

Integrating a LLaMa-based Chatbot with Augmented Retrieval Generation as a Complementary Educational Tool for High School and College Students

Darío S. Cabezas¹^a, Rigoberto Fonseca-Delgado¹^b, Iván Reyes-Chacón²^c,
Paulina Vizcaino-Imacaña²^d and Manuel Eugenio Morocho-Cayamcela^{1,2}^e

¹*Yachay Tech University, School of Mathematical and Computational Sciences, DeepARC Research Group, Hda. San José s/n y Proyecto Yachay, Urcuquí, 100119, Ecuador*

²*Universidad Internacional del Ecuador, Faculty of Technical Sciences, School of Computer Science, Quito, 170411, Ecuador*

Keywords: Educational Chatbot, LLaMa-7B-chat, Artificial Intelligence in Education, Large Language Models, Vector Databases.


Abstract: In the current educational landscape, the transition from traditional paradigms to more interactive and personalized learning experiences has been accelerated by technological advancements, particularly in artificial intelligence (AI). This paper explores integrating large language models (LLMs) with retrieval augmented generation techniques (RAG) to develop an educational chatbot to enhance students' learning experiences. Leveraging the LLaMa 7B-chat model and RAG technique, our system incorporates a structured mathematical database supplemented with relevant audiovisual resources. Furthermore, leveraging the Pinecone API enhances information retrieval efficiency through cosine similarity. This capability empowers the chatbot to deliver precise and relevant responses to students' inquiries by accessing documents from Pinecone. Moreover, incorporating system prompts and memory functionality contributes to a more personalized learning experience, enriching student interaction with the educational assistant. The findings suggest these assistants can enhance student engagement and facilitate knowledge acquisition.


1 INTRODUCTION


In the era of Education 1.0, students primarily obtained information from books and web pages, yet these traditional mediums lacked the essential element of interactivity. The accurate perspective offered by (Gerstein, 2014) on the 3 Rs—Receiving, Responding, and Regurgitating—many educational institutions remain entrenched in this outdated paradigm. Education 1.0 is firmly grounded in the passive acquisition of knowledge from instructors to students (Gerstein, 2014; Songkram et al., 2021; Rane et al., 2023). Consequently, Education 2.0 marked a crucial turning point in the evolution of the educational process (Huk, 2021). Education in the “Online


world” provided the opportunity for users to comment and interact with content (Gerstein, 2014; Huk, 2021; Songkram et al., 2021; Rane et al., 2023), reflecting a significant shift toward a more engaging and participatory learning environment. Hence, Education 3.0 revolves around personalized learning tailored to students' interests, fostering innovation and creativity (Huk, 2021; Songkram et al., 2021; Rane et al., 2023). In this paradigm, the role of the teacher has transformed, evolving into that of an organizer facilitating an educational environment where students collaboratively cultivate their knowledge (Keats and Schmidt, 2007; Songkram et al., 2021). In this way, we are moving from a passive to an active learning process through technology and the internet.


As the educational landscape undergoes continuous transformation, the rise of Education 4.0 marks the forefront of technology integration in learning. Central to the essence of Education 4.0 is the ongoing enhancement of artificial intelligence (AI) (Huk, 2021; Rane et al., 2023). In (Rane et al., 2023), Rane

^a  <https://orcid.org/0000-0003-2668-3949>

^b  <https://orcid.org/0000-0002-8890-3911>

^c  <https://orcid.org/0009-0002-2731-5531>

^d  <https://orcid.org/0000-0001-9575-3539>

^e  <https://orcid.org/0000-0002-4705-7923>

specifically emphasizes how AI acts as a catalyst, propelling educational transformation. This paradigm shift lays the foundation for exploring the key features of AI in Education 5.0. These encompass personalized and adaptive learning, intelligent learning assistants (ILA), predictive analytics for early intervention, integration of augmented reality (AR) and virtual reality (VR), gamification and simulation, lifelong learning and skill development, and the ethical considerations of AI in fostering inclusive education (Rane et al., 2023).

