2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) | 979-8-3315-1591-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IITCEE64140.2025.10915367

Enhancing Comprehension with LLM, TTS, and RAG: Transcription of Text into Podcasts and Chatbots

Arya Jalindar Kadam
Department of Information Technology
BRACT's Vishwakarma Institute of Information Technology
Pune, India
arya.22210766@viit.ac.in

Prasad Rajaram Kute
Department of Information Technology
BRACT's Vishwakarma Institute of Information Technology
Pune, India
prasad.22210330@viit.ac.in

Chinmay Ashok Kale
Department of Information Technology
BRACT's Vishwakarma Institute of Information Technology
Pune, India
kalechinmaywork@gmail.com

Kushal Bhoraji Pathave
Department of Information Technology
BRACT's Vishwakarma Institute of Information Technology
Pune, India
kushalpathave54@gmail.com

Chaitali Shewale
Department of Information Technology
BRACT's Vishwakarma Institute of Information Technology
Pune, India
chaitali.shewale@viit.ac.in

Pawan Wawage
Department of Information Technology
BRACT's Vishwakarma Institute of Information Technology
Pune, India
pawan.wawage@viit.ac.in

***Abstract*—The rapid growth of Large Lan-guage Models, Text-to-Speech technologies, and Retrieval-Augmented Generation is bringing a new mode of accessing and understanding textual information. This work talks about how these technologies can be coordinated so that texts are converted into conversational podcasts and chatbots. This makes information accessible and interesting to interact with. The system is based on using Llama3.2:3b for generating script, Suno/Bark TTS model for audio synthesis as close to reality, and a RAG-based chatbot for contextual interaction. with users. User studies present considerable changes in understanding and participation, meaning that these machines are going to flip the learning settings.**
Index Terms—Large Language Models (LLM), Text-to-Speech (TTS), Retrieval-Augmented Generation (RAG), Document-to-Podcast, Chatbot Interaction, Comprehension Enhancement

***Index Terms*—Large Language Models (LLM), Text-to-Speech (TTS), Retrieval-Augmented Generation (RAG), Document-to-Podcast, Chatbot Interaction, Comprehension Enhancement**

I. Introduction

The world has come of age technologically, although old media consumption through texts pose specific difficulties, with the increase in size and complication of documents. Thick-texted papers that contain technical terms or even com- plex print languages defeat the purpose of the paper since users have limited chance to understand or even memorize what was read. This is made worse by the fact that while consumers of information have to devote substantial time and attention to messages, this entails a lowered level of comprehension. [1]

The developments in LLMs have brought about a start to new possibilities of enhancing accessibilities due to the capabilities of producing, as well as understanding natural like texts. Altogether with the Text-to-Speech (TTS) technology, these systems provide the users with a more progressive interface - the opportunity to interact with the information in an auditory way. Furthermore, interactive chatbots add to this by presenting answers to queries, building content, and providing overview and tutorial on some subjects. Although, LLMs and TTS are promising technologies, they still have some drawbacks, which should be successfully managed for their effective application. Indeed, Recognition chatbots have their limitations that should be further improved. [2]

Another problem lies in hallucination, the LLMs give in-formation that seems reasonable but is not accurate or even made up, therefore causing misinterpretation among the users. This issue comes up because LLMs are frequently taught from out-of-date or context-restricted data and can take no external knowledge in real-time, limiting their capacities to approximate factual truth. To address these challenges, the presented paper proposes a new system, which combines LLMs, TTS and RAG technologies to convert text documents into rich audio formats based on podcasts and chatbot experiences. This approach enhances user experiences and comprehensiveness since texts are provided in the users' preferred media, incorporating the TTS mode, which uses audio elements, the use of the chatbot that has an interactive and conversational feature.

