# Large Language Model-Based Sport Coaching System Using Retrieval-Augmented Generation and User Models

CRISTIAN COMENDANT, University of Twente, The Netherlands

Large language models (LLM) have advanced at a high pace in recent years. By using big datasets, they are capable of understanding and generating human-like language. Models like OpenAI's generative pre-trained transformer (GPT) use deep learning techniques to produce relevant responses in different fields. These models generate natural language applicable for use by artificial agents, such as social robots or chatbots. The challenge now is to personalize these responses for individual users. For this, user models must be used to capture user preferences and behaviors and offer a solution to this challenge.

This study designed two LLM swimming coaching systems, both incorporating user models, with one system additionally utilizing a Retrieval-Augmented Generation (RAG) system. RAG increase the quality of the output of a LLM by leveraging contextual or real-world knowledge. Over a three-week period, these systems provided guidance and feedback to improve the swimming performance.

Our results showed that participants using the LLM system with RAG significantly enhance their freestyle stroke technique, as evidenced by a reduction in the number of strokes needed to swim a 25 meters lap. This demonstrates the potential of integrating LLMs with RAG and user models to improve personalized coaching in sports.

Additional Key Words and Phrases: Retrieval-Augmented Generation, AI Coaching System, Personalization

## 1 INTRODUCTION

### 1.1 Motivation

In the world of sports coaching, nothing is as challenging as giving personalised advice to athletes. This means taking into account the unique needs, goals, and characteristics of each individual athlete. It consists of physical condition, skill level, athletic achievement history, personal preferences, or health conditions. This approach improves performance, prevents injuries, and maintains motivation. In essence, personalised training advice is a key part of every athlete's career. [7]

Nonetheless, traditional coaching methods often face limitations in delivering personalized recommendations because these methods rely mainly on training programs that have limited access to detailed performance data and are restricted by the time and effort required from a coach who manages multiple athletes. In addition, they are highly dependent on subjective judgments, leading to potential inconsistencies and biases. [3]

Large language models (LLMs) can represent the perfect solution to these challenges by providing data-driven insights, scalability, and personalized feedback[14] since they can analyse vast amounts of data to provide understanding, manage personalized guidance for multiple athletes simultaneously, and offer real-time feedback and adjustments. However, LLMs often generate hallucinations. One solution to increase the factual correctness of the generated responses is to use retrieval-augmented generation (RAG) [19]. RAG is a technique for improving the reliability and precision of generative AI models by fetching factual information from external sources. [9]

The final aim of the paper is summarized in the following research question:

**RQ: To what extent does a large language model, with and without retrieval-augmented generation, improve freestyle stroke technique, within a 3-week time frame?**

The subsequent sections are organized as follows: Section 2 describes the experimental setup, detailing the tasks required from both the system and the participants, participant interaction methods, and data collection processes. Section 3 outlines the study's results, emphasizing significant insights from the collected metrics and participant feedback. Section 4 presents an analysis and interpretation of these results, exploring the implications and acknowledging the limitations of the study. Finally, Section 5 concludes the paper by summarizing the main findings and proposing directions for future research and enhancements.

### 1.2 Related Work

Previous research in sports coaching and artificial intelligence (AI) has established the foundation for the creation of individualized coaching programs for athletes. To increase athlete performance, traditional coaching approaches have depended on a real coach who is responsible for individual training for each athlete based on their goals and abilities. This takes plenty of time. Recent advances in AI, particularly LLMs, have opened new opportunities for personalized coaching at scale. One study [12] explored the use of LLMs in sports coaching, demonstrating their ability to generate personalized training plans based on users' schedules and goals. The study highlighted the potential of AI coaching systems to adapt to individual needs and preferences, leading to improved training outcomes.

Another article [13] researched how a personalized AI coach can be used to provide guidance and feedback to cyclists based on their individual capabilities and training. Researchers observed that the system performed "equally to or better than the control training plans in 14 and 24 week training periods," being evaluated as better in 4 out of 5 test components, including training load, resting time quantity, resting time distance, and efficiency. They reported "a higher statistical difference in the results of the experts' evaluations between the control and virtual coach training plans" favouring the virtual coach.

In the sphere of personalized interaction with LLMs, some studies [18] [2] investigated ways to tailor unique responses for individual users. The first study, [18], explores fine-tuning and zero-shot reasoning approaches for subjective text perception tasks.

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

Another study investigated the application of user models in AI-based coaching systems, demonstrating that personalized input significantly benefits the effectiveness of the coaching. The study found that personalization improved user engagement, training effectiveness, and overall satisfaction by tailoring guidance to individual needs and preferences, resulting in more accurate and effective coaching outcomes. [2]

Even though these studies have made important improvements in the area of AI-powered sports coaching systems, there is still a gap for research that focuses specifically on swimming or on using RAG. While personalization, fine-tuning, and zero-shot approaches have been successfully tested, the significance of RAG in enhancing AI coaching systems remains underexplored. This research aims to create a swimming coaching system for swimmers that helps them improve their freestyle stroke technique and performance outcomes by using RAG and user models.

