Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 9 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i09.55653

PAPER

Development of a Virtual Assistant Based on LLMs for the Knowledge Domain in Biomedical Metrology

Yamilet Carreon, Miguel Chicchon(✉)

Universidad Tecnológica del Perú, Lima, Perú

c11201@utp.edu.pe

ABSTRACT

Accurate measurements are essential for effective diagnosis and treatment in the healthcare sector. However, there is limited training in biomedical metrology for Biomedical Engineers, which may hinder their performance. This study evaluated the knowledge of large language models (LLMs) in biomedical metrology to develop a specialized virtual assistant that supports these professionals. The effectiveness of the LLMs was assessed based on the accuracy and coherence of their responses using the CBET Certification exam and metrics such as Rouge-L, F1 score, and cosine similarity. The Llama 3.2-3B Mini model, optimized with retrieval-augmented generation (RAG), showed an increase in the F1 score from 0.402 to 0.526, a Rouge-L score of 0.497, and a cosine similarity of 0.657, demonstrating its ability to generate relevant and accurate responses. The developed virtual assistant represents a promising tool for improving the training and performance of biomedical engineers, ensuring access to precise and reliable information, thereby strengthening safety in the healthcare sector. Our source code is publicly available at https://github.com/yamilet2662/assistant_biome.

KEYWORDS

virtual assistant, large language models (LLMs), retrieval-augmented generation (RAG), hugging face spaces, biomedical metrology

1 INTRODUCTION

Biomedical metrology is an essential discipline in healthcare, as it ensures the accuracy and reliability of measurements used for the diagnosis and treatment of patients. Incorrect measurements can lead to inaccurate diagnoses, inadequate treatments, and ultimately, severe consequences for patient health [1]. However, the lack of specific training in this area for biomedical engineers represents a significant challenge, especially in regions where biomedical metrology education is still in its early stages and is not sufficiently integrated into academic programs [2]. This deficiency affects the quality of healthcare, as professionals lack the necessary knowledge to interpret the data provided by medical equipment correctly.

Carreon, Y., Chicchon, M. (2025). Development of a Virtual Assistant Based on LLMs for the Knowledge Domain in Biomedical Metrology. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(9), pp. 125–137. https://doi.org/10.3991/ijoe.v21i09.55653

Article submitted 2025-03-21. Revision uploaded 2025-05-17. Final acceptance 2025-05-17.

The World Health Organization (WHO) highlights the importance of proper training in the management and regulation of biomedical equipment, emphasizing that this knowledge is fundamental for sound decision-making in healthcare [3].

Some initiatives have been implemented, such as the establishment of specialized laboratories for the design and calibration of biomedical equipment in universities [4]. However, these efforts are insufficient to meet the demand for adequate training in biomedical metrology. Additionally, biomedical engineering professionals face challenges in staying up-to-date with technological advancements and sector regulations. The lack of resources and limited attention to the development of this essential area of engineering further exacerbate the problem.

Artificial intelligence (AI), particularly large language models (LLMs) such as GPT in its various versions, could play a crucial role in improving the training and support of biomedical engineers. LLMs are systems trained on large volumes of data, which enable them to understand and generate accurate responses in natural language. This capability can be leveraged to design virtual assistants specialized in biomedical metrology, providing access to precise information and recommendations as if guided by an expert in the field. By integrating a language model, such as ChatGPT, with the retrieval-augmented generation (RAG) technique, it is possible to enhance the model's ability to retrieve relevant information from specific databases, thereby increasing its accuracy and applicability in real-world scenarios [5].

There is an urgent need for accessible and accurate tools to support biomedical engineers in analyzing and understanding metrological data. The lack of training in this area affects the quality of measurements and compromises patient safety. Given that LLMs have proven effective in various fields, from medical diagnosis to customer service in specialized sectors, their application in biomedical metrology could revolutionize how biomedical engineers are trained and make decisions [6]. Furthermore, implementing this technology could help improve the calibration and regulation of medical equipment, making training more accessible in low-resource countries or rural areas where training opportunities are limited [7].

This study aims to evaluate the knowledge domain of LLMs in the field of biomedical metrology, with the goal of developing a specialized virtual assistant that serves as a support tool for biomedical engineering professionals. The effectiveness of different LLMs will be evaluated to identify the one that provides the most accurate and coherent responses regarding the fundamental concepts and principles of biomedical metrology. Additionally, the RAG technique was implemented to optimize the performance of the selected model. The development of this virtual assistant is expected to significantly contribute to improving the training and performance of biomedical engineers by enabling them to access precise and reliable information. In contrast to previous research that applied RAG techniques in general domains, this study proposes an application specialized in biomedical metrology, a field that is still little explored in terms of AI solutions. The proposed approach stands out for combining computational efficiency, knowledge personalization, and accessibility, representing a concrete contribution to the professional training of biomedical engineers, especially in educational contexts with limited resources. The remainder of this paper is organized as follows: Section 2 details the literature review; Section 3 describes the methodology used for dataset creation, RAG implementation, and the metrics employed for evaluation; Section 4 presents the results; Section 5 discusses the findings; and finally, Section 6 provides the study's conclusion.