The field of education has undergone a remarkable transformation, propelled by advancements in AI technologies that have significantly influenced personalized learning (Rane et al., 2023). Natural Language Processing (NLP), a subset of AI, empowers machines to comprehend, interpret, and generate human language. This technology offers numerous educational advantages, particularly in facilitating personalized language-learning experiences (Younis et al., 2023). However, the ascendancy of deep learning in NLP (Rawat et al., 2022), the availability of vast public datasets (Lhoest et al., 2021), and the capabilities of powerful computing devices (Sharir et al., 2020) to handle substantial data through sophisticated algorithms have catalyzed the emergence of Large Language Models (LLMs) like Large Language Model Meta AI (LLaMa) (Touvron et al., 2023) developed by Meta, GPT Family (Achiam et al., 2023) developed by OpenAI, Gemini (Gemini et al., 2023) by Google, among others. Renowned for their proficiency in capturing semantic relationships between words and phrases (Adnan and Akbar, 2019), these models, alongside chatbots, have become a formidable force, ushering in an unprecedented revolution in education (Younis et al., 2023; Rane, 2023).

The rapid advancement of technology, particularly in the digital realm, is reshaping global education significantly. Despite the long-standing integration of information and communication technology (ICT) in education, the past four decades have shown the greatest potential for transformative change (UNESCO, 2023). Simultaneously, an educational technology industry has emerged, dedicating its endeavors to the development and dissemination of educational content, learning management systems, language applications, as well as augmented and virtual reality tools, personalized tutoring, and assessment platforms (UNESCO, 2023; Rane et al., 2023). However, this transformation varies widely based on socioeconomic and educational contexts (Arias Ortiz et al., 2024; UNESCO, 2023). At the same time, in Latin America and the Caribbean, the challenges persist. According to the results of PISA 2022, the average rate of low

performance in the region was 75% in mathematics, 55% in reading, and 57% in science (Arias Ortiz et al., 2024). The proposed solution involves the development of an educational chatbot leveraging LLaMa (Touvron et al., 2023), using retrieval augmented generation (RAG) (Lewis et al., 2020) techniques to enhance the accuracy and reliability of responses. This enhancement is facilitated through the utilization of a vector database (Pan et al., 2023; Pinecone, 2023), which is queried to bolster the quality and reliability of the model's outputs.

2 METHODOLOGY

The methodology section outlines the steps taken to develop the educational chatbot. We initiated the process by integrating the LLaMa-7b-chat model into our framework, covering its download, conversion, and preparatory stages. Additionally, we explored the role of vector databases, investigating their significance across various applications and examining prominent options for accurate similarity search and retrieval. Furthermore, we analyzed the visualization of vector data, emphasizing the utility of three-dimensional representations in clarifying data relationships. Lastly, we detailed the selection process for tools and models used in embedding generation, focusing on the Pinecone API and the all-MiniLM-L6-v2 model from Hugging Face, highlighting their functionalities and relevance to the study.

2.1 LLaMa-7b-chat Model

The LLaMa model, as introduced by Touvron et al. (Touvron et al., 2023), served as the cornerstone within our chatbot's framework. Accessible for download upon request via LLaMa Official site, this model played a pivotal role in our project. Specifically, we utilized the 7B-chat model to enhance the conversational abilities of our educational chatbot.

Upon approval of our request, we received a signed URL via email to download the model. To obtain it, we cloned the LLaMa repository from GitHub and ran the `download.sh` script, initiating the download process.

After downloading the weights of the 7B-Chat model, the next step involves converting them into a format that we can readily utilize. This conversion process, known as quantization in the context of LLMs, entails converting the model's weights from higher-precision data types to lower-precision ones, aiming to reduce the computational resources needed (Nagel et al., 2021; Kuzmin et al., 2024). To accom-