## 2 METHOD OF RESEARCH

### 2.1 Experimental Setup

Participants interacted with a specialized tool designed for this study. The setup included an underwater camera (GoPro Hero 12) and an external webcam. The underwater camera communicated with the laptop via a Wi-Fi signal.[1]

Both cameras captured images of participants swimming, which were used as input for the multi-modal large language system to provide feedback aimed at improving their swimming technique.

Participants viewed the feedback on the laptop screen and interacted with the platform in two ways: through speech or text. For verbal communication, they used waterproof headphones to ask questions and receive auditory feedback from the AI system. This allowed for real-time interaction without the need to leave the pool. For textual input, participants used the laptop's interface, allowing them to type queries or provide feedback. (Figure 1)

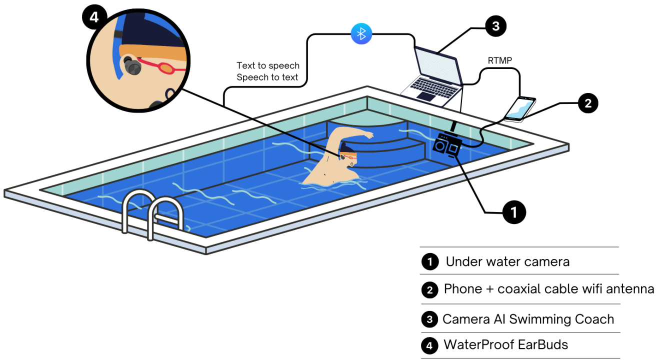


Fig. 1. System overview

### 2.2 LLM-based Swim Coach

*2.2.1 System Overview.* The AI Coaching System, an AI chatbot, was developed to facilitate this research by providing personalized swimming guidance. There were two assistants: LLM integrated with a user model and RAG (LLMUR) and LLM integrated only with a user model (LLMU). The core functionalities of the system were identical for both groups, with a significant distinction for LLMUR: the ability to augment the AI's knowledge base. Participants in the LLMUR could upload relevant PDF files or automatically incorporate videos from YouTube channels into the vector store.

The system is primarily divided into two main components: the OpenAI server and the local application. The local application consists of several processes: image processing, video processing and speech processing. For the transcription of audio files, the Whisper AI model[2] was utilized to understand the video content, the LLaVA model[3] summarized the frames that were split at one-second intervals. On the OpenAI server side, two assistants were implemented. One of these assistants was integrated with a RAG system. Both assistants used the generative pre-trained transformer-4o (GPT-4o) model to generate responses (Figure 2).

*RAG Model:* The AI responses are aimed to be improved by using the RAG technique, which incorporates data that is not present in the GPT-4o LLM's general knowledge base. When receiving a user query, the OpenAI Embedding API initially transforms it into a query vector. The query vector is then used to retrieve relevant information from a vector store containing preindexed embeddings of document chunks. The vector store returns the most relevant chunks based on their similarity to the query vector. The GPT-4o model receives these chunks, which give it extra context. With this data, the LLM model may return replies that are more precise and appropriate. By using updated data, this method tries to reduce the risk of hallucination and repetition. [5]

*User Model:* Each participant completed a questionnaire during their first training session. They were asked to provide their age, weight, height, swimming level, primary goals, current training regimen, and any health conditions. This information was sent to the assistant only once and was used as a context by the AI to personalize the training sessions. The assistants were instructed to always remember the user model (see Appendix A and B).

#### *2.2.2 System Functionality.*

*Data Upload and Processing:* Participants in the RAG group had the ability to upload offline and online data. The aim was to create a fully customizable, up-to-date database. The main challenge was determining the sources that have a constant flow of updated data, which a normal LLM would not normally have access to. One solution was to link a YouTube channel into the vector store, because there is always new data which is uploaded and a standard LLM was not trained on. To make these videos understandable for LLM, all videos had to be converted to text. This is because a vector store understands only text data, relying on text embeddings, which are high-dimensional vector representations of text.

Furthermore, this method allowed participants to add data from their preferred swimming coaches or idols, providing up-to-date information and a motivational component. This allowed participants

[1] Stable connection was ensured through an insulating coaxial cable.