2 LITERATURE REVIEW

Large language models are advanced systems trained on extensive amounts of data designed to provide text comprehension and generation capabilities for complex tasks. These models can interpret and generate text coherently, making them valuable tools for processing and analyzing biomedical data [8]. Recently, studies have been conducted on the application of LLMs in multiple areas. One such evaluation focused on the performance of LLMs in the construction sector knowledge domain. This test was conducted on models such as GPT and LLaMA to investigate their performance in technical knowledge, specifically in the ASHRAE Certified HVAC Designer exam, which is related to heating, ventilation, and air-conditioning systems. The results demonstrated that some LLMs have the potential to assist professionals in designing these systems [9].

Similarly, another study utilized LLMs for multidimensional writing evaluation, comparing the performance of GPT-3.5, GPT-4, and Claude 2 in written text assessment tasks. This analysis highlights the advantages of these models over human-assigned scores [10]. Additionally, it has been shown that the performance of these models is sensitive to the instructions they receive. In this context, a study was conducted to evaluate the immediate sensitivity of language models, aiming to analyze their performance from different perspectives [11].

In the medical field, a chatbot optimized using the LLaMA language model was developed. After a fine-tuning process using datasets derived from conversations between patients and doctors, ChatDoctor significantly outperformed other LLMs, such as ChatGPT, in terms of accuracy, retrieval, and F1 score. This model demonstrated remarkable capability in answering questions about current diseases, improving the efficiency of medical diagnoses, and expanding access to quality medical consultations [12].

Conversely, a study demonstrated that implementing a RAG architecture can reduce hallucinations generated by the models. In this case, the accuracy of GPT-4 in generating preoperative instructions increased from 80.1% to 91.4%. This improvement was achieved using the Hugging Face library, which facilitates access to various pre-trained models [13]. Similarly, another study evaluated the performance of LLMs in sentiment analysis using the RAG architecture and observed significant improvements in model performance, with an increase of up to 18% in the F1 score [14].

In the educational field, a virtual assistant based on language models has been proposed to automate administrative tasks in educational institutions. In this study, the LLaMA 2-7B and Mistral 7B models were evaluated and optimized using the RAG technique. The authors concluded that LLaMA 2-7B offers a more viable solution for this purpose owing to its accuracy and efficiency [15]. In addition, an advanced educational robotic system has been developed that, through LLMs such as ChatGPT and Llama, teaches the correct writing of the alphabet. Owing to its computer vision algorithms and design, there has been a noticeable improvement in the learning skills of students [16].

3 MATERIALS AND METHODS

3.1 Dataset

To select the appropriate LLM, an exhaustive collection of information related to biomedical metrology and LLMs was conducted. Additionally, specific data that demonstrated expertise in Biomedical Metrology were selected. Subsequently, a specific

dataset was constructed based on the Biomedical Metrology Certification from the Certification Institute for Healthcare Technology (CBET). This dataset was structured and classified to facilitate its use for the fine-tuning and optimization of the LLM.

The closed-book question-and-answer method was used to evaluate the LLMs without access to additional resources. Fifty multiple-choice questions were used, each requiring an explanation for the selected answer, based on the CBET exam guide. These questions were designed to test the ability to recall key concepts in Biomedical Metrology.

The prompt was generated using a request template and consisted of two parts: the instruction and the question. The instruction explained the task the LLMs needed to perform, which was: \"Answer the following multiple-choice questions from the CBET Certified Technician exam and explain the reasoning behind your choice\" for this study. The question referred to one of the 50 single-choice exam-style questions in the CBET study guide. Notably, some LLMs provide responses that are influenced by prior prompts. To avoid this influence on the test results, no prompts were supplied beforehand. Each test was conducted in a new dialogue session on the platform, considering the randomness of the LLM responses [9].

3.2 Large language models

Various language models were evaluated to select the most suitable model in terms of accuracy, efficiency, and technical feasibility within the metrology domain. **ChatGPT-4o**, developed by OpenAI and released in early 2024, is notable for its advanced reasoning and contextual understanding capabilities, although it has certain access restrictions [17]. **Llama 3.1-70B**, a high-performance model with extensive text generation capabilities, multilingual translation, and open-source availability [18]. **Claude 3.5 Sonnet**, developed by Anthropic in 2024, is recognized for its natural generation capabilities and precision in reasoning and coding tasks [19]. **Mistral 8x7B**, a model based on a multi-expert approach, with an efficient architecture for generation tasks [20]. **DeepSeek V3**, a recent open-source model focused on efficiency and precise generation, with evaluation results in educational knowledge domains comparable to closed-source models [21]. **Llama 3.2-3B Mini.Q4 k M.gguf**, a quantized model derived from Llama-3.2-3B-Instruct, offering greater efficiency and accessibility for deployment on resource-limited devices [22].

