us to guide the models effectively without the need for extensive retraining or fine-tuning [19]. There are several different prompt techniques, in this research we will focus on using "Few-shot". The models selected above will use this prompt technique for guidance in creating a Physiotherapist Agent for processing biomechanical data using RAG to retrieve relevant information from a knowledge based on the question of the user and Text To SQL methodologies to retrieve the biomechanical data from the database.

3.3.1 Few-Shot Prompting.

Few-shot prompting is a prompt technique that involves providing the model with a few examples of interaction between the user and the model. This helps the model understand the context and how the conversation between them and the user should be. By providing examples to the LLM, the model adjusts its responses accordingly and knows what kind of questions it expects from the user [2].

Prompt: As an expert sports physiotherapist, your task is to analyze the provided biomechanical data from wearable devices for a summarized run. The goal is to identify data points that are under or above the expected range based on the context provided in the document. Mention only the outliers in the data and provide what it indicates.

3.4 Developing RAG and Text-to-SQL System

The LLMs will be tested in the following proposed RAG system to determine the optimal size between them that balances accuracy and resource usage when processing biomechanical data. RAG is a technique that allows LLMs to use documents of different structures to be used as guidance when answering questions from the user. These documents can be in any format and for any domain. In this proposed system, the LLMs will be using a PDF document that provides information about the different data points from the data generated from the previous section as guidance to generate results. The information that this document will contain is about what the ideal values of the specific data points are and what they indicate. The proposed RAG system works as follows: The system takes the PDF document and splits it into smaller chunks via a splitting algorithm. These small chunks will then be converted into vector embeddings using the HuggingFaceEmebeddings and they will be stored in a Chroma vector database, which helps with the retrieval of relevant chunks based on the user's queries. The LLMs will be using a custom prompt template using the Few Shot Engineering technique from the other section above. When the user asks a question relevant information will be retrieved from the knowledge base and will be fed into the user prompt to be used as guidance for the LLMs to produce relevant responses. Additionally, the system will have two more models one that will identify which data points need to be retrieved and one model that will generate SQL queries to retrieve the data points identified from the database. After the model retrieves the data it will add the data to the user's query and send it to the Physiotherapy Agent Model to be analyzed. These additional models will not be tested and evaluated since it is outside the scope of this research but they are necessary for the application to function properly.

The following is a flow diagram where it shows how the proposed application works:

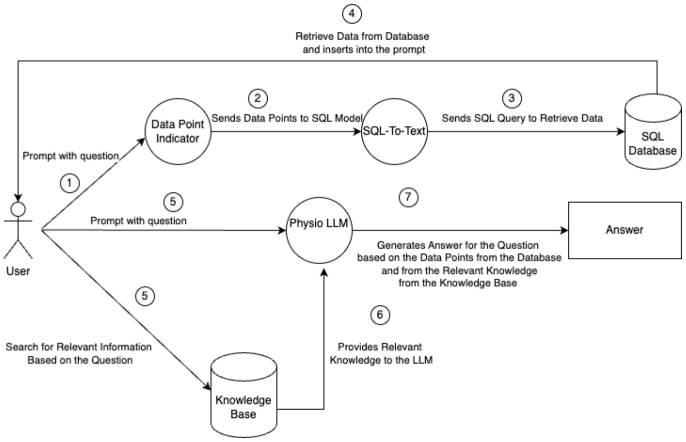
4
Retrieve Data from Database and inserts into the prompt
2
Sends Data Points to SQL Model
3
Sends SQL Query to Retrieve Data
Data Point Indicator
SQL-To-Text
SQL Database
Prompt with question
1
5
Prompt with question
Physio LLM
7
Answer
User
Generates Answer for the Question based on the Data Points from the Database and from the Relevant Knowledge from the Knowledge Base
5
6
Search for Relevant Information Based on the Question
Provides Relevant Knowledge to the LLM
Knowledge Base

Fig. 3. This figure shows a flow diagram of how the proposed RAG system works with numbers that represent each step of the process.

3.5 Document Selection

Document selection is a very critical part of implementing an RAG, as the document is responsible for the model to provide relevant and accurate responses. These documents can be books and documents relevant to the use case of the system. However, academic books might be difficult for the model to process, since they usually contain pictures, tables, and unstructured information and contain a lot of pages of content, which makes the vectorization embedding process complicated. For this system, a guideline document was generated by using GPT-4o LLM. This document generated by the AI model, provides guidelines about each data point, explaining what each value indicates and what it can cause. While this might not be the ideal document selection, it allows us to evaluate how well the models respond to the content and how effectively they follow the guidelines provided.

