**Enhancing Learning using Retrieval Augmented Generation: A Survey**

Vishal Vijay Khot, Shreyash Premchand Shinde, Varenyam Bhushan Nikam, Prashant Kumar Shukla

1. Computer Department, Pimpri Chinchwad College of Engineering and Research, Pune, Maharashtra, India
2. Computer Department, Pimpri Chinchwad College of Engineering and Research, Pune, Maharashtra, India
3. Computer Department, Pimpri Chinchwad College of Engineering and Research, Pune, Maharashtra, India
4. Computer Department, Pimpri Chinchwad College of Engineering and Research, Pune, Maharashtra, India

Email:{[vk3800870@gmail.com](mailto:vk3800870@gmail.com), [shreyashps123@gmail.com](mailto:shreyashps123@gmail.com), [varenyamnikam004@gmail.com](mailto:varenyamnikam004@gmail.com), [Shuklaprashant123488@gmail.com](mailto:Shuklaprashant123488@gmail.com) }

**Abstract.**

This survey paper investigates the role of Retrieval Augmented Generation models in improving student learning by effectively capturing and describing video content. The RAG model leverages multimodal data by combining text, visual, and audio data to generate question-based content. By analyzing video content based on speech and visual metadata, the model can capture and record important content in videos, saving teachers and students time in entering important information. This approach refines the content of the video, making it easier to extract important details and useful insights. Integrating AI-driven RAG models into learning models opens up new possibilities for personalized learning. Students receive AI-generated content that is accurate and tailored to their learning needs, focusing on the most important aspects of the video lessons. This technique not only makes the video easier to use, but also keeps students engaged by creating questions that encourage deeper understanding of the material. The results of this research demonstrate the potential of the RAG model to transform the learning process by increasing the retrieval of important information and improving access to important content. The integration of various devices into this AI-based approach represents a significant advance in learning by enabling interactive and effective learning.

**Keywords:** Retrieval Augmented Generation (RAG), Multimodal Data, Personalized Learning, Question-Based Content, Video Content Analysis, Speech Metadata, Visual Metadata, AI-Driven Learning, Interactive Learning, Content Refinement.

# Introduction:

Recent advancements in AI have shown significant promise in enhancing education, particularly through the integration of Retrieval-Augmented Generation (RAG) systems. For instance, an intelligent chatbot tutoring system leveraging RAG and custom LLMs addresses limitations in traditional tutoring methods and general-purpose LLMs by delivering accurate, personalized, and contextually relevant assistance, ultimately boosting student engagement and academic performance [[15](#bookmark=id.wf11wsipoh9g)]. Similarly, OwlMentor, an AI-powered tool, provides university students with interactive learning features such as document-based chats, automatic question generation, and quiz creation, demonstrating its potential to support understanding of complex scientific texts [[17](#bookmark=id.yozs4lhsc02j)]. In higher education, RAG-based virtual teaching assistants and teaching aids have shown potential for improving learning experiences, though challenges such as ethical considerations, reliability, and hallucinations necessitate careful safeguards and policy compliance [[2](#bookmark=id.kr1v8yqv4803)]. Furthermore, the introduction of Education Specific Retrieval-Augmented Generative AI (ES-RAG AI) emphasizes the importance of tailoring AI solutions to educational needs, fostering transparency, accountability, and enhanced learning outcomes, akin to the calculator’s historical adoption in education [[5](#bookmark=id.d58lc8qyuvg7)]. Together, these studies highlight the transformative role of RAG in education, paving the way for responsible and effective integration of AI technologies to enrich learning experiences.

In recent years, interest in artificial intelligence generated content (AIGC) has been increasing. There are many content creation tools designed to create different types of content in various formats. These include large language models (LLMs) such as the GPT series[[7](#bookmark=id.ggrld4f3v219)][[8](#bookmark=id.h1lwikld6p5i)][[16](#bookmark=kix.5cyc4xz1z8it)] and LLAMA series for text and code generation, DALL-E and Stable Diffusion for images, and Sora [[13](#bookmark=id.2wkt5fyagicu)] for video. The term "AIGC" implies that the content is created by a standard design, not by human or process control. These designs demonstrate outstanding performance through the use of innovative techniques, extensive infrastructure, and high-quality materials. For example, sequence-to-sequence processing has shifted from using short-term temporal (LSTM) networks to Transformer-based models, and image processing has shifted from differential networking (GANs) to latent propagation models. Furthermore, the underlying structure, once numbering in the millions, is now in the millions. These advances are supported by extensive, high-quality data that provides rich information for optimizing nonstandard models.