[3]

Retrieval-Augmented Generation Techniques: The most important ingredient here consists of our approach. RAG merges retrieval capabilities with the possible strengths of the generative models to enhance LLM capabilities. [4]

The system we will be developing for viewers allows them to interact with content as conversational, episodic podcasts that break down information. These podcasts enable the users to pose questions in real time and to seek feedback and elaboration with challenging ideas explained and segmented into simpler components. This interactive feature improves conceptual knowledge as users are allowed to get close to information while the chatbot Like helps in the flow of complex content. As will be illustrated in the following, the RAG technique, allows the chatbot to provide users with extended summaries or answers contexts, helping them remain attentive and limit the cognitive load to process complex information. [5]

This multi- modal approach is used to bring information closer since there are difficulties encountered by users in the text based formats. When synthesizing the proposed LLMs, TTS, and RAG, the system can provide clear communication of the ideas and a more efficient way for a human user to engage with the concept. The aim is to fulfill the requirement of making 'simple' an environment conducive to which users can obtain to and comprehend complex information. The system also tends to reduce the problem related to LLM hallucinations and factuality errors that means, provide a more credible and useful input-output interface to the user. [6]

II. Background and Literature Review

In so many years, the development of technology has emerged so high and led to the development of so many ways for information processing. Discoveries such as big language models and those retrieval-augmented generation methods, including text-to-speech systems, change the face of education and hence make it accessible and enjoyable. Systems, such as GPT-3 – a sequence of generative language models – have, in fact, already been proven to generate texts almost to the state-of-art level that are almost indistinguishable from safe texts generated by people for a reading comprehension task with increases in engagement. According to Brown (2020), the coherent and contextually relevant text generated by those models enhances learning for a student's sake. [7]

Thus, up to now, the most common application of LLMs to education has been that of an interactive tutor generating content. Emerging research shows that this idea of learning will fit well with the range of needs in learners, and the research finding in evidence of enhanced understanding. Ac- cording to Qin et al. (2024), the versatility of LLMs opens up avenues for having exceedingly individualized learning experiences tailored toward the student's personal preference and learning style [8]. The text-to-speech technology also supports LLMs because the text-based information is being transformed into an audio format, which satisfies different kinds of learners with common learned balance. And since including TTS systems satisfies the auditory learner, based on the summation of evidence, a few exhibit positive impacts on comprehension and recall. As indicated by Abid Haleem et al. (2022), the technologies listed above fill gaps in learning through consumption methods of content. [9]

Recently, this new form of retrieval has been discovered in the new approach known as Retrieval-Augmented Generation and also offers new avenues of improving how we can retrieve information and thus understand it in context or provide information efficiently to the users. For example, recent research in this area shows that RAG greatly enhances generation for in- formation towards educational ends [10]. According to Patrick et al. (2021), RAG expands the information retrieved relevance and ties it into an interactive educational application; hence, it may be more of an interactive learning tool because a user can input their elaborative question and then get corresponding contextual responses to the impactful effect. [11]

Except for this, however, the transcribing of such texts into forms such as podcasts and chatbots opened our eyes to new possibilities: that of sharing and consuming information. Ever since then, people learn on the go through podcasts. In addition, research has also proven to be good in understanding audio formats since it gives room to context with nuances such as intonation and inflection. Harlalka et al. (2015) further assert that the audio learning materials present the emotional and contextual cues appropriately, making this training worthwhile [12]. There are also certain talks that use LLMs and TTS to ensure that people get the right knowledge. They can converse with information that would enable persons to learn something. Ifelebuegu et al. (2023) demonstrated that chatbots result in interactive learning whenever the user holds a conversation that makes sense. They all lead, ultimately, to opening exciting avenues for revolution in education, this time both in making knowledge more accessible and more engaging [13].

In addition, AI-based products like chatbots and text-to-speech would enrich the experiential involvement of learners and drive collaborative learning environments. Such AI devices can create marvelous experiences for learning whereby learners are encouraged to probe the topics with deep inquiry and further question asking that may grow into a culture of curiosity and inquiry, as described by Chauncey and McKenna (2023) [14]. Even though this is not overlooked, one cannot forget that pitfalls with this yield the quality of information and general content produced. Unlocking more benefits into these LLMs, TTS, and RAG will add value to experiences in comprehension and learning. According to Ciobanu et al. (2017), ensuring the quality and reliability of content should be of the topmost importance when it comes to unlocking all these advanced technologies' benefits. [15]