3.3 Implementation of retrieval-augmented generation

Once the most suitable model was selected, its responses were optimized for the specific domain of biomedical metrology through the implementation of the RAG technique. This technique allows the language model to be fine-tuned by retrieving relevant information, which helps to improve the accuracy and relevance of the generated responses.

The evolution of RAG is divided into three stages: naive, advanced, and modular.

- **Naive RAG** follows a basic process that includes the stages of indexing, retrieval, and generation. However, it has limitations related to accuracy, incomplete retrieval, and redundancy in responses.
- **Advanced RAG** addresses these deficiencies by optimizing the aforementioned stages. It improves indexing, embeddings, and the handling of noisy information through reclassification techniques, thereby increasing response quality.

- **Modular RAG** introduces a flexible architecture based on specialized modules, such as search, memory, routing, and RAGFusion. These modules enhance both information retrieval and generation. Unlike previous approaches, Modular RAG allows for the reorganization and adjustment of its components based on the specific task, thus increasing the system's accuracy and adaptability [23]. Therefore, Modular RAG was used in this study.

During implementation, key parameters, such as the value of top_k (from 3 to 5), were adjusted in the FAISS search, which allowed more relevant fragments of the corpus to be retrieved. Additional pre-processing was applied to eliminate redundancies and non-reporting data, and the reclassification step in the RAG fusion module was enhanced. These adjustments improved the contextualization and accuracy of the model when answering questions in the biomedical domain.

The all-MiniLM-L6-v2 architecture of sentence transformers was used to generate embeddings, given its efficiency and ability to semantically represent short texts. Prompt engineering consisted of "Question + Justify your choice" style templates, eliminating any previous context or training cues to ensure pure and unbiased answers.

The RAG process (see Figure 1) begins with the user's query input. The system preprocesses the query by removing characters, converting uppercase to lowercase, etc., and generating embeddings, which, in this case, were created using Sentence Transformers.

Simultaneously, a database was constructed to extract important fragments related to the queries. A filter was applied to the retrieved information, eliminating irrelevant content and generating embeddings.

In the next phase, relevant information was retrieved using a search index, such as FAISS, which vectorizes the previously extracted information [24]. The selected information, converted into vectors, was organized by the model (Llama 3.2-3B Mini.Q4 K M.gguf). Finally, the model delivered a response to the user through the application interface.

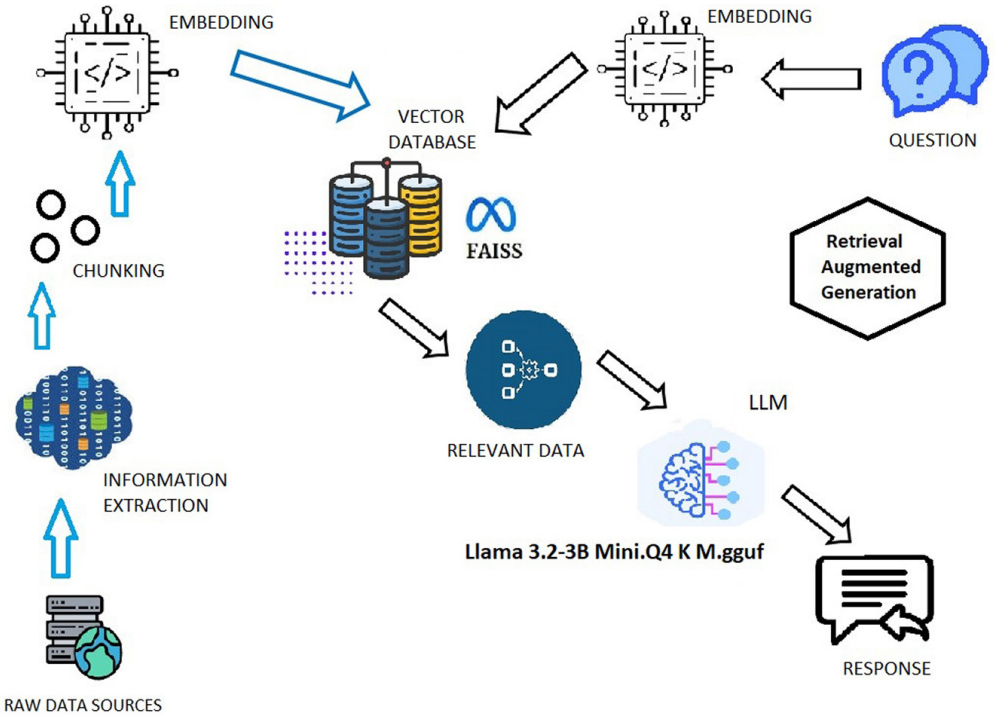

Fig. 1. Diagram of retrieval-augmented generation

3.4 Metrics

After determining the appropriate dataset, different language models were evaluated to assess their performance based on their knowledge of biomedical metrology. In this regard, several LLMs were selected, and tests were conducted using a previously constructed dataset. The results related to the accuracy and precision of their responses were analyzed to determine the most suitable model for incorporation as a specialized virtual assistant in this type of metrology.