4 EXPERIMENT

This section covers the steps that were taken to conduct the experiment. This involves setting up the necessary hardware, finding evaluation metrics, and designing a dataset to evaluate these metrics.

4.1 Hardware Setup

To be able to deploy and evaluate the LLMs used in this study, necessary hardware and software are required to be able to run the models smoothly and meet their computational demands.

(1) **GPUs:** 3 x NVIDIA RTX A6000 48GB
(2) **CPU:** Intel(R) Xeon(R) w9-3495X
(3) **Memory (RAM):** 8 x 64GB 4800 MHz DDR4 RAM
(4) **Storage:** 2TB NVMe SSD

4.2 Evaluation Metrics

Evaluating the performance of LLMs in this research poses a significant challenge. The reason behind this issue is because LLMs sometimes generate hallucinations. For this reason, traditional methods are not suitable when it comes to evaluating them. To determine the optimal size balancing accuracy and resource usage the following metrics will be used:

(1) **Accuracy in Outlier Detection**: This metric determines the ability of the models to identify outliers within the biomechanical data based on the relevant information from the RAG system.
(2) **Response Time**: Response time is defined as the time it takes for the model to respond from the moment the user asks a question until the user hears back an answer from the LLM.
(3) **GPU Usage**: It is the amount of GPU Usage it takes from the moment the user asks a question to the model until the model generates an answer.
(4) **RAM Usage**: It is the amount of RAM Usage it takes from the moment the user asks a question to the model until the model generates an answer.

By focusing on these metrics, this study aims to evaluate the LLMs, determining which model offers the best balance between accuracy and resource usage for processing biomechanical data in physiotherapy applications.

4.3 Designing dataset to measure metrics

To evaluate the performance of models accurately, a well-designed dataset is essential. The dataset for evaluating the metrics contains a set of questions and answers. The questions in the dataset are questions that users might ask the models like for performance or symptoms that they might be experiencing including the related data from the database. The answers in the dataset are the identified outliers within the data what they indicate and what is expected for LLMs to respond. This dataset will be used to see how well the models follow the instructions from the document and also it will serve as a benchmark to evaluate the selected metrics.

Question 1:	**Question 2:**
Is my heart rate within a normal range during my run? KM 1: - Heart Rate: Avg 120 BPM, Max 130 BPM KM 2: - **Heart Rate: Avg 200 BPM, Max 250 BPM** KM 3: - Heart Rate: Avg 132 BPM, Max 150 BPM	My Hamstring kind of hurt during my last run. Is there anything wrong? KM 1: - Left Hamstring Maximum Effort: 280 Nm - Right Hamstring Maximum Effort: 290 Nm - Left Hamstring Average Effort: 168 Nm - Right Hamstring Average Effort: 169 Nm KM 2: - Left Hamstring Maximum Effort: 270 Nm - Right Hamstring Maximum Effort: 287 Nm **- Left Hamstring Average Effort: 53 Nm** - Right Hamstring Average Effort: 165 Nm KM 3: - Left Hamstring Maximum Effort: 264 Nm - Right Hamstring Maximum Effort: 287 Nm - Left Hamstring Average Effort: 156 Nm - Right Hamstring Average Effort: 164 Nm
Answer 1:	**Answer 2:**
Your heart during the run seems to be functioning normally, except in KM 2 it seems your average and max heart rate is higher than the normal range, this might be an indication of poor conditioning, overexertion, or stress.	Your hamstring data seems to be in normal ranges (Maximum and Average Effort). In KM 2, however, your Left Hamstring Average Effort of 53 Nm seems to be lower than expected. This may be because of an indication of underuse of the hamstring,

Table 2. This table provides examples of possible questions of users and expected answers of the models based on the biomechanical data stored in the database. It illustrates how the proposed system processes and analyzes the data to provide relevant feedback for sports physiotherapy.

5 EVALUATION

This section evaluates the performance of the LLMs using the proposed system based on the selected evaluation metrics and the designed dataset.