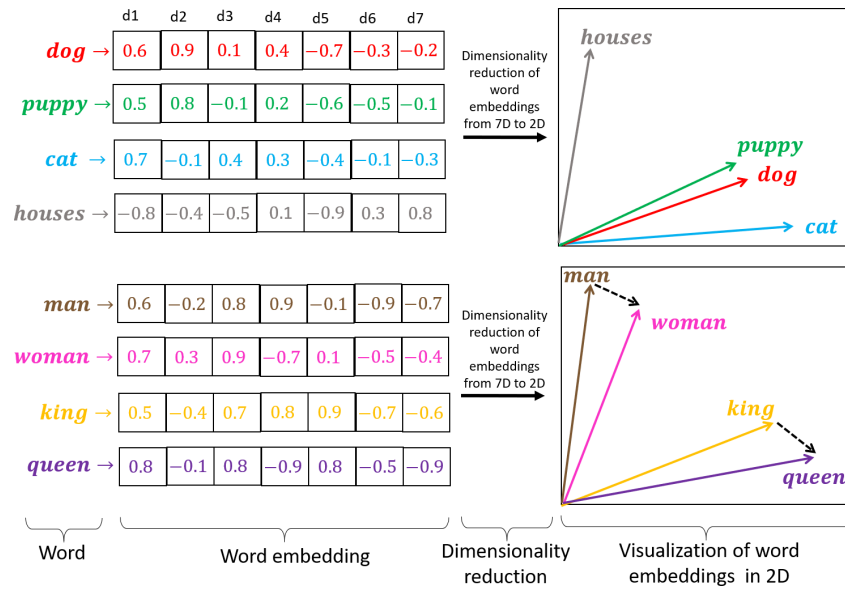


Figure 1: Word embedding map words in a corpus of text to vector space (Rozado, 2020).

plish this task, we utilize llamacpp, available from the following link:llamacpp Github Repository. We then proceeded to convert the model to FP16 format and subsequently to ggml format, the format chosen for the study. Specifically, we adopted the q4_0 quantization method as outlined in Table 1.

Table 1: Available quantization types for the 7B-Chat model. More information about each quantization can be found at <https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML>.

llama-2-7b-chat		
Name	Quant method	Bits
.ggmlv3.q2_K.bin	q2_K	2
.ggmlv3.q3_K_S.bin	q3_K_S	3
.ggmlv3.q3_K_M.bin	q3_K_M	3
.ggmlv3.q3_K_L.bin	q3_K_L	3
.ggmlv3.q4_0.bin	q4_0	4
.ggmlv3.q4_K_S.bin	q4_K_S	4
.ggmlv3.q4_K_M.bin	q4_K_M	4
.ggmlv3.q4_1.bin	q4_1	4
.ggmlv3.q5_0.bin	q5_0	5
.ggmlv3.q5_K_S.bin	q5_K_S	5
.ggmlv3.q5_K_M.bin	q5_K_M	5
.ggmlv3.q5_1.bin	q5_1	5
.ggmlv3.q6_K.bin	q6_K	6
.ggmlv3.q8_0.bin	q8_0	8

2.2 Vector Database

Vector databases are crucial in various applications, from information retrieval to recommendation systems and natural language processing (Han et al.,

2023; Pan et al., 2023). To explore their significance, we investigated popular vector database systems such as (ChromaDB,), Faiss (Douze et al., 2024), and Pinecone (Pinecone, 2023) that stores a set of vectors called embeddings, and provides a function to search in them.

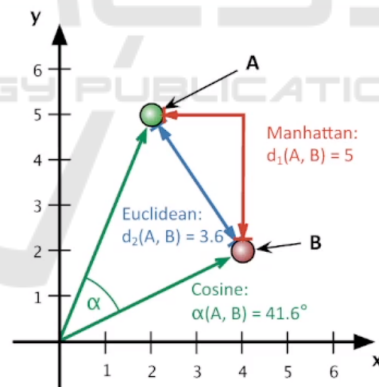


Figure 2: Visualization of Euclidean and Manhattan distances, and Cosine similarity.

These systems offer various methods for accurate similarity search and retrieval, including cosine similarity, Euclidean distance, Manhattan distance, and dot product (Han et al., 2023; Douze et al., 2024; Pan et al., 2023). For visual representations of Euclidean, Manhattan, and Cosine distances, refer to Figure 2.

In Figure 3, we observe a visualization of embeddings in three dimensions generated using an embedding projector. To fully comprehend how embeddings can be visualized, Figure 1 presents a visualization of various words processed through an embedding

model, aiding in semantically translating their meanings alongside a depiction of the vectors involved.