The following metrics were used:

- **Rouge-L:** Evaluation based on the longest common subsequence (LCS). This metric was used to evaluate the accuracy and completeness of generated responses. It measures the overlap of n-grams between the response generated by the LLM and the correct response. It should be noted that Rouge-L does not require words to be contiguous, only to keep order, which better captures the structure of the text [25].
- **Recall:** This metric measures the percentage of words in the reference response that are present in the generated response. It was calculated by dividing the number of correctly generated words by the total number of words in the correct response.
- **F1 score:** This refers to a combination of precision and recall scores, considering that false-positive and false-negative scores are important for interpreting the effectiveness of classifications [26].

$$F1 = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{1}$$

- **Cosine similarity:** This metric evaluates the similarity between two vectors in a vector space. In natural language processing (NLP), texts are transformed into vectors, and cosine similarity measures the closeness of these texts based on the angle between the vectors. Unlike metrics such as Rouge-L or F1-Score, which focus on word matches, cosine similarity is based on the orientation of the vectors, allowing for the evaluation of text alignment in terms of content, regardless of length or scale [26].

$$Cosine\,Similarity\ (A,B) = \frac{\mathrm{A} \times \mathrm{B}}{\|\mathrm{A}\| \times \|\mathrm{B}\|} \tag{2}$$

3.5 Design and implementation

Design and development of the user interface. The first phase focused on the design and development of the user interface (UI). This involves designing the appearance of the assistant in the web application. This includes defining the functional modules integrated into the app, such as the home window, chat window, and text input bar, as shown in Figure 2.

For this purpose, **Streamlit** was used, which is a Python-based framework for creating web applications oriented toward data and AI [27]. Additionally, **Google Colab**, a cloud-hosted service that requires no setup, was used to program and execute the Python code directly from the browser [28].

3.6 Integration of the LLM into the application

Model deployment. Hugging Face was used for the integration of the app interface and model with RAG. Hugging Face is an AI company specializing in the development of NLP models and providing access to various pre-trained models. The Hugging Face platform was utilized, where models can be hosted and tools for their creation and evaluation are available [29].

API configuration. An application programming interface (API) is a service contract between two applications that communicate with each other through requests and responses [30]. In this case, a web API was created that functioned as a processing interface between the web server and web browser.

Testing and evaluation. A testing process was performed to ensure the proper functioning of the application. Initially, the model performance was evaluated in its base state. Subsequently, additional tests were conducted after integrating the RAG into the virtual assistant. All results from both evaluations were collected and analyzed to identify areas for improvement and make adjustments, if necessary.

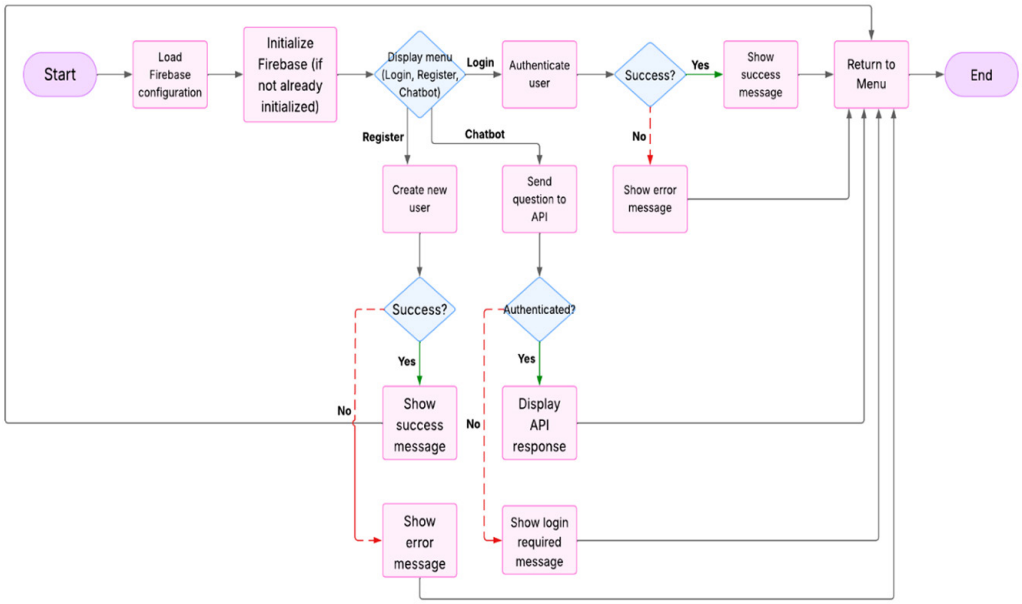

Fig. 2. User interface flow diagram

4 RESULTS

4.1 Evaluation of LLM Effectiveness

Original LLM. To evaluate the effectiveness of the LLMs, the following metrics were used: Rouge-L, F1 score, and cosine similarity. The results are presented in Table 1, which shows that a high value in the Rouge-L metric does not always imply a higher percentage of correct answers. For example, the Llama 3.2-3B Mini model showed a high Rouge-L score (0.3034), indicating that its responses were structurally similar but not necessarily correct.