5.1 Performance of Metrics

These are the results after evaluating the metrics on the dataset

Model	Accuracy	GPU Usage	RAM Usage	CPU Usage	Response Time
Phi3	0.281	24.034	9.01	3.22	15.39
Llama3	0.458	39.1	7.94	14.73	21.71
Llama3:70b	0.549	77.78	13.06	35.66	42.75

Table 3. This table shows the performance of the LLMs on the following specified metrics: accuracy, GPU, RAM, CPU Usage and Response Time

5.2 Analysis of Metrics

The performance metrics show the trade-off between accuracy and resource usage for each model

(1) **Accuracy**: Llama3:70b shows the highest accuracy, making the ideal model for processing biomechanical data. In second place, Llama3 comes with a 0.99 difference, showing capabilities that it can handle and process this kind of data. Lastly, Phi3 came in last place with the lowest accuracy, showing that its results are unreliable.

(2) **Resource Usage**: Similarly, Llama3:70b shows the highest usage of RAM, GPU, CPU, and Response time making the model that consumes the most resources. This high consumption could be a significant limitation in an environment with limited computational resources. In contrast, Llama3 and Phi3 require fewer resources, with Phi3 being the most resource-efficient but at the cost of lower accuracy.

5.3 Determining the Optimal Size

To determine the optimal size between the models, we need to apply a weighted scoring to balance accuracy with resource usage. This involves normalizing the metrics, assigning weights, and computing the composite score.

5.3.1 ***Normalization of Metrics.*** The first step to finding the optimal size that balances accuracy and resource usage is to normalize each metric. Normalization is an essential step when weighted scoring is applied. It allows us to convert metrics of different scales into a mutual scale. In this case, the scale will be from 0 to 1.

To convert everything into a scale of 0 to 1, the following formula will be used for each metric:

$$\text{Normalized Metric} = \frac{\text{Metric Value}}{\text{Maximum Value}}$$

The following table shows the metrics after Normalization is implemented for each one of them:

Model	Accuracy	GPU Usage	RAM Usage	CPU Usage	Response Time
Phi3	0.51	0.30	0.69	3.09	0.36
Llama3	0.83	0.50	0.60	0.41	0.50
Llama3:70b	1.00	1.00	1.00	1.00	1.00

Table 4. The tables show the performance of metrics after normalization was implemented on every single metric to be in the scale from 0 to 1 using the formula used above.

5.3.2 *Selection of Weights.*

In this subsection, we will go through the process of applying weights to each metric and provide a reason behind the choice. This step is critical to ensure that each metric contributes appropriately based on its importance.

These are the weights chosen for each metric:

- **Accuracy**: 50%
- **GPU Usage**: 20%
- **RAM Usage**: 20%
- **CPU Usage**: 5%
- **Response Time**: 5%

5.3.3 *Justification of Weights.*

- **Accuracy:** The primary objective of this research is to determine the model that offers the best balance between accuracy and computational resources for processing biomechanical and biomechanics data. Keeping this in mind, accuracy is given the highest weight of 50% since it directly affects the reliability of the data analysis.
- **GPU Usage:** GPUs are critical for deploying and running LLMs, as LLMs are dependent on GPUs for their computing power. A higher GPU capacity leads to faster processing times and better performance. This metric is weighted is weighted at 20% for these reasons.
- **RAM Usage:** RAM is another important component for running and deploying. LLMs. These models require extensive memory to be able to run smoothly and meet the computational requirements of the model. Insufficient memory can lead to lower performance and failure in processing. Therefore, RAM usage is weighted equally with GPU usage of 20% to show its importance in ensuring smooth model execution.
- **CPU Usage:** While CPUs are important, usually LLMs do not rely on CPU that much but they are more dependent on GPU and RAM. CPU impacts the overall system performance, but its role is less important compared to GPU and RAM when it comes to running LLMs, this is why is weighted at 5%.
- **Response Time:** Response time is important for real-world applications, it is much less critical for this study compared to accuracy and GPU and RAM Usage. While response time enhances the user experience, it is not a primary factor when it comes to evaluating LLMs for performance, for that reason its weight is 5% as well the same as the CPU Usage.

5.3.4 *Composite Score Calculation.*

Using the normalized values and assigned weights, we can calculate the composite score for each model, which will allow us to determine which model has the ideal size for balancing accuracy and resource usage. To calculate the composite score, the following formula will be used:

$$\text{Composite Score} = \sum_{i=1}^{n} (\text{Weight}_i \times \text{Normalized Metric}_i)$$

The following graph compares the composition scores of each model.