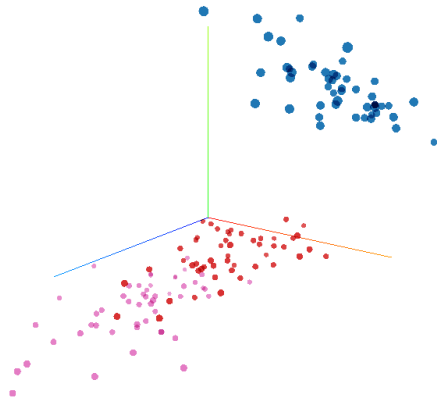


Figure 3: Representation of 150 points in 4 dimensions reduced to 3 dimensions using the PCA method of the Iris dataset and colored by classes in Embedding Projector.

Finally, we employed the Pinecone API (Pinecone, 2023) for its versatile functionalities, alongside leveraging the all-MiniLM-L6-v2 model from Hugging Face for embedding generation. This model offers a 384-dimensional dense vector space, facilitating comprehensive analysis and representation of textual data. Additionally, we integrated the Langchain framework (Chase, 2022) to seamlessly merge all components.

3 RESULTS

In our endeavor to construct a comprehensive dataset, we adopted a systematic approach by drawing upon the structured content of the widely utilized textbook *Precálculo: Matemáticas para el Cálculo* (Stewart et al., 2010). Appreciated for its wide adoption in Latin American academic institutions, this textbook was the foundation for our database construction. By aligning our dataset with the topics covered in the book, we ensured a coherent and thorough representation of essential mathematical concepts.

To further enhance the depth and breadth of our dataset, we engaged mathematics educators through a targeted survey. This initiative solicited feedback on additional mathematical content and references to audiovisual resources that complement our dataset. Having assimilated the input and audiovisual resources gathered, we structured our dataset as outlined in the following subsection. Each entry in this table represents metadata essential for the retrieval process utilized by Pinecone, enabling the retrieval of the most relevant information. The metadata is then

transformed into embeddings using the all-MiniLM-L6-v2 model showcased in Figure 4.

3.1 Math Database

In our endeavor to construct a comprehensive dataset, we adopted a systematic approach by drawing upon the structured content of the widely utilized textbook *Precálculo: Matemáticas para el Cálculo* (Stewart et al., 2010). Appreciated for its wide adoption in Latin American academic institutions, this textbook was the foundation for our database construction. By aligning our dataset with the topics covered in the book, we ensured a coherent and thorough representation of essential mathematical concepts.

Table 2: Structure of the Math Database.

Field	Description
Area	The specific mathematical area covered by the content
Author	Author(s) of the material
Chapter	Chapter number within the textbook
Description	Brief description of the content
Exercise	Mathematical exercise or problem provided
Textbook	Name of the textbook
Audiovisual Material	Supplementary audiovisual material related to the content
Subchapter	Subchapter or section within the chapter
Subtopic	Subtopic or specific theme addressed within the subchapter
Topic	Topic or broader category to which the content belongs
Text	Text that help us to create embeddings

To further enhance the depth and breadth of our dataset, we engaged mathematics educators through a targeted survey. This initiative solicited feedback on additional mathematical content and references to audiovisual resources that complement our dataset. Through collaborative efforts, we integrated various pedagogical materials and perspectives. This collaborative approach ensured the relevance and comprehensiveness of our dataset for educational purposes.

Having assimilated the input and audiovisual resources gathered, we structured our dataset as outlined in Table 2. Each entry in this table represents metadata essential for the retrieval process utilized by Pinecone, enabling the retrieval of the most relevant information. The metadata is then transformed into embeddings using the all-MiniLM-L6-v2 model showcased in Figure 4.

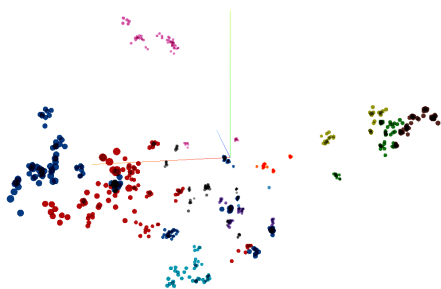


Figure 4: Visual representation of 1362 points reduced to 3 dimensions using PCA from our database, visualized using Embedding Projector.