[2] https://github.com/openai/whisper

[3] https://llava-vl.github.io/

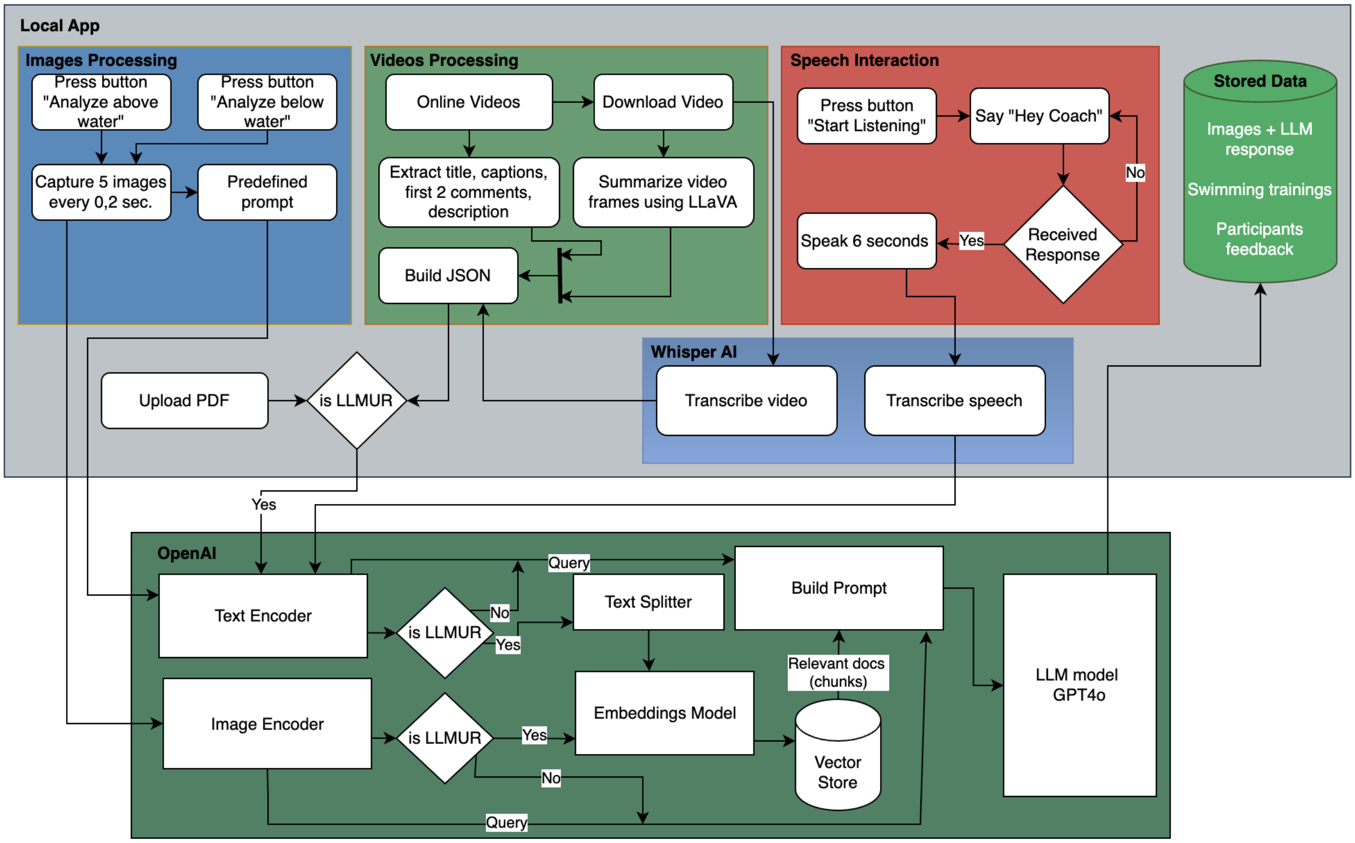


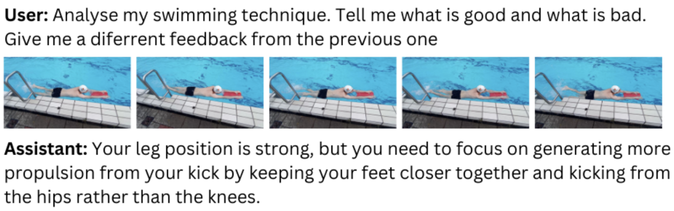
Fig. 2. System overview

to model their training routines after those of their role models, adapted to their individual power, speed, and endurance levels.

Subsequent steps have been undertaken to develop the architecture:

- **Video Processing:**
  - Participants uploaded a YouTube channel ID, and the system retrieved the first 10 videos from the channel.
  - Videos were downloaded locally, and metadata such as title, video URL, description, captions, and the first two comments were extracted.
- **Transcription and Visualization:**
  - Videos were transcribed using Whisper AI to convert speech to text. This allowed GPT-4o to refer users to specific parts of the video containing the most relevant information, preventing users from watching the entire video.
  - Videos were split into frames, and each second, Llava summarized the images to extract key visual details.
- **Data Integration:**
  - The processed text and visual data were saved in a JSON file, converted to embeddings, and stored in the vector store, enabling the LLM to retrieve and use the most relevant information for swimming training guidance.

*Image Analysis:* During the experiment, the tool semi-automatically took pictures of the participants. One shot consisted of five pictures, captured with a delay of 0.2 seconds, mimicking a short video to capture participants' movement in the water. These images were analysed by the GPT-4o after the user instruction was given. The assistant responded with possible adjustemts of the participant's swimming technique. (See Figure 3)


**User:** Analyse my swimming technique. Tell me what is good and what is bad. Give me a diferrent feedback from the previous one

**Assistant:** Your leg position is strong, but you need to focus on generating more propulsion from your kick by keeping your feet closer together and kicking from the hips rather than the knees.

Fig. 3. An Example of image input and LLM output.

Subsequently, the output was saved in a PDF file, which included the images and the assistant's output. The PDFs were necessary for later use by experts to analyse the accuracy and factual correctness of the data.

*2.2.3 Prompt Engineering.* The prompts were designed to generate the most relevant responses from the LLM. The application utilized two assistants, LLMU (see Appendix B) and LLMUR. The instructions for both assistants were identical, with one small difference: the LLMUR assistant's prompt included an additional rule, "Retrieve relevant data from your documents."