**User:**Users typically initiate interaction with the RAG system by providing queries or prompts. The user’s request is the starting point for generating relevant responses and content based on the information received. For example, in an educational system, users may be students looking for content for video lessons or answers to specific questions..

**Retriever:**  
The retriever is responsible for searching a large amount of data to find the most relevant data that matches the user's query.These component functions rely on the knowledge behind the browser to access nonparametric memory, where data, text, or other information is stored. It uses methods such as similarity search, embedding based retrieval, or keyword matching to extract relevant information or data for further processing.

**Data Source:**The data source is where the user goes to find relevant information. It can be composed of different types of data, such as text, documents, video metadata, and even multimodal data, including visual and audio content. For example, in the case of a video lecture, the data can include indexed speech, subtitles, and image metadata to help the browser find relevant times for the user’s comments.

**Response:**The response is the final output to the user. It represents the data generated from the retrieval and generation components. In the context of RAG, the response is important and relevant because it contains real information provided by the sources, rather than relying on the prior knowledge of the LLM. For example, in learning, the answer can be a summary or explanation according to the content of the video received according to the user's query.

**LLM:**This section is explained in the Terminologies Section. Together, these components form a system that combines the retrieval of factual data with generative capabilities, enhancing the quality of the responses.

**A diagram of a flowchart

Description automatically generated**

Fig. 1. Retrieval Augmented Generation Architecture

**Terminologies:**

**Large Language Models (LLMs):**Large language models (LLMs) are artificial intelligence tools designed to understand, generate, and process human language. These models are typically built using deep learning techniques, particularly transformer architectures that allow them to process large datasets and learn language patterns. Trained on general knowledge, LLM can predict the next word in a sentence, complete sentences, translate words, and generate human-like responses to text messages. Masters technologies like ChatGPT have revolutionized word processing by making text more understandable and human-like. These models have become powerful tools for tasks such as text generation, translation, and interactive development. However, LLMs face significant challenges, particularly with regard to bias in teaching materials. Addressing these issues is important to ensure fairness and equitable use. While the future of the LLM is promising, continued research is essential to improve its interpretation, security, and ethics. Balancing technological advances with new responsibilities will be critical to their success.[[11](#bookmark=id.7q5zsq9dqylb)][[19](#bookmark=kix.9cn77cwxvzba)]

**Retrieval Augmented Generation (RAG*)*:**Retrieval-Augmented Generation (RAG) is a combination of two basic functions: retrieval and generation. This method is particularly suitable for generating accurate and more informed answers by providing important external information before generating the text. Retrieval: The system searches large databases (such as academic or knowledge databases) to find information relevant to the user's query. RAG gathers facts or information from external sources, rather than relying solely on a prior knowledge model. Generation: After receiving relevant information, the system combines the returned information with the knowledge model to produce a coherent answer. This step ensures that the answer is both accurate and meaningful. Various methods have been explored to implement the RAG model, making it suitable for many applications in different formats.[[12](#bookmark=id.d6hiu0m2thf9)]

**Multimodal Retrieval and Generation:**Multimodal representation learning involves processing and integrating information from multiple modalities (such as text, images, and audio) to create representations that are shared across multiple forms of information. Early approaches to multimodal processing relied on models designed for specific tasks. Recently, the model has shifted to the pre-training-fine tuning paradigm, which has shown great promise. This includes pre-training models on big data that use unsupervised or self-supervised tasks to capture general knowledge and can be fine-tuned for specific tasks. Transformer-based architectures, particularly the integrated model that combines multiple transformers with a unified interface, have become common in many predefined directions. These models represent the state of the art (SOTA) for performing cross-modal tasks due to their ability to learn rich representations. As the field of deep learning deepens, so does the effectiveness of structuring activities that contain different types of information. The versatility of this model becomes evident as multimodal datasets and data covering everything from e-commerce product names to social media posts and short videos continue to grow. Despite this progress, most existing algorithms still focus on single-modal representation learning or word recognition for cross-modal retrieval tasks.   
However, with the advancement of visual language modeling, large language models (LLM), retrieval augmented generation (RAG), and multimodal LLM, new opportunities for representation of different genres and backgrounds have emerged[[14](#bookmark=id.dg6cba948njo)][[18](#bookmark=kix.tr5ayuefcow2)].