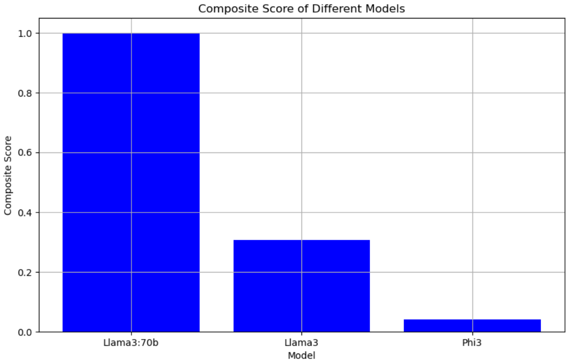

Fig. 4. This figure shows the different Composite Score of each LLM, which can provide useful information about which model balances accuracy and resource consumption the most

Based on the composite score of each model it seems the larger model balances accuracy and computational resources better than any other model. Proving that the larger model category is the optimal size when it comes to analyzing biomechanical running data.

6 DISCUSSION

The evaluation results cover important insights into the accuracy and resource usage of different LLMs in processing biomechanical data for physiotherapy applications.

6.1 Conclusion

The study explored different parameter size LLMs in a physiotherapy agent application for analyzing biomechanical data using RAG and Text-To-SQL methods. The goal of the study was to determine what is the optimal size of the models that were tested for balancing accuracy and computational resources.

6.1.1 *Answer to RQ1:*

The study concluded that larger models like Llama3:70b offer the best balance between accuracy and resource usage, as indicated by the highest composite score. While these models require much more resources than the other models, the higher accuracy compensates for the resource utilization, making it more ideal to process biomechanical data in physiotherapy applications. Medium-size models like Llama3 are a great alternative to balance performance but the accuracy measure in the experiment was below 50% making it unreliable for processing data. The Small Models have very low accuracy but they do not require expensive hardware.

6.1.2 *Answer to RQ2:*

- **Larger Models:** These models have the highest accuracy out of the rest of the models but require much more substantial computational resources. They are suitable for places where there is no lack of resources and high accuracy is required.
- **Medium Models:** Medium models offer a balanced performance with better accuracy than the smaller models and demand less computational resources than the larger models. They are ideal for applications where super-accurate responses are not needed and they are constraints in terms of the availability of computational resources.
- **Small Models:** These models provide lower accuracy than the other categorized models, making them less reliable. They are very resource-efficient and generate fast responses, compared to the other models. Smaller models are ideal in environments where accuracy is not important but quick responses and minimal computational resources are.

6.1.3 ***Practical Implications.*** The findings and the results of this research have provided valuable insights for the development of physiotherapist applications using LLMs. It has offered an in-depth analysis of different size models analyzing biomechanical running data. The study presented information about the trade-offs between the accuracy and computational requirements of these models and provided advice on what kind of environments they are suitable for.

6.2 Limitations

This subsection covers the challenges and the limitations of the used methodologies in the research, focused on the Text-To-SQL agent and model hallucinations

6.2.1 ***Challenges with the RAG System and SQL Integration.*** The proposed Retrieval-Augumented Generation (RAG) system integrated with SQL did not function as expected. There was an issue with the SQL model generating inaccurate queries, resulting in the data not being retrieved from the database.

To address this issue, the relevant data was instead added directly to the prompt of the user. While this approach does not give off a feeling of a real system, it still allows the biomechanical to be analyzed and processed from the LLMs.

6.2.2 ***Model Hallucinations.*** Model hallucination refers to when LLMs generate a response that has nothing to do with the question that the models were asked. This phenomenon impacted the evaluation of the models, specifically when accuracy was measured. The models sometimes produced responses that were incorrect or irrelevant to the question, making it difficult to evaluate the real accuracy.

6.3 Future Work

To further improve the findings and address the current limitations of this research, the study identified several areas for future work. These include developing realistic datasets, using fine-tuning techniques, and improving the Text-To-SQL Agent.

6.3.1 ***Expanding Data and Evaluation.*** By creating and developing realistic datasets to test the models, we can obtain better real-world insights. Using real-world data from wearable devices will allow us to test the models in real-world conditions, providing a more precise evaluation.

6.3.2 ***Fine-Tuning Model Training.*** The models have the potential to perform better by using fine-tuning techniques on specific biomechanical running datasets. This involves retraining parts of the model or adding new layers and training them with biomechanical data, helping them understand specific data patterns [21]. Fine-tuning has shown in studies, that it can provide more accurate results and relevant insights [15].