*2.2.4 Speech Interaction.* The user could interact with the AI via speech-to-text (STT). To activate this function, the user first pressed the "Start Listening" button on the tool. After that, to start asking questions to the AI, the user had to say "Hey coach." Once the bot responded, the user had 6 seconds to speak before their speech was converted to text by Whisper AI and sent to the GPT-4o model.

It is important to note that the user could ask and receive feedback from the AI via STT only when they were close to the laptop, as the Bluetooth connection was limited to a range of approximately 5 meters. This meant that while in the water, the user's interaction with the AI through STT was constrained by this distance.

*2.2.5 Human - AI Interaction.* Participants interacted with the chatbot during their sessions to ask questions, seek clarifications, and request modifications to their training plans. This interaction was intended to simulate a real-world coaching scenario in which athletes could receive immediate personalized feedback.

## 2.3 Conditions (Independent Variables)

The independent variable in this study is the type of system used by participants. There are two types: LLMUR and LLMU. The division was made to compare how the presence or absence of external information retrieval affects the personalization of the training, overall participant's performance, and trustworthiness.

## 2.4 Measurements (Dependent Variables)

The dependent variables, which are directly affected by the independent variable, include both objective and subjective measurements:

*2.4.1 Objective Measurements.*

- **Time:** The duration taken by participants to swim the 25 meters distance, measured at the end of the first and last training sessions to assess improvements over time.
- **Distance per Stroke:** The distance covered by participants per stroke, calculated using the formula:

$$\text{Distance per Stroke} = \frac{\text{Total Distance}}{\text{Number of Strokes}}$$

  This metric was also recorded at the end of the first and last training sessions to determine improvements in swimming efficiency.

*2.4.2 Subjective Measurements.*

- **User Feedback:** Participants' feedback after each training session includes evaluating the AI-generated responses on a 10-point scale, using the 18-item Physical Activity Enjoyment Scale (PACES) [1] on a 7-point scale to determine enjoyment levels, and the Borg CR-10 scale [17] to measure perceived exertion. At the end of all nine training sessions, participants completed the Working Alliance Inventory - Short Revised (WAI-SR) form on a 5-point scale [6].The results were used to compare the long-term impact of using LLMU and LLMUR, focusing on goal agreement related to improving stroke technique.
- **Expert Evaluation:** Assessments provided by two experts on the factual correctness and relevance of all AI-generated messages. These evaluations helped to ensure the accuracy and usefulness of the feedback given to participants.

## 2.5 Participants

Ten participants were divided using random sampling without replacement into two groups to evaluate the effectiveness of an LLM with and without RAG in improving swimming technique. The first group interacted with an LLM integrated with both a user model and RAG, while the second group interacted with an LLM integrated only with a user model. Participants were selected from our private network. However, the selection was based on specific inclusion criteria: male individuals who train three times a week, with each training session lasting 45 to 60 minutes. Individuals who did not meet these criteria were excluded from the study. The participants were aged between 20 and 23 years, with a mean (M) age of 21 years and a standard deviation (SD) of 0.94. 70% of participants were beginners, while 30% were intermediate swimmers. All participants had prior solid experience interacting with AI models.

Participants signed a consent form before taking part in the study. Additionally, participants were compensated for their time and effort by offering a fruit to regain energy. This study was approved by the Ethics Committee of the University of Twente.

## 2.6 Procedure

In total, participants were required to complete 9 training sessions, all of which took place during the summer.

Due to time constraints, participants took three training sessions consecutively, with a 5-minute break between each session. After these consecutive sessions, they had to take at least a 24-hour break before starting the next training session to ensure enough recovery. On average, the resting time mean between three consecutive sessions was 3 days (SD = 1.39), and the time slots were decided based on their availability.

Both groups of people used the same application, with a critical distinction: participants in the RAG group were granted an additional feature that enabled them to input relevant information, thereby personalizing their training experience.

Prior to the commencement of the training swimming sessions, participants received a briefing outlining the study's objectives, their expected involvement, and detailed instructions on system operation. Throughout the training sessions, participants had complete freedom to interact with the AI. They were told to adhere to the guidance provided by the AI and had the autonomy to request modifications to their training. After each training session, participants were required to complete a feedback form.

## 3 RESULTS

The results were analyzed using Python language [15]. To compare the different conditions, we used the parametric independent samples t-test [8] and the nonparametric Wilcoxon rank sum test

(WRST) [16]. The choice of statistical test was based on the following criteria: the t-test was used if the data set was normally distributed (Shapiro-Wilk test [4]) and homogeneity of variance (Levene test [10]) was met.

During the research, the researcher captured and the GPT-4o model analyzed a total of 1,030 photos, resulting in 206 PDF files, each containing five images. Following each training session, participants completed a feedback form, resulting in a collection of 90 feedback forms. Performance metrics were recorded twice for each participant, once at the beginning and once at the end of the training period.

## 3.1 Quantitative Data

## 3.2 Subjective Measurements

To assess the effectiveness of the system, subjective survey measurements were used, consisting of user feedback.

The overall AI evaluation metrics are depicted in Table 1. To obtain these values, the mean metric of each individual participant was computed. The "Personalization" metric resulted in significantly higher scores for LLMUR (M = 7.27 and SD = 1.55) compared to LLMU (M = 5.15, SD = 0.63), t(8) = 2.83 and $p < 0.05$. All other AI evaluation metrics revealed no significant differences between the two groups.