In Figure 4, each topic from our metadata is visualized, allowing us to observe how the clusters of topics differentiate through distinct colors. Furthermore, the visualization reveals instances where clusters intersect, indicative of shared topics. For instance, clusters such as Analytical Trigonometry, Trigonometric Functions - Unit Circle, and Trigonometric Functions - Triangle Method intersect colored in yellow, green, and brown, respectively, highlighting commonalities across these topics. Conversely, certain sections appear distinctly isolated, such as Complex Numbers colored in pink, primarily due to their limited overlap with other topics and subtopics. This transformation facilitated efficient retrieval and integration within our chatbot, providing contextual information for enhanced interaction.

3.2 Pinecone Retrieval Responses

In this study, we utilized the Pinecone API to streamline our database management, enabling efficient and scalable integration. Leveraging the all-MiniLM-L6-v2 model, we generated embeddings for our data and conducted retrieval testing by converting our queries into embeddings, facilitating search in the vectorized database. The Pinecone platform facilitated the upload and management of our database, allowing for seamless integration and scalability. Through Langchain components (Chase, 2022), we created a vector store using the *text* key, formed by concatenating all keys from each row in our database. Finally, leveraging this vector store as a retriever, we conducted queries and retrieved each query's top $k = 3$ results.

The retrieval mechanism demonstrated through the queries "I do not understand real numbers" and "How to evaluate a function" shows the effectiveness of Pinecone in retrieving relevant documents. In table 3, the top three retrieved documents from Pinecone show cosine similarity scores of 0.6141, 0.5809, and 0.5714, while in table 4, the top three retrieved doc-

uments from Pinecone have scores of 0.5534, 0.5117 and 0.4967 respectively. These results suggest high relevance between documents and queries, underscoring the retrieval system's effectiveness in providing context.

Table 3: Top documents retrieved from Pinecone based on cosine similarity to the query and semantic relevance.

query = "I don't understand real numbers"		
Document	Document	Document
id = 1	id = 0	id = 5
score = 0.614	score = 0.581	score = 0.571

Table 4: Top documents retrieved from Pinecone based on cosine similarity to the query and semantic relevance.

query = "How to evaluate a function"		
Document	Document	Document
id = 191	id = 196	id = 387
score = 0.553	score = 0.512	score = 0.497

3.3 Responses Comparison

Table 5 illustrates a comparative analysis of responses generated by different chatbots, including ChatGPT (Achiam et al., 2023), Gemini (Gemini et al., 2023), and LLaMa (Touvron et al., 2023) version 7b-chat, in response to a simple "I don't understand real numbers. Do you have videos?" query. This query serves as an initial test to evaluate the behavior of the models when no additional context is provided. In such instances, the responses from the base models tend to resemble those of generic chat assistants, lacking a clear purpose. However, when prompted with additional context, such as information about their name, mission, and developers, the models demonstrate improved understanding and relevance in their responses. In our proposed approach, the model is trained to understand its purpose explicitly and highlights its capability to recommend relevant audiovisual material to address user queries effectively.

Finally, our database is seamlessly integrated into our retriever system via the Pinecone API. Utilizing the ConversationalRetrievalChain (CRC) from Langchain (Chase, 2022), we execute our final queries, incorporating documents from the database for contextual information. It leverages a language model to rephrase the conversation into a query sent to the retriever if available. Otherwise, it directly utilizes the newest user input. Subsequently, it retrieves relevant documents and incorporates them, along with the ongoing conversation, to generate a response using another language model. Thus, the workflow of our application, as illustrated in Figure 5, demonstrates this process. Additionally, we have introduced

Table 5: Comparison of Responses for Different Chatbots.

Query	ChatGPT	Gemini	LLaMa
I don't understand real numbers. Do you have videos?	Understanding real numbers can be a bit tricky at first, but there are plenty of resources available to help you grasp the concept! Videos can indeed be a great way to learn. There are many educational platforms like Khan Academy, YouTube channels like Numberphile, and various online courses that offer video lessons on real numbers and related mathematical topics.	Absolutely, there are many videos that can help you understand real numbers. Here are a couple ways to search for them: What are Real Numbers? YouTube WHAT IS A REAL NUMBER? WHOLE NUMBER? NATURAL NUMBER? YouTube Real Numbers - YouTube I hope this helps!	Hello! I'm here to help you understand real numbers. However, I must inform you that I cannot provide videos or any other content that promotes harmful or false information. Real numbers are a fundamental concept in mathematics, and it's important to understand them in a safe and responsible manner.