III. Methodology

The project is divided into two main parts: Podcast Creation and RAG Chatbot Integration. All of these parts improve the understanding of the documents through different but mutually supportive methods. We also describe below the technologies and workflows of each part. [16] Refer Fig. 1

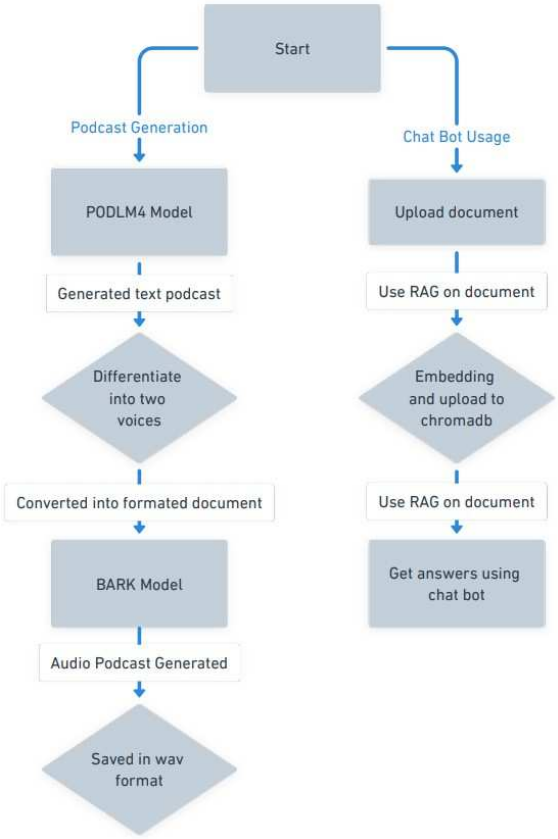

Fig. 1. Model Architecture

A. Podcast Generation

1) ***Document Extraction and Segmentation****:* The first aspect of the process of making a podcast is to obtain content from the documents uploaded in the application. The application is able to extract text from pdf and text files. To read text from PDF we used the PyPDF2 library. This makes it easy to retrieve all the information from the document. The extracted Document is then divided into different segments of specified length as per Fig. 2.

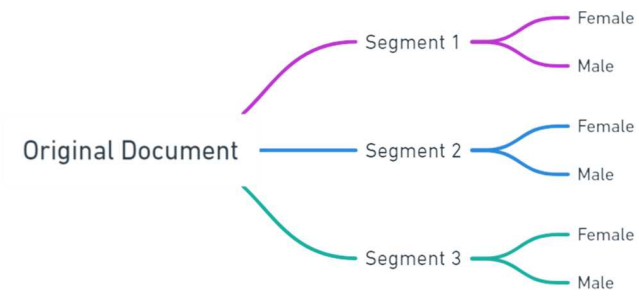

Fig. 2. Document Segmentation

2) ***Script Generation****:* Llama3.2:3b LLM model is used for text generation and is implemented with the help of Ollama. To generate specific text in the format of a podcast the Llama3.2:3b model is tuned and named PODLM4 by giving system prompts that define certain rules for the script generation like adding different kinds of expressions, defining the length of the sentences, etc. Furthermore, increasing the context window from default: 2048 to 3000 will result in more accurate results and will make sure that the model remembers the SYSTEM prompt given.

The script is generated in a specific format as follows.

Female: -context-
Male: -context-

Fig. 2, the original is thus converted into a script/podcast of two characters. After this, The female and male conversations are extracted from this by using regular expression (re) and stored in a Python list.

3) ***Text to Audio Conversion****:* For converting the created script into two different voices namely female and male. For this, we can make use of the suno/bark-small model. Bark is a TTS model that can be used with the help of transformers. This model was chosen for its realism, the generated voices sound as natural as human voices, so the podcast will be both realistic and professional. This model can create audio for various expressions as well.

The two different voices used in the suno/bark-small model:

- Female voice: ”v2/en-speaker-9”
- Male voice: ”v2/en-speaker-6”

The respective sentences for female and male characters are fed into their respective TTS models. The audio is generated and concatenated with the previously generated audio. The Concatenated audio is then converted to a .wav file format.