The overall PACES evaluation is depicted in Figure 4. The general scores between LLMUR (M = 5.52, SD = 0.42) and LLMU (M = 5.25, SD = 0.47) were not significantly different, t(34) = 1.78, $p > 0.05$. However, the fun metric of the PACES scale was significantly higher for LLMUR (M = 5.37 and SD = 0.62) compared to LLMU (M = 4.57 and SD = 0.32), t(8) = 2.57 and $p < 0.05$. All other metrics on the PACES scale did not show significant differences between the two groups.

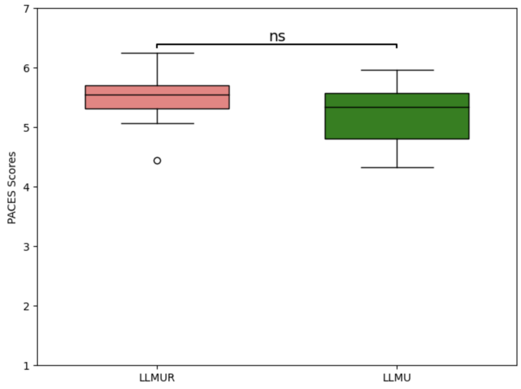


Fig. 4. Comparison of Overall PACES Evaluation Metrics

The personalization metric over nine training sessions between LLMUR and LLMU is presented in Figure 5. LLMUR score was significantly higher (M = 7.26, SD = 1) compared to LLMU (M = 5.15 and SD = 0.76), t(16) = 5.01 and $p < 0.001$.

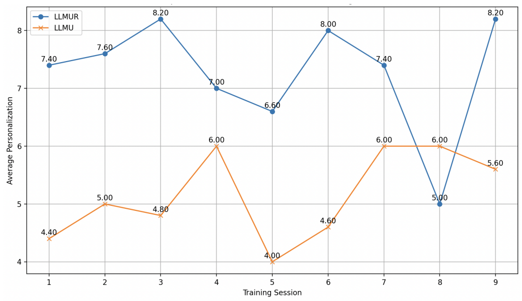


Fig. 5. Comparison of Personalization Over 9 Training Sessions

The motivation metric over nine training sessions between LLMUR and LLMU is illustrated in Figure 6. LLMU score was significantly higher (M = 6.13, SD = 0.65) compared to LLMUR (M = 5.31 and SD = 0.90), W = 10 and $p < 0.05$.

To find personalization and motivation values during nine training sessions, the mean for each metric was computed per training session. This involved taking the feedback given for that metric by each participant at each session, then averaging those values.

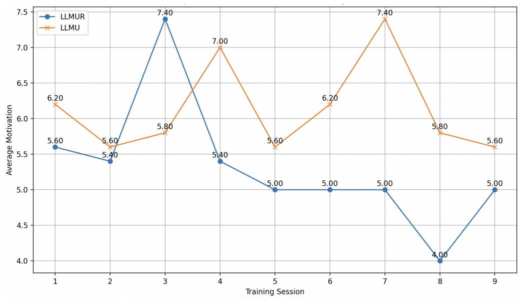


Fig. 6. Comparison of Motivation Over 9 Training Sessions

The WAI-SR results are depicted in Table 2. The LLMUR shows a higher reliability on the "Task" scale ($\alpha = 0.77$) compared to LLMU $\alpha = 0.65$). Additionally, a t-test indicates a significantly higher value for LLMUR (M = 3.81, SD = 0.75) compared to LLMU (M = 3.61, SD = 1.35), t(14) = 2.74 and $p < 0.05$. All other WAI-SR metrics did not show significant differences between the two groups.

The Borg CR-10 scale results showed that the perceived exertion scores for LLMUR (M = 6.89, SD = 0.72) and LLMU (M = 6.14, SD = 1.22) were not significantly different, W = 18 and $p > 0.05$.

## 3.3 Objective Measurements

In addition to subjective measurements, we used objective quantitative measurements to assess the participants' performance.

Participants in the LLMUR group (M = 1.78, SD = 1.06) and the LLMU group (M = 1.63, SD = 1.41) did not show a statistically significant difference in time improvement, W = 15, $p > 0.05$ (Figure7). However, the stroke count reduction (distance per stroke) was significantly higher for LLMUR (M = 1.40, SD = 0.55) compared to LLMU (M = 0.00, SD = 1.22), W = 22, $p < 0.05$ (Figure 8).

Table 1. Mean [M], Standard Deviation [SD], and statistics for AI Evaluation Metrics

| Measurement | LLMUR *M* (SD) | LLMU *M* (SD) | Statistics |
|---|---|---|---|
| Usefulness | 7.60 (.68) | 7.49 (.79) | t(8) = .23, p = .82 |
| Satisfaction | 7.18 (.83) | 7.27 (.95) | t(8) = -.17, p = .87 |
| **Personalization** | **7.27 (1.55)** | **5.15 (.63)** | **t(8) = 2.83, p = .02** |
| Understanding * | 7.00 (1.24) | 6.38 (.55) | W = 17.50, p = .34 |
| Motivation | 5.31 (1.77) | 6.17 (.68) | t(8) = -1.01, p = .34 |
| Intensity | 7.09 (.57) | 6.56 (.95) | t(8) = 1.07, p = .31 |
| Volume | 6.84 (.82) | 6.49 (.92) | t(8) = .64, p = .54 |
| Resting Time | 5.93 (.81) | 5.60 (1.04) | t(8) = .55, p = .59 |