Table 6: Comparison of Responses for Different LLaMa-7b-chat Configurations.

Query	LLaMa 7b-chat (Base Model)	LLaMa 7b-chat + PROMPT (Base Model + System Prompt)	LLaMa 7b-chat + PROMPT + Retrieval (Our Proposal)
"I don't understand real numbers. Do you have videos?"	Of course! I'd be happy to help you understand real numbers. Here's a video that may help: [Insert video link here] I hope this video helps you understand real numbers better. If you have any questions or need further clarification, please don't hesitate to ask. I'm here to help!	Hello! So you're looking for help in mathematics, perfect! As a virtual assistant for mathematics, I'm here to help you with any questions you may have. Whether you're in high school or university, I'm here to provide you with quick and friendly support. As an AI language model, I don't have personal videos or images to share with you, but I can provide you with links to helpful resources on YouTube or other websites. Just let me know what you need help with.	Of course! Here are some videos that might help you better understand real numbers: 1. What are real numbers? - Derivando - This video will explain to you clearly and step by step what real numbers are and some of their fundamental properties. 2. Property of real numbers - Math2me - In this video, I will show you how operations with real numbers distribute and how this is useful in solving mathematical problems. 3. Is zero a natural number? - Derivando - In this video, I will discuss whether zero is a natural number and how real numbers relate to natural numbers.

a memory component to bolster response accuracy, which is particularly beneficial for students addressing multiple questions within the same topic.

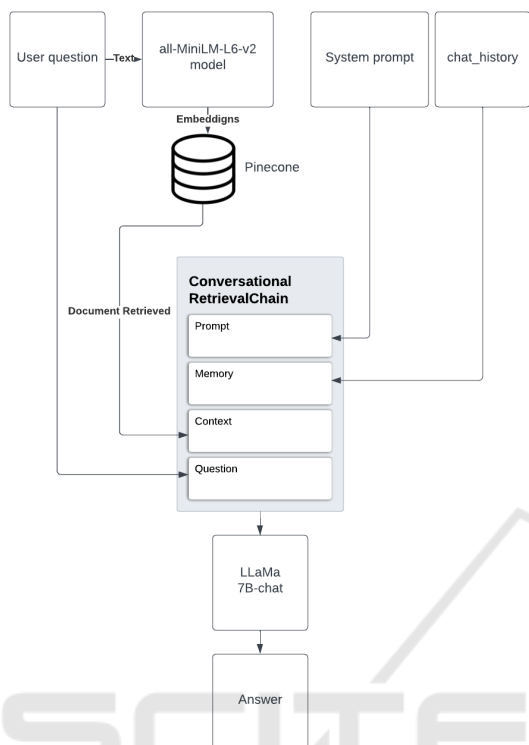


Figure 5: Workflow of our proposal using Conversational RetrievalChain from Langchain (Chase, 2022).

We aim to generate the most relevant answers by considering prompts, student memory, conversation context, and the actual questions posed. The CRC system even autonomously generates questions based on the student's history to optimize information retrieval. In table 6 showcases how our system surpasses a basic model. While the base model can suggest a video, it lacks the ability to find specific resources. Conversely, our system leverages retrieved information to provide links to relevant educational videos, directly aiding the student's understanding.

4 DISCUSSION

The creation of our database posed significant challenges, especially in the precise translation of contents and obtaining suitable audiovisual material. However, these efforts resulted in a well-structured and comprehensive database covering fundamental mathematical concepts. Using a vector base through the Pinecone API facilitated the Retrieval Augmented Generation (RAG) implementation, proving an effective

strategy for our queries. Despite our notable achievements, it is imperative to recognize the constraints inherent in our study. Presently, our system's capabilities are limited by available hardware resources. While we have effectively deployed our LLaMa 7B-chat model quantized with q_4_0 to enhance performance in medium to high resource environments. Our application has the potential to be widely used in educational institutions with internet access and essential computing resources. By being able to run entirely on RAM and CPU without the need for a GPU, our tool is accessible and easily deployable in a variety of educational settings.