6.3.3 ***Improvement Solutions of Text-To-SQL Integration.*** The proposed system of this research could be improved by boosting the inaccuracies of the Text-To-SQL model. Potential methodologies that could enhance the Text-To-SQL Agent are trying different LLMs, redefining the database schema, and also trying different prompt techniques and styles. Testing the mentioned methodologies can benefit the proposed system by making it function as proposed.

7 ACKNOWLEDGEMENT

This study would not have been possible without the help of WearM.AI hardware and financial support. The access to high-end hardware by WearM.AI was crucial to evaluate and analyze the Large Language Models used in this research.

REFERENCES

[1] Xavier Amatriain. Prompt design and engineering: Introduction and advanced methods. *arXiv preprint arXiv:2401.14423*, 2024.

[2] Rabiul Awal, Le Zhang, and Aishwarya Agrawal. Investigating prompting techniques for zero-and few-shot visual question answering. *arXiv preprint arXiv:2306.09996*, 2023.

[3] Nilesh Barla. Deploying large nlp models: Infrastructure cost optimization. *https://neptune.ai/blog/nlp-models-infrastructure-cost-optimization*, 2024.

[4] G Bharathi Mohan, R Prasanna Kumar, P Vishal Krishh, A Keerthinathan, G Lavanya, Meka Kavya Uma Meghana, Sheba Sulthana, and Srinath Doss. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems*, pages 1–24, 2024.

[5] Paolo Bonato. Advances in wearable technology and applications in physical medicine and rehabilitation, 2005.

[6] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024.

[7] Yan Chen and Pouyan Esmaeilzadeh. Generative ai in medical practice: in-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26:e53008, 2024.

[8] Amr Elmeleegy. Demystifying ai inference deployments for trillion parameter large language models. *https://developer.nvidia.com/blog/demystifying-ai-inference-deployments-for-trillion-parameter-large-language-models/*, 2024.

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[10] Matevž Hribernik, Anton Umek, Sašo Tomažič, and Anton Kos. Review of real-time biomechanical feedback systems in sport and rehabilitation. *Sensors*, 22(8):3006, 2022.

[11] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

[12] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022.

[13] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Xing Li. The power of local llms: When private deployment outshine commercial giants like chatgpt. *https://www.linkedin.com/pulse/power-local-llms-when-private-deployment-outshine-giants-xing-li-ptoac/*, 2024.

[15] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. Data-efficient fine-tuning for llm-based recommendation. *arXiv preprint arXiv:2401.17197*, 2024.

[16] Dave Ver Meer. Number of chatgpt users and key stats. *https://www.namepepper.com/chatgpt-users*, 2024.

[17] Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. *arXiv preprint arXiv:2401.06775*, 2023.

[18] Partha Pratim Ray. Can llms improve existing scenario of healthcare? *Journal of Hepatology*, 80(1):e28–e29, 2024.

[19] Jiho Shin, Clark Tang, Tahmineh Mohati, Maleknaz Nayebi, Song Wang, and Hadi Hemmati. Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks. *arXiv preprint arXiv:2310.10508*, 2023.

[20] Mihaela Tomova, Martin Hofmann, Constantin Hütterer, and Patrick Mäder. Assessing the utility of text-to-sql approaches for satisfying software developer information needs. *Empirical Software Engineering*, 29(1):15, 2024.

[21] Kushala VM, Harikrishna Warrier, Yogesh Gupta, et al. Fine tuning llm for enterprise: Practical guidelines and recommendations. *arXiv preprint arXiv:2404.10779*, 2024.

[22] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018.

[23] Michael Weinberger. The rise of local llm-based ai applications: A new opportunity for cost- and privacy-conscious businesses. *https://www.linkedin.com/pulse/rise-local-llm-based-ai-applications-new-opportunity-cost-michael-3czec/*, 2024.

[24] Richard W Willy. Innovations and pitfalls in the use of wearable devices in the prevention and rehabilitation of running related injuries. *Physical Therapy in Sport*, 29:26–33, 2018.

[25] Ali K Yetisen, Juan Leonardo Martinez-Hurtado, Barış Ünal, Ali Khademhosseini, and Haider Butt. Wearables in medicine. *Advanced Materials*, 30(33):1706910, 2018.

[26] Angelo Ziletti and Leonardo D'Ambrosi. Retrieval augmented text-to-sql generation for epidemiological question answering using electronic health records. *arXiv preprint arXiv:2403.09226*, 2024.