* Understanding metric is not normally distributed and therefore WRST

Table 2. Mean [M], Standard Deviation [SD], Cronbach's alpha [$\alpha$], and statistics for WAI-SR scale

| WAI-SR Scales | LLMUR (*N* = 5) | | LLMU (*N* = 5) | | Statistics |
|---|---|---|---|---|---|
| | *M* (SD) | $\alpha$ | *M* (SD) | $\alpha$ | |
| Bond | 3.33 (1.18) | .74 | 3.36 (1.01) | .76 | t(14) = -.06, p = .96 |
| **Task** | **3.81 (.61)** | **.77** | **2.95 (.94)** | **.65** | **t(14) = 2.74, p = .02** |
| Goal * | 4.03 (.75) | .77 | 3.61 (1.35) | .90 | W = 35.5, p = .70 |
| Total * | 3.72 (.87) | .85 | 3.34 (1.12) | .88 | W = 42.5, p = .26 |

* Goal and Total metrics are not normally distributed and therefore WRST

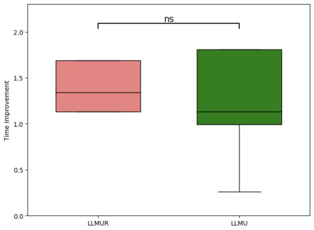


Fig. 7. Time Metric

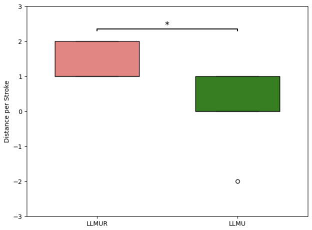


Fig. 8. Distance per Stroke Metric

### 3.4 Qualitative Data

The user model had different impacts on the personalized training experience for participants. Beginners and intermediate swimmers received similar guidance, but the volume and intensity were slightly higher for the intermediate group.

All participants' conversations and the images analyzed by the model were shared with two experts: the head coach of the WS Twente Team and the second national swimming coach of the Republic of Moldova. The experts were instructed to review all the provided data. They were asked to assess the correctness and usefulness of the system and to provide their feedback. To ensure that the coaches' opinions were not biased against one particular system, the LLMUR was presented as "Group 1" and the LLMU was presented as "Group 2".

The experts' responses were recorded through written feedback forms. After their examination, it was highlighted that both assistants (LLMUR and LLMU) could be useful for coaches by giving insights that they might omit. The tool is suitable for beginner and intermediate swimmers, but not for competitive swimmers. The LLMUR instructions were more descriptive because they provided valid and relevant Youtube videos. The tool provided incorrect observations despite linguistically correct advice. (See Appendix C)

By taking into account their feedback we counted all the incorrect observations and found that errors occurred 25% of the time. One example of an inaccuracy was when the AI received images where only the underwater body was visible, but gave instructions on how to keep the elbow above the water.

## 4 DISCUSSION

In this study, we investigated the impact of a large language model using user models and a RAG system on human swimming performance. Our goal was to determine if personalized training provided by LLMUR leads to better outcomes compared to LLMU over a three-week period. We implemented two distinct assistant systems, which participants used to improve their swimming techniques.

Our results, as presented in the research paper, show a significant reduction in stroke count for participants who used the LLMUR system compared to those who used the LLMU system. This indicates that the LLMUR system, which allows users to augment the AI's knowledge base, effectively improves swimming technique. As noted by the experts, the swimming trainings were similar in complexity for both groups: LLMUR and LLMU. However, the key factor that significantly helped the LLMUR group develop a better swimming foundation was the inclusion of relevant videos that allowed participants to visualize certain aspects of the training. Therefore, the methodological assistance was ampler. Both groups showed minor improvements in time metric, but the difference was not statistically significant. This is likely because, for beginner swimmers, improvement in time is more closely related to their strength than technique alone. However, improvements in technique, as evidenced by reduced stroke counts, are expected to lead to time reductions over a longer period.

After analysing the WAI-SR results, we identified higher reliability on the "Task" scale for the LLMUR compared to the LLMU. This indicates that LLMUR participants agreed with the specific tasks assigned during the training. This occurred primarily because participants knew the sources of the training materials and were more inclined to trust the factual accuracy of the training and feedback generated.

A significantly higher difference was discovered in the overall personalization metric for LLMUR compared to LLMU. This occurred mainly because the LLMUR database was augmented with swimming materials that the normal LLMU did not have access to, providing a broader range of information.

Interestingly, while the personalization over all nine training sessions was significantly higher for LLMUR, the motivation over the duration of the nine training sessions was significantly higher for the LLMU group. LLMU group demonstrated higher levels of interest and engagement. However, their technique was poorer due to the system's inability to provide sufficient instructional material to enhance their skills. Thus, "... positive thinking may not have the same impact on performance as task-relevant content ..." [11].

The perceived exertion levels, measured using the Borg CR-10 scale, were similar between the two groups. This result matches our expectations, given that the training programs were designed primarily for beginner swimmers, as indicated in Section 2.5 on Participants. Both assistants provided appropriate training for this skill level, resulting in comparable exertion levels.