Table 1. Statistical results (mean and standard deviation) of Rouge-L, F1 score, and cosine similarity based on CBET evaluation for each original model

LLMs	LLMs Average Metrics		
	Rouge-L	F1 Score	Cosine Similarity
ChatGPT-4o	$0.2739_{(0.0571)}$	$0.3481_{(0.0575)}$	$0.6056_{(0.1185)}$
Llama 3.1-70B	$0.2669_{(0.0591)}$	$0.3677_{(0.0653)}$	$0.6096_{(0.1132)}$
Claude 3.5 Sonnet	$0.2878_{(0.0492)}$	$0.3105_{(0.0450)}$	$0.5768_{(0.1328)}$
Mistral 8x7B	$0.2761_{(0.0680)}$	$0.3855_{(0.0683)}$	$0.5749_{(0.1346)}$
DeepSeek V3	$0.3188_{(0.0621)}$	$0.3959_{(0.0632)}$	$0.5974_{(0.1243)}$
Llama 3.2-3B_Mini	$0.3034_{(0.0706)}$	$0.4023_{(0.0781)}$	$0.6087_{(0.1146)}$

In contrast, considering that the F1 score combines the precision and recall metrics, Table 1 shows that the Llama 3.2-3B Mini model is capable of correctly identifying a significant number of correct answers. Similarly, it is evident that the Claude 3.5 Sonnet model has the lowest F1 score (0.3105), reflecting its inferior performance in terms of accuracy when generating correct answers. In this case, Llama 3.2-3B Mini exhibited the highest F1 score (0.402), indicating better performance in selecting correct answers.

Regarding the cosine similarity metric, which measures the semantic similarity between the generated and reference responses, Table 1 shows that the Llama 3.2-3B Mini model achieved a high value (0.6087). This suggests that the model generates responses that are semantically similar to the reference responses, although they are not always accurate or correct in terms of content.

In contrast, the Claude 3.5 Sonnet (0.5768) and Mistral 8x7B (0.5749) models exhibited the lowest cosine similarity values. This could be because the models used language that was less aligned with the expected responses.

LLM with RAG (retrieval-augmented generation). Once the effectiveness tests of the evaluated models were conducted, the Llama 3.2-3B Mini Model was selected to apply the RAG technique to improve its accuracy. This choice was based on its balanced performance in terms of the evaluation metrics F1 score, Rouge-L, and cosine similarity, which demonstrated that it generated consistent responses that were semantically and structurally aligned with the reference answers.

To carry out this process, we began by constructing a specific dataset based on information related to biomedical metrology. This dataset includes key topics such as anatomy and physiology, public health safety in healthcare settings, fundamentals of electricity and electronics, health technology and function, solving health technology problems, and health information technology.

Subsequently, a retrieval index was constructed using Facebook AI Similarity Search (FAISS), an open-source library created for similarity search and clustering of dense vectors. FAISS can be used to build an index and perform searches with remarkable speeds and memory efficiencies. This technique improved the retrieval of relevant documents and ensured that the model generated more precise and contextualized responses than the base model.

The retrieval index was constructed based on a relevant dataset for biomedical metrology. The chosen model (Llama-3.2-3B_Mini.Q4_K_M.gguf) was integrated to optimize the performance of the responses.

Evaluation of the effectiveness of the virtual assistant. To compare the performance of the original LLM and the LLM with RAG, 50 CBET-based questions were

presented to both models, and the F1 score, Rouge-L, and cosine similarity were calculated. To determine whether there was a significant difference between the means of the metrics obtained from the original LLM and LLM with RAG models, statistical tests were conducted for two related datasets. Initially, the normality assumption of the metrics was verified using the Shapiro-Wilk test, revealing that only the F1 score metric of the original LLM did not meet this assumption. Based on this information, the paired samples t-test was applied to the cosine similarity and Rouge-L metrics, while the Wilcoxon test was employed for the F1 score metric. The results, with a significance level of 0.05, indicated a significant difference between the means of the metrics obtained from the two models. Table 2 presents the results of the statistical tests performed.

Table 2. Results of statistical tests on the means of the metrics obtained from evaluating the Llama-3.2-3B_Mini.Q4_K_M.gguf model with and without the RAG system

Normality Test (Shapiro-Wilk)		
Metrics	**Original LLM**	**LLM with RAG**
F1 Score	P-value = 0.0233 < 0.05	P-value = 0.0783 > 0.05
Rouge-L	P-value = 0.2928 > 0.05	P-value = 0.0839 > 0.05
Cosine Similarity	P-value = 0.0803 > 0.05	P-value = 0.5371 > 0.05
Paired Sample Test		
F1 Score	Wilcoxon-Test: P-value = 0.0438 < 0.05	
Rouge-L	T-Test: P-value = 2.5677×10^{-9} < 0.05	
Cosine Similarity	T-Test: P-value = 3.8182×10^{-5} < 0.05	

As shown in Table 3, a significant increase in the F1 score from 0.402 to 0.526 is evident, indicating that the optimized model is more effective in generating relevant responses in the field of biomedical metrology. The use of RAG complements the model's internal knowledge with retrieved information, thus enhancing accuracy.