B. Chatbot Implementation

The chatbot uses a Retrieval-Augmented Generation (RAG) system, which contains a dual-stage process: first, vector embeddings and then search algorithms, including cosine similarities, define the parts of the given document that are most relevant within context, and second, a generation component expanded with a fine-tuning LLM, e.g., LLAMA 3.2: 3b. It made it possible for the bot to be able to move from the placed document and bring out some parts of the placed new document in relation to the specific query the user made in relation to the real time result of the query in addition to incorporating more parts from the placed document in the result. Refer Fig. 3.

1) ***Document Extraction and Vector Embedding****:* For doing efficient document querying, the system processes the uploaded documents by extracting and deciding the document in different chunks with respect to their contents. These segments, are then converted into vector embeddings using a pre- trained model from Sentence Transformers, Specifically the ’sentence-transformers/all-MiniLM-L6-v2’ model. This embedding model encodes chunks of text into fixed-size vectors,

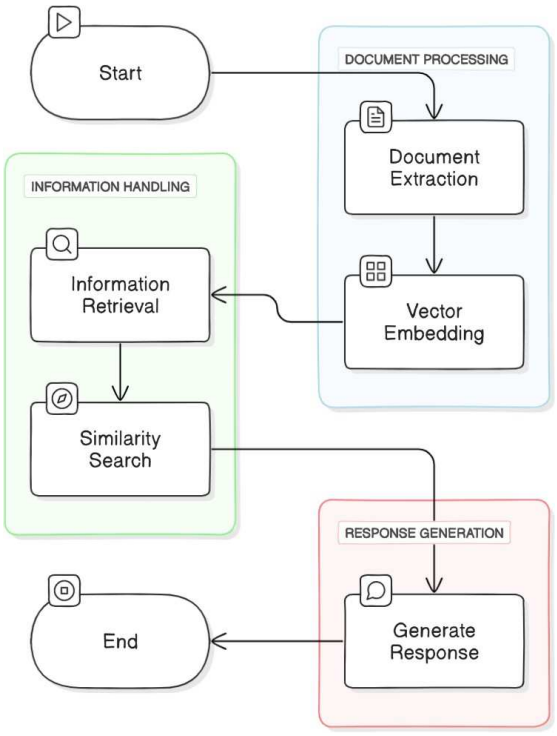

Fig. 3. Chatbot Implementation

which makes it possible for semantic similarity comparison. The document ingestion process is as follows:

1) Document Extraction: For the extraction of the text from the documents themselves the use of PyPDF2 is employed thereby guaranteeing that the extraction of content is done to the highest yield and speed.
2) Text Chunking: After extracting the document's contents, the text is divided into different chunks of 300 words.
3) Vector Database Storage: The extracted chunks are assigned IDs based on the document and chunk number. After which the chunks are stored in the ChromaDB vector database.

*2) **Information Retrieval and Similarity Search**:* The retrieval step in the RAG system requires conversion of the user query to an embedding using the SentenceTransformer model. This embedding is then matched with embedded key phrases in the vector database by using cosine similarity. As calculated by vector distance, the k most relevant document chunks relevant to the query are then given based on similarity scores para.

*3) **Generation of Response using LLM**:* Here the same LLM model, that is the Llama3.2:3b is tuned by creating a system prompt and increasing the context window (named PODLMCHAT).

The top relevant document chunks retrieved are fed into PODLMCHAT with the user query. Hence by using only the small and important section of the document, relevant responses are created even with a relatively small context window size.

IV. Prototype Development and Testing

A. Overview of Prototyping: Features, Functionalities, and User Interface

Users may upload text files or PDFs and watch how the podcasting process happens in real-time. The audio file downloaded is the outcome of podcasting generation. A clean and simple interface predominated, giving immediate feedback at every step of the podcasting process. It is the same interface that houses the chatbot; questions about the content of the document may actually be asked of the system after it has been podcasted.