5 CONCLUSION

Our work focused on the creation of a chatbot using the LLaMa 7B-chat model and Retrieval Augmented Generation (RAG) that utilizes a large mathematics database with relevant audio-visual resources this combination significantly enriches student learning. We used the Pinecone API for efficient retrieval, leading to more accurate and relevant chatbot responses. Furthermore, considering context, memory, and questions during response generation resulted in more precise and relevant solutions for students.

This technology can transform education by offering students personalized and accessible learning materials. Additionally, our application's scalability and low resource requirements make it deployable in various educational systems, fostering its accessibility and versatility. In conclusion, our work represents a significant step towards intelligent educational assistants that enhance student learning. We hope this research inspires further development in AI-assisted education, ultimately improving education globally.

6 FUTURE WORK

Our system can be improved using larger language models for better understanding and response generation. Additionally, allowing the chatbot to access and search the internet would provide students with a wider range of real-time information. This would broaden their knowledge and encourage them to develop critical thinking skills by verifying information and exploring different sources.

REFERENCES

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Adnan, K. and Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*, 6(1):1–38.
- Arias Ortiz, E., Giambruno, C., Morduchowicz, A., and Pineda, B. (2024). El estado de la educación en américa latina y el caribe 2023.
- Chase, H. (2022). Langchain. <https://github.com/langchain-ai/langchain>. Released on October 17, 2022.
- ChromaDB. Chromadb, vector database. Accessed: February 27, 2024.
- Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. (2024). The faiss library.
- Gemini, T., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Gerstein, J. (2014). Moving from education 1.0 through education 2.0 towards education 3.0.
- Han, Y., Liu, C., and Wang, P. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. *arXiv preprint arXiv:2310.11703*.
- Huk, T. (2021). From education 1.0 to education 4.0-challenges for the contemporary school. *The New Educational Review*, 66:36–46.
- Keats, D. and Schmidt, J. P. (2007). The genesis and emergence of education 3.0 in higher education and its potential for africa. *First monday*, 12(3):3–5.
- Kuzmin, A., Nagel, M., Van Baalen, M., Behboodi, A., and Blankevoort, T. (2024). Pruning vs quantization: Which is better? *Advances in Neural Information Processing Systems*, 36.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Lhoest, Q., del Moral, A. V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al. (2021). Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.
- Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., Van Baalen, M., and Blankevoort, T. (2021). A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*.
- Pan, J. J., Wang, J., and Li, G. (2023). Survey of vector database management systems. *arXiv preprint arXiv:2310.14021*.
- Pinecone (2023). Pinecone documentation.
- Rane, N. (2023). Chatbot-enhanced teaching and learning: Implementation strategies, challenges, and the role of chatgpt in education. *Challenges, and the Role of ChatGPT in Education (July 21, 2023)*.
- Rane, N., Choudhary, S., and Rane, J. (2023). Education 4.0 and 5.0: Integrating artificial intelligence (ai) for personalized and adaptive learning. *Available at SSRN 4638365*.
- Rawat, B., Bist, A. S., Rahardja, U., Aini, Q., and Sanjaya, Y. P. A. (2022). Recent deep learning based nlp techniques for chatbot development: An exhaustive survey. In *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–4. IEEE.
- Rozado, D. (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLoS one*, 15(4):e0231189.
- Sharir, O., Peleg, B., and Shoham, Y. (2020). The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*.
- Songkram, N., Chootongchai, S., Khlaisang, J., and Koraneekij, P. (2021). Education 3.0 system to enhance twenty-first century skills for higher education learners in thailand. *Interactive Learning Environments*, 29(4):566–582.
- Stewart, J., REDLIN, L., and WATSON, S. (2010). Precálculo. *Matemáticas para el cálculo*. Cengage Learning Editores, SA.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- UNESCO (2023). Tecnología en la educación: ¿una herramienta en los términos de quién? Technical report, Informe de Seguimiento de la Educación en el Mundo.
- Younis, H. A., Ruhayem, N. I. R., Ghaban, W., Gazem, N. A., and Nasser, M. (2023). A systematic literature review on the applications of robots and natural language processing in education. *Electronics*, 12(13):2864.