Moreover, an increase in Rouge-L and cosine similarity was observed, suggesting that the responses generated with RAG were better structurally and semantically aligned with the reference answers. This validates the effectiveness of integrating RAG, demonstrating an improvement in the quality of responses in biomedical metrology.

Table 3. Statistical evaluation metric results (mean and standard deviation) based on CBET evaluation for the Llama-3.2-3B_Mini.Q4_K_M.gguf model with and without the RAG system

Metrics	**Model (Llama-3.2-3B_Mini.Q4_K_M.gguf)**	
	Original LLM	**LLM with RAG**
F1 Score	$0.4023_{(0.0781)}$	$0.5267_{(0.1654)}$
Rouge-L	$0.3034_{(0.0706)}$	$0.4974_{(0.1685)}$
Cosine Similarity	$0.6087_{(0.1146)}$	$0.6576_{(0.1540)}$

The results indicated an improvement in the accuracy of the responses, validating the effectiveness of integrating the RAG technique. In this sense, a significant improvement in the quality of response generation was observed for topics related to biomedical metrology.

4.2 Assistant UI

User interface. The UI was developed using Streamlit, and the modules consisted of a home screen, a chat window, and bottom text input bar. As shown in Figure 3, the home screen includes a login and a brief introduction to the assistant's functions. The chat window is conversational in style, where the user can observe real-time interactions with the virtual assistant. The bottom text input bar is used to enter questions and send them to the assistant based on the selected large language models.

The integration of the RAG model API with the app interface in Streamlit was performed using Hugging Face Spaces, allowing the application to process user queries and generate real-time responses based on the retrieved information.

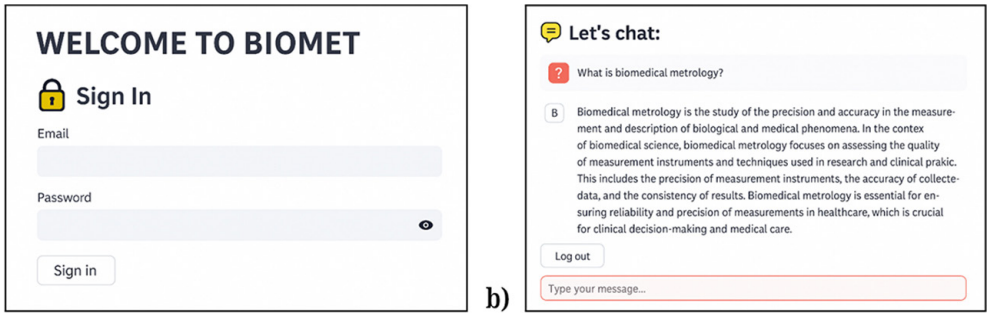

Fig. 3. Graphical interface (a) Login with registration (b) Conversational interface BiometChat-User

5 DISCUSSIONS

The integration of the RAG technique with the Llama 3.2-3B Mini model allowed for a significant improvement in the quality of responses in the field of biomedical metrology. The main benefits include an increased F1 score by 30.9%, indicating greater accuracy in the answers; reduction of hallucinations through the use of a controlled knowledge base; efficient implementation in resource-constrained environments through the use of a quantified model; and improved semantics and structure of responses, with greater consistency to the reference material.

The results obtained from the evaluation of the Llama-3.2-3B_Mini.Q4_K_M.gguf model with the implementation of RAG demonstrated a significant improvement in response generation within the domain of biomedical metrology. In particular, the increase in the F1 score suggests greater accuracy in identifying correct answers and their justification, while the improvement in Rouge-L and cosine similarity reflects better structuring and semantic alignment with the reference responses.

It is worth noting that Llama-3.2-3B_Mini.Q4_K_M.gguf is a quantized model, making it efficient in terms of memory and inference speed. Despite this optimization, it maintains a competitive performance compared to larger and heavier models. This aligns with a study in which personalized chatbots were developed using RAG to optimize their performance, showing a massive improvement from the base model (0.55) to the RAG-optimized model (0.91). This suggests that access to external information sources enhances the education and creation of an authentic knowledge base for the models [31].

The integration of FAISS as a search and retrieval engine optimized the selection of relevant information, mitigating the knowledge limitations of the base model. This contributed to reducing errors and improving the contextualization of generated responses. This optimization process follows the recommendations described by

Vidivelli et al. [31] and Chandrasekhar et al. [13], who demonstrated that the use of modular RAG architectures and tuning of parameters, such as top_k, can significantly increase the accuracy of models in specialized domains. Furthermore, the evaluation demonstrated that the use of RAG was effective in providing a more precise reference framework, avoiding the generation of incorrect responses.