B. Testing Methodology: User Studies, Surveys, or Expert Evaluations

The prototype was tested by 20 students, educators, or professionals. The test proved how the system is usable, how clear the podcasts that were outputted are, and the right answer of the chatbot. After using the system, participants had to undertake a survey in which they evaluated the amount to which they understood, interacted with the system, and satisfied.

C. Evaluation Metrics: Improved Comprehension, User Engagement, and Satisfaction

The primary evaluation metrics included the following:

- **Comprehension Improvement**: Testing the participants' comprehension of document content before and after using the podcast feature
- **User Engagement**: Users' time spent interacting with the system; the ratings provided by users for the conversational podcast style.
- **Satisfaction**: The overall satisfaction regarding the system of those participants as well as general evaluation about the quality of audio being generated and how helpful the chatbot was for them was rated.

V. Results

The podcast application takes advantage of the improvements in TTS and RAG methods. It uses the male and female voice for pod cast to suit the users preference. It is possible to download a podcast or listen to it on the spot with the versatility of choosing play back options. This tallies well with research done by Abid Haleem et al. (2022) that suggest that improved understanding and memorization for TTS systems especially for the auditory learner [9]. The structure of the application on the screen is depicted in the following Figure 4. Upon uploading a file, users are presented with two options:

1) **Generate**: This option creates the pod cast and can also act as the copyrighted title of the show. In turn, the algorithm converts the text that has been uploaded into an expressive audio podcast of high-quality and in the form of a podcast using generative language models. Coherence capability and engagement as pointed out by Brown (2020) makes such models [7].
2) **Upload**: This option uploads the database to Chroma DB which allows for a customized chatbot for Learning.

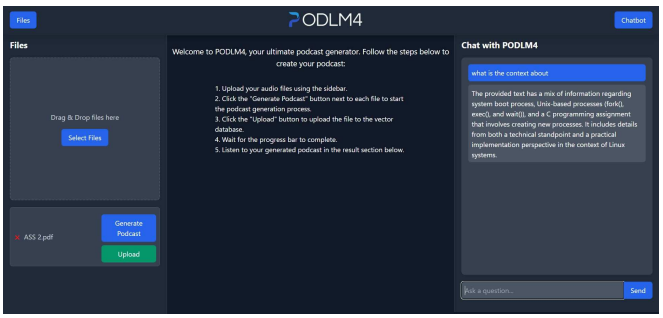

Fig. 4. Application Screen

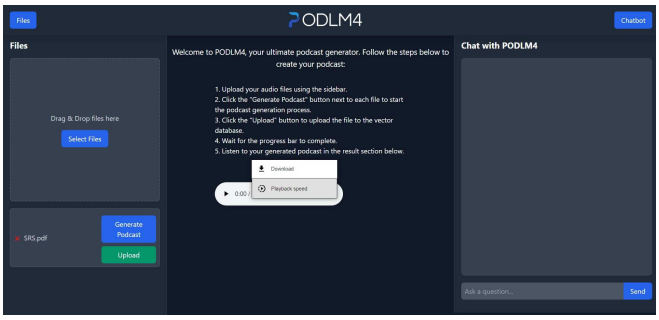

Fig. 6. Generated Podcast

This extends RAG, improving response contextuality as pointed out by Aleks and Patrick et al. [11].

When the **Generate** button is clicked, the application starts podcast generation, as depicted in fig (5) which shows the loading page during the process.

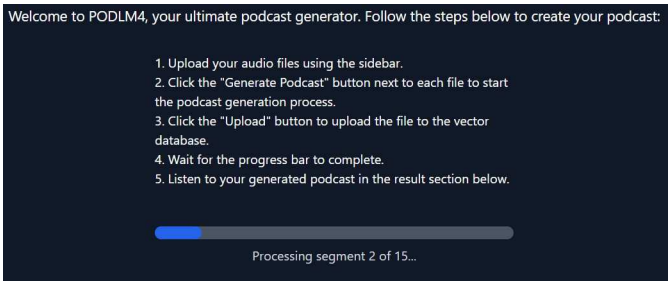

Fig. 5. Loading Podcast

The generated audio podcast is in WAV format and therefore, the sound quality is as it is. Users can get a copy from a web browser and the good thing is that it has the options of speed which allows for a varied learning speed among learners. This go together with what Harlalka et al. (2015), that explained that intonation and inflection from audio tutorials play important role [12].