We found no significant differences in overall PACES results between the two groups, LLMUR and LLMU. However, the fun metric was significantly higher for the LLMUR group. This could be due to the interactive features of the LLMUR system, which allowed users to personalize their training experience and view relevant YouTube videos, making the learning process more engaging.

Similar to previous studies, which have shown the benefits of personalized AI in improving user engagement and training outcomes, our results demonstrate that the inclusion of external, verifiable information sources significantly boosts the user trust and performance improvements in sports coaching applications.

### 4.1 Limitations

Several limitations were encountered during this study, which may have impacted the results and their generalizability.

The small number of pictures captured during some sessions was due to a slow Ethernet connection. These problems affected the connections between the API and the local system, as well as the live transmission from the underwater camera.

It is hard to say the accuracy of these data because the sample population consisted only of 10 people. A higher number of participants might give more accurate results.

The study included only male participants due to a scarcity of female. Including women in the research would have provided a more holistic picture of the data for both genders.

These limitations should be considered when interpreting the results of this study. Future research could benefit from addressing these technical and operational challenges to ensure more consistent data collection and more vigorous analysis.

### 4.2 Future Work

For future applications implementing RAG, it is advisable to modify the current system to automatically update the vector store whenever the content creator uploads a new post or video. This will ensure that AI responses remain current and diminish repetitions and hallucinations.

Additionally, linking the RAG system with other social media platforms could improve the relevance and accuracy of the AI's responses by integrating a wider array of up-to-date content. This could facilitate more personalized and effective training recommendations.

Future research could also involve replicating this experiment with competitive swimmers. Improving technique for competitive swimmers requires a more sophisticated model capable of fine-tuning minor technical aspects, such as arm elbow angle or foot position. Furthermore, more data about the swimmer would be required, including medical records, sleep tracking (quality and quantity), and dietary information. These enhancements can be achieved by integrating RAG with real-time databases, such as those provided by fitness tracking devices.

This integration would allow for a more holistic approach to training, incorporating real-time health and performance data to deliver highly personalized and accurate guidance.

It will be useful for future research to make the device completely autonomous. To accomplish this, the camera might be strategically positioned and coded to precisely follow the swimmer when swimming into the water.

## 5 CONCLUSION

Answering the research question, we find that people using LLMUR significantly increased their freestyle stroke technique, by reducing the number of strokes. The LLMUR provided a wider range of materials which improved the learning process.

The outcomes demonstrated that compared to LLMU, LLMUR was able to offer more personalized training. Even if the training methods used by the two experts were general in nature, LLMUR's explanation was more visual because it provided video materials extracted from the vector store. Users tended to have higher reliability to the "tasks" provided by the LLMUR because they knew the source of the information which in the end resulted in greater trust. LLMU managed to obtain higher scores in the "motivation" metric over the duration of nine training sessions, meaning that people reported a higher level of interest. Despite the significantly higher "personalization" metric for the LLMUR across all sessions, LLMU performance turned out to be weaker because there was not enough content in the system to help them improve their skills.

It is essential to note that the feedback and guidance provided by both LLMUR and LLMU are beneficial only for beginner and intermediate swimmers. The current technology is not yet suitable for advanced swimmers.

Future research should address the technical limitations, include a more diverse participant sample, allocate a longer period of time to conduct the research and explore the application of RAG and user models in competitive swimming.

## ACKNOWLEDGMENTS

The author expresses gratitude to his supervisor, Sebastian Schneider, for his help and support during the research project. The author also thanks all participants who took part in the study, as well as the experts Renato Markovinovic and Sergiu Postică.

## REFERENCES

[1] Cheng Chen, Susanne Weyland, Julian Fritsch, Alexander Woll, Claudia Niessner, Alexander Burchartz, Steffen CE Schmidt, and Darko Jekauc. 2021. A short version of the physical activity enjoyment scale: development and psychometric properties. *International Journal of Environmental Research and public health* 18, 21 (2021), 11035.

[2] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2023. When large language models meet personalization: Perspectives of challenges and opportunities. *arXiv preprint arXiv:2307.16376* (2023).

[3] C Cushion. 2013. Coaching and coach education. In *Science and Soccer*. Routledge, 211–229.

[4] R Dudley. 2023. The Shapiro–Wilk test for normality.

[5] Paulo Finardi, Leonardo Avila, Rodrigo Castaldoni, Pedro Gengo, Celio Larcher, Marcos Piau, Pablo Costa, and Vinicius Caridá. 2024. The Chronicles of RAG: The Retriever, the Chunk and the Generator. *arXiv preprint arXiv:2401.07883* (2024).

[6] Robert L Hatcher, Karin Lindqvist, and Fredrik Falkenström. 2020. Psychometric evaluation of the Working Alliance Inventory—Therapist version: Current and new short forms. *Psychotherapy Research* 30, 6 (2020), 706–717.

[7] Scott R Johnson, Pamela J Wojnar, William J Price, Timothy J Foley, Jordan R Moon, Enrico N Esposito, Fred J Cromartie, et al. 2011. A coach's responsibility: Learning how to prepare athletes for peak performance. *The Sport Journal* 14, 1 (2011), 1–14.