Additionally, an paper mentions that beyond applying RAG to produce an LLM specialized in a specific topic, it is possible to do so by varying certain parameters, such as the number of **top_k**, which determines the number of top n-grams considered during the search. This would result in a more reliable model capable of maintaining high fidelity and accuracy in its response [13]. This study corroborates that using a higher **top_k** value provides the model with greater reliability in its response. In the present study, the **top_k** value was increased from three to five, which contributed to the improvement in the metric results of the RAG-optimized model.

Although the results are promising, this study has some limitations. The evaluated model (Llama 3.2-3B Mini) has size and context constraints that can affect the response depth in complex scenarios. In addition, the wizard is hosted on the Hugging Face Space, so optimization for scalability is required. For future work, a more complete environment is envisaged to optimize the model using additional fine-tuning and to integrate it with interactive educational platforms.

6 CONCLUSIONS

The development of a virtual assistant using the Llama 3.2-3B Mini model optimized with RAG for the domain of biomedical metrology demonstrates the potential of these technologies to improve access to accurate information in specialized areas. The RAG technique helped mitigate knowledge limitations of the base model and reduce errors by providing access to external domain-specific information. The integration of FAISS for document retrieval and sentence transformers for embedding generation optimized the relevance of the responses, while the implementation of Firebase ensured secure and personalized user management in the application. Additionally, deploying the API with Docker on Hugging Face Spaces and using Streamlit as the interface ensured an accessible solution.

In future work, the optimization of the model through more specific training and the implementation of strategies to improve computational efficiency and system scalability will be considered using techniques such as fine-tuning. Moving forward, the focus will be on long-term stability and the expansion of the existing knowledge base, with the goal of providing a more robust and accurate tool for biomedical engineering professionals specializing in biomedical metrology.

7 REFERENCES

[1] V. Becerra, V. Téllez, G. Peñaloza, and M. Castro, "Vital signs assistant for prehospital care," *Pädi Scientific Bulletin of Basic Sciences and Engineering of the ICBI,* vol. 11, pp. 152–160, 2023. https://doi.org/10.29057/icbi.v11iEspecial2.10720

[2] M. V. A. Cabrera, E. F. M. Ortega, and R. G. Hernández, "La educación metrológica en la formación técnica: Una mirada crítica y propositiva desde la figura profesional agropecuaria," *Prohominum*, vol. 6, no. 3, pp. 8–24, 2024. https://doi.org/10.47606/acven/ph0257

[3] R. Benitez, R. Uresti, and T. González, "Challenges of metrology in clinical engineering in Mexico," *SOMIB Mexican Society of Biomedical Engineering*, 2017. dx.doi.org/10.24254/CNIB.17.8

[4] J. Chang *et al.*, "Masi: A mechanical ventilator based on a manual resuscitator with telemedicine capabilities for patients with ARDS during the COVID-19 crisis," *HardwareX*, vol. 9, p. e00187, 2021. https://doi.org/10.1016/j.ohx.2021.e00187

[5] A. Zeichick, "What is enhanced recovery generation (RAG)?" Oracle, 2023. [Online]. Available: https://www.oracle.com/pe/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/

[6] Andina News Agency, "Peru will launch its first medical device production laboratory," 2022. [Online]. Available: https://andina.pe/agencia/noticia-peru-pondra-funcionamiento-primer-laboratorio-produccion-dispositivos-medicos-913103.aspx

[7] U. Mutilba and G. Kortaberria, "Towards automated and integrated metrology in production," *Tekniker Member of Baque Research y Technology Alliance*, 2023. [Online]. Available: https://www.tekniker.es/media/uploads/noticias/TSN39_Tekniker_Articulo_Metrologia.pdf

[8] B. Yan *et al.*, "On protecting the data privacy of large language models (LLMs) and LLM agents: A literature review," *High-Confidence Computing*, vol. 5, no. 2, 2025. https://doi.org/10.1016/j.hcc.2025.100300

[9] J. Lu *et al.*, "Evaluation of large language models (LLMs) on the mastery of knowledge and skills in the heating, ventilation and air conditioning (HVAC) industry," *Chinese Roots Global Impact,* vol. 6, no. 14, 2024. https://doi.org/10.1016/j.enbenv.2024.03.010

[10] X. Tang, H. Chen, D. Lin, and K. Li, "Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments," *Heliyon*, vol. 10, no. 14, p. e34263, 2024. https://doi.org/10.1016/j.heliyon.2024.e34262

[11] J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, and K. Chen, "ProSA: Assessing and understanding the prompt sensitivity of LLMs," *arXiv preprint arXiv:2410.12405,* 2024. https://doi.org/10.48550/arXiv.2410.12405

[12] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge," *Cureus*, vol. 15, no. 6, 2023. https://doi.org/10.7759/cureus.40895

[13] A. Chandrasekhar, J. Chan, F. Ogoke, O. Ajenifujah, and A. B. Farimani, "AMGPT: A large language model for contextual querying in additive manufacturing," *Additive Manuf. Lett.,* vol. 11, p. 100232, 2024. https://doi.org/10.1016/j.addlet.2024.100232