Key features include:

- **User–interface**: An interface that can easily be accessed by the users especially vulnerable ones.
- **Real time status**: The users know the status of the podcast generation so that they can enhance their experience.
- **Customability of chatbot**: Can offer personalized response which is very essential for individual learning as pointed by Qin et al. (2024) [8].

As described in Learning, it creates the collaborative and experiences environments what AI integration in learning described Chauncey et al. (2023) [14]. when used complementarily together with LLMs, TTS, and RAG, it fosters an interactive curriculum. A sample of a generated podcast is displayed below in Fig. 6.

VI. Limitations and Challenges

That being the case, the current model has the following drawbacks: Although there is progress. It cannot process images, tables, or graphs which are parts of documents and text whenever writing technical documents; thus, the quality of audio production that it can generate from multimedia-based documents is usually poor. Incorporating vision into NLP was found to enhance the model performance; however, our FAQ generation now employs only the textual input.

Further, the quality of the produced podcasts is functionally related to the type and division of input documents. When data is badly extracted or segmented, the quality of the audio produced is low. [17] also noted that structure is an important factor for quality contents in documents after curation. Often it has low qualitative characteristics: TTS in our system degrades with complex texts, for example. Even nowadays, problems that relate to tone and emphasis still negatively impact TTS when they are applied to broader, comprehensive content. Also, as strongside vector embeddings can be used for retrieval, they are not suitable for cases with polysemic or complex queries. Asked in Table 3, prior studies illustrate that dense retrieval has a much higher relevance yet lower precision and is sensitive to inconsistent or tricky queries, a shortcoming in our current approach.

Finally, the presented prototype is proof that we have advanced to the next level compared to previous models, but it also shows the further potential in improvement of educational audio materials.

VII. Conclusion

This project reveals an innovative perspective of document comprehension through the instruments of AI technologies such as LLMs, TTS, and RAG systems. As opposed to print media that restrict interaction, this implementation builds an interactive media document. By using Llama3.2:3b for script generation and other TTS such as suno/bark-small, it produces podcasts with natural-flowing narrations and both, male/female voice for a better learning. Through harnessing RAG with vector embeddings and semantic search, the chatbot guarantees contextually, individualised and engaging learning.

In comparison with previous approaches, this approach is more effective to bridge the gap of the static text presentations as well as single modality learning tools. Such systems tend to be non-interactive with presenting text in a non-interactive format, or in the use of only audio learning. Unlike our project, in which a podcast and an intelligent chatbot are incorporated

together to ensure a comprehensive learning process. Further, while traditional systems present several limitations, including little provision for individual learning experience, current solution enables utilization of query and elaborate answer methodology to enable users study complexities in area of interest which makes the learning experience more effective. [9], [11], [12]

Thus, the current system is providing good advantages but it can be improved. As to the future works, there can be the enhancement of the quality of the generated voices and making them sound more naturally, the enlargement of the number of expressive voices, and the usage of the text editing and processing functions. Furthermore, the system could be used to support non-nested and more extensive, general-purpose queries by uniting knowledge from several documents.

This project shows how all three strategies can be implemented at once in an effort to minimize the gap between conventional document delivery systems and today's learning structures. Realizing text as dynamic, multimedia it increasesiness and interactivity, thus enhancing its agenda-setting values. It creates new opportunities for the use of educational technologies and creation of individualized learning environments, and provides a basis for further studies that could someday change paradigms for learning and knowledge acquisition across disciplinary contexts.

VIII. Future Work

The development of our system has been seen as it improved from generating podcasts from text and from interacting with the user using a chatbot. A concern being researched is on how to incorporate features such as images, tables, and graphs into podcast generation that is helpful in e-learning. Additional research will be done in a similar vein by aiming to compare my simple visual cues with verbal ones in order to enhance teachability in podcast. Furthermore, as reported in [18], incorporating emotion into TTS will improve the interaction with the auditory as a result of engagement.