[8] Tae Kyun Kim. 2015. T test as a parametric statistic. *Korean journal of anesthesiology* 68, 6 (2015), 540–546.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[10] Brian B Schultz. 1985. Levene's test for relative variation. *Systematic Zoology* 34, 4 (1985), 449–456.

[11] Maureen L Shewchuk. 1985. *Aspects of thought factors and their effects on performance in swimming.* Ph. D. Dissertation.

[12] Donghoon Shin, Gary Hsieh, and Young-Ho Kim. 2023. PlanFitting: Tailoring Personalized Exercise Plans with Large Language Models. *arXiv preprint arXiv:2309.12555* (2023).

[13] Alessandro Silacci, Redha Taiar, and Maurizio Caon. 2020. Towards an AI-based tailored training planning for road cyclists: a case study. *Applied Sciences* 11, 1 (2020), 313.

[14] Xinming Tu, James Zou, Weijie Su, and Linjun Zhang. 2024. What Should Data Science Education Do With Large Language Models? *Harvard Data Science Review* 6, 1 (jan 19 2024). https://hdsr.mitpress.mit.edu/pub/pqiufdew.

[15] Guido VanRossum and Fred L Drake. 2010. *The python language reference.* Vol. 561. Python Software Foundation Amsterdam, The Netherlands.

[16] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution.* Springer, 196–202.

[17] Nerys Williams. 2017. The Borg rating of perceived exertion (RPE) scale. *Occupational medicine* 67, 5 (2017), 404–405.

[18] Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized Large Language Models. *arXiv preprint arXiv:2402.09269* (2024).

[19] Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396* (2023).

## A USER MODEL PROMPT

*"This is my information: Age: {age} Gender: {gender} Height: {height} cm Weight: {weight} kg Swimming Experience and Goals: Skill Level: {skilllevel} Primary Goals: {goals} Current Training Regimen: Do you have a coach or trainer? {hascoach} If yes, how often do you train with them? {trainingfrequency} Preferences and Habits: Preferred Swim Time: {swimtime} Health and Fitness Background: Health Conditions: {healthconditions} "*

## B LLMU ASSISTANT INSTRUCTIONS

*"You are a swimming expert coach. You provide feedback and guidance to improve the freestyle stroke. If the user asks questions that are not related to swimming, do not respond. Always remember the first message from the user as that one defines their user model. This information is essential for making new training plans and measuring progress. You will provide personalized swimming training based on the user's information only when requested by the user, for example, 'Start the training.' Your answer is always short and to the point, 1-2 sentences. Focus only on one detail. If you are asked to provide a video, it should not be longer than 15 seconds. Only the training description is longer in text. Don't just give tips; also make personalized swimming training based on the user information, specifying the distance to swim and time interval. For each exercise, the user will give you some pictures so that you can also analyse their technique. Each training session lasts for 15 minutes. In total, a participant will do 9 training sessions. After each session, ask for user feedback. Do not respond if the images are not related to swimming."*

## C EXPERT TRANSCRIPTS

### C.1 Feedback - expert one

*"After analysing all the trainings, I can conclude that the methodology and trainings provided in both categories are suitable for beginner and intermediate swimmers. They offer a good foundation on how to swim. When comparing Group 1 and Group 2, I observed that sometimes the guidance provided was incorrect. However, while the instructions from both groups were almost the same in terms of complexity, and relevance, the feedback from Group 1 was a bit more descriptive and helpful. For both categories, the assistant tried to provide YouTube videos to visualize the freestyle stroke. However, most of the URLs provided in Group 2 were invalid or irrelevant. In contrast, Group 1 did a better job by providing valid YouTube videos, which were most of the time relevant to the training provided or the questions asked by the participants. While I understand that this research targeted beginner swimmers, the application is not yet suitable for competitive swimmers. However, it might be useful for helping a coach during training by providing insights that a coach might omit."* by Sergiu Postica

### C.2 Feedback - expert two

*"A majority of the participating swimmers can be characterized as nonexperienced swimmers, with little to no previous exposure to either technical or (swim) endurance type of training; the same applies to group 2 as well. With that in mind, the advice provided to the swimmers by the "Assistant," being general in nature, might (at this stage of the methodology development) turn out to be either a useful addition to an (experienced) coach's tools arsenal or, otherwise, a good starting guidance to a self-coaching beginner swimmer. The more experienced/skilled the swimmer, the more 'nuances' in evaluating his/her technique are necessary, which will always require an experienced 'coaching eye.' That being said, regarding the 'accuracy/correctness/quality' of the advice provided by the Assistant and basing my judgment on the corresponding images of the swimmer performing their swim tasks, it appears to me that the Assistant in a number of cases' sees things incorrectly, while providing linguistically correct (and seemingly logical) observation/advice. The same conclusions/observations apply here as for group 1. An additional observation would be that the advice provided to the swimmers in the corresponding groups seems to be 'somewhat different in nature', as if the two groups were 'trained' using different information/data sources."* by Renato Markovinovic

## D APPENDIX D

During the preparation of this work , the author used Grammarly[4] to improve the readability of the work. After using the tool, the author reviewed and edited the content as needed and takes full responsability for the content of the work.

[4] https://www.grammarly.com/