[14] S. Khaled, E. H. Mohamed, and W. Medhat, "Evaluating large language models for arabic sentiment analysis: A comparative study using retrieval-augmented generation," *Procedia Comput. Sci.,* vol. 244, pp. 363–370, 2024. https://doi.org/10.1016/j.procs.2024.10.210

[15] U. H. Khan, M. H. Khan, and R. Ali, "Large language model based educational virtual assistant using RAG framework," *Procedia Comput. Sci.*, vol. 252, pp. 905–911, 2025. https://doi.org/10.1016/j.procs.2025.01.051

[16] S. Vinoth Kumar, R. B. Saroo Raj, J. Praveenchandar, S. Vidhya, S. Karthick, and R. Madhubala, "Future prospects of large language models: Enabling natural language processing in educational robotics," *International Journal of Interactive Mobile Technologies,* vol. 18, no 23, pp. 85–97, 2024. https://doi.org/10.3991/ijim.v18i23.51419

[17] Z. Huang, Z. Wang, S. Xia, and P. Liu, "OlympicArena Finals: Claude 3.5 Sonnet vs. GPT-4o," *arXiv preprint arXiv:2406.16772v2*, 2024. https://arxiv.org/html/2406.16772v2

[18] Facebook company, "Introducing Llama 3.1: Our most capable large-scale language model yet," 2024. [Online]. Available: https://about.fb.com/ltam/news/2024/07/presentamos-llama-3-1-nuestro-modelo-de-lenguaje-a-gran-escala-mas-capaz-hasta-la-fecha/

[19] Anthopic, "Introducing Claude 3.5 Sonnet," 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[20] A. Jiang *et al.*, "Mixtral of experts," Mistral.AI, 2024. [Online]. https://arxiv.org/pdf/2401.04088

[21] DeepSeek-AI, "DeepSeek-V3 technical report," 2024. https://arxiv.org/html/2412.19437v1#S1

[22] J. Ou, Y. Chen, B. Xiong, Z. Wang, and W. Tian, "Accelerating mixture-of-experts language model inference via plug-and-play lookahead gate on a single GPU," *Comput. Stand. Interfaces,* vol. 95, no. 103996, 2025. https://doi.org/10.1016/j.csi.2025.103996

[23] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2024. https://doi.org/10.48550/arXiv.2312.10997

[24] LangChain "Faiss | Introduction | LangChain," 2024. [Online]. Available: https://python.langchain.com/docs/integrations/vectorstores/faiss/

[25] A. Citarella, M. Barbella, M. Ciobanu, F. Marco, L. Biasi, and G. Tortora, "Assessing the effectiveness of ROUGE as unbiased metric in Extractive vs. Abstractive summarization techniques," *J. Comput. Sci.,* vol. 87, p. 102571, 2025. https://doi.org/10.1016/j.jocs.2025.102571

[26] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores," *Appl. Intell.*, vol. 52, pp. 4961–4972, 2022. https://doi.org/10.1007/s10489-021-02635-5

[27] Streamlit, "Basic concepts of Streamlit," 2024. [Online]. Available: https://docs.streamlit.io/get-started/fundamentals/main-concepts

[28] Google colab, "Welcome to colaboratory," 2024. [Online]. Available: https://colab.research.google.com/?hl=en-GB

[29] Hugging Face, "The AI community building the future," Hugging face.co, 2024. [Online]. Available: https://huggingface.co/

[30] Amazon, "¿What is an application programming interface (API)?" 2024. [Online]. Available: https://aws.amazon.com/es/what-is/api/

[31] S. Vidivelli, M. Ramachandran, and A. Dharunbalaji, "Efficiency-driven custom chatbot development: Unleashing langchain, RAG, and performance-optimized LLM fusion," *Comput., Mater. & Continua,* vol. 80, no. 2, pp. 1–10, 2024. https://doi.org/10.32604/cmc.2024.054360

8 AUTHORS

Yamilet Carreon completed her B.S. in Biomedical Engineering from the Universidad Tecnológica del Perú, Lima, Peru, in 2024. Her study interests cover a wide range of areas such as machine and deep learning, computer vision, remote sensing, programming for healthcare applications and the use of emerging technologies in medical systems (E-mail: u18203594@utp.edu.pe).

Miguel Chicchon completed his B.S. degree in Electronic Engineering at the Universidad Nacional de Ingeniería, Lima, Peru, in 2009, and obtained his M.S. degree in Informatics specializing in Computer Science from the Pontificia Universidad Católica del Perú (PUCP), Lima, in 2019. Currently, he is pursuing a Ph.D. degree in Engineering at PUCP. Since 2024, he has held the position of Professor in Robotics at the Universidad Tecnológica del Perú. His study interests encompass a wide range of areas such as machine learning, deep learning, reinforcement learning, computer vision and remote sensing (E-mail: c11201@utp.edu.pe).