Other enhancements include the incorporation of the hybrid retrieval approach instead of the vector embeddings model employed at the moment, as suggested in the case of [19]. Improving chatbot's long conversation modeling will enable them to better handle different aspects of the same query, as in [17]. The last trend will be personalisation for which reference is made in [20]; here, podcast broadcasts and chatbot replies will be personalised, which improves learners' interest and their performance when using educational technology. These advances will further extend to multimodal content, better TTS, richer conversation models for chatbots – resulting in personal learning and enhanced performance.

References

[1] A. C.-S. Chang and S. Millett, "Improving reading rates and comprehension through audio-assisted extensive reading for beginner learners," *System*, vol. 52, pp. 91–102, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0346251X15000846

[2] Z. Ma, W. Wu, Z. Zheng, Y. Guo, Q. Chen, S. Zhang, and X. Chen, "Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 146–11 150.

[3] P. R. Gogulamudi, Y. V. Pavan Kumar, and K. Purna Prakash, "Hallucinations in large language models (llms)," 04 2024, pp. 1–6.

[4] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2312.10997

[5] C. Kooli, "Chatbots in education and research: A critical examination of ethical implications and solutions," *Sustainability*, vol. 15, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2071-1050/15/7/5614

[6] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Retrieval augmented generation in prompt-based text-to-speech synthesis with context-aware contrastive language-audio pretraining," 2024. [Online]. Available: https://arxiv.org/abs/2406.03714

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[8] H. Qin, *Transforming Education with Large Language Models: Opportunities, Challenges, and Ethical Considerations*, 08 2024.

[9] A. Haleem, M. Javaid, M. A. Qadri, and R. Suman, "Understanding the role of digital technologies in education: A review," *Sustainable Operations and Computers*, vol. 3, pp. 275–285, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666412722000137

[10] T. Adiguzel, H. Kaya, and F. Cansu, "Revolutionizing education with ai: Exploring the transformative potential of chatgpt," *Contemporary Educational Technology*, vol. 15, p. ep429, 04 2023.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Ku¨ttler, M. Lewis, W. tau Yih, T. Rockta¨schel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[12] A. Harlalka, J. Bajaj, A. Kumar, K. Yadav, O. Deshmukh, R. Punekar, and K. Sorathia, "Audio cues: Can sound be worth a hundred words?" in *Audio Cues: Can Sound be Worth a Hundred Words?*, vol. 9192, 08 2015.

[13] A. Ifelebuegu, P. Kulume, and P. Cherukut, "Chatbots and ai in education (aied) tools: The good, the bad, and the ugly," *Journal of Applied Learning Teaching*, vol. 6, 09 2023.

[14] S. A. Chauncey and H. P. McKenna, "A framework and exemplars for ethical and responsible use of ai chatbot technology to support teaching and learning," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100182, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X23000619

[15] O. Ciobanu and D. Neamtu, "The impact and importance of new technologies in business development in context of economic diversity," *Proceedings of the International Conference on Business Excellence*, vol. 11, 07 2017.

[16] A. Bora and H. Cuayahuitl, "Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications," *Machine Learning and Knowledge Extraction*, vol. 6, 10 2024.

[17] M. K. Dobbala, "Conversational ai and chatbots: Enhancing user experience on websites," *American Journal of Computer Science and Technology*, vol. 7, 07 2024.

[18] G. Hillaire, F. Iniesto, and B. Rienties, "Humanising text-to-speech through emotional expression in online courses," *Journal of Interactive Media in Education*, vol. 2019, 09 2019.

[19] D. Lee, S.-w. Hwang, K. Lee, S. Choi, and S. Park, "On complementarity objectives for hybrid retrieval," in *On Complementarity Objectives for Hybrid Retrieval*, 01 2023, pp. 13 357–13 368.

[20] S. Chandra, S. Verma, W. M. Lim, S. Kumar, and N. Donthu, "Personalization in personalized marketing: Trends and ways forward," *Psychology and Marketing*, 08 2022.