# Literature Review:

# RAG for audio:

**Audio Generation:**

Recent designs emphasize the integration of multimodal learning and self supervised techniques to achieve high-quality audio products. New techniques such as AudioGen 2.0[[9](#bookmark=id.9pfjf0p5ldr9)] leverage transformer-based architectures to improve the quality of production using retrieval-based models. This model uses large-scale pre-trained dense retrievers like CLAP 2.0[[4](#bookmark=id.wr4v6m9ojfet)] to retrieve highly relevant caption-audio pairs, which enhances the semantic consistency of the generated audio. Next is a powerful audio generator that can produce detailed and recognizable audio-based descriptions. Other models such as SoundGen have introduced adaptive retrieval mechanisms that update during training, allowing for dynamic improvements in retrieval quality based on real-time learning. Make-An-Audio 2 uses a similar retrieval approach but incorporates advanced generative adversarial networks (GANs)[[1](#bookmark=id.lb2mcyxb36rl)][[3](#bookmark=id.3j65d0gg1g9o)] to improve the accuracy of sound reproduction based on different models.

**Audio Captioning:**

Recent developments in audio captioning emphasize multimodal fusion and use transformers to better understand the context. New models such as RECAP 2.0[[10](#bookmark=id.d6hiu0m2thf9)] now integrate the CLAP-X receiver, which performs multiple noise processing by learning from large datasets. This model uses different learning methods to optimize subtitles and audio files. Models such as AudioCapT5 use a T5-based architecture that allows the integration of different components, including CLAP-X and Audio CLIP, to generate consistent subtitles. These retrievers are now augmented with self-supervised learning methods that refine caption relevance and diversity. For example, the likes of listen2cap use a large audio sampler from Audiolm to create distinct words that capture small sound changes. In addition, new research is transforming audio into various media (e.g., spectrograms, embeddings) and then using large language models (LLMs) like GPT-4 to create complex subtitles. This method uses deep retrieval models like CLIP-Audio to optimize audio and text in a shared latent space to improve retrieval and rendering quality

1. **RAG for video:**

**Video Generation:**

Recent advances in video processing retrieval augmented generation (RAG) focus on the combination of multiple learning modes and pre-training models to improve video integration. Modern models such as Make-A-Video use dense clients such as CLIP-ViT to retrieve relevant videos and text from large databases based on input instructions. These retrieved clips are used to organize electronic-based video generators such as Video Transformer and latent propagation video models to create content-rich and consistent video output. The aspect of the cross-modal transformer allows for the proper merging of retrieved video clips and data files, thus improving the temporal consistency and semantic alignment of the generated videos. In addition, models such as Video Gen 3.0 integrate GAN-based architectures with latent propagation models to improve video quality and realism, while Video LM enables the use of elaborate retrieval mechanisms by leveraging large-scale pre-trained video language models.

**Video Captioning:**

Video subtitles are optimized using pre-trained visual and self-learning techniques. Modern models like VidCapGPT and VideoT5 use backup tools like CLIP-ViT and Align-ViT to store relevant data and information from big data sources, increasing the accuracy and diversity of the text. This model incorporates Transformer-based architectures like T5 and GPT-4 to generate sequential and accurate content. The latest system also includes a number of dynamic changes that pay attention to combining video, audio, and text to achieve better compatibility across formats. For example, VidCAP 2.0 leverages deep retrieval models like ViT-CLIP to align video frames and captions in a shared latent space. This improves the ability to capture subtle details in the film, providing detailed and accurate information. Video2Text also uses contrastive learning to better align video embeddings with text, creating more descriptive and context-aware subtitles. Existing models can handle complex situations such as multi-scene videos by extracting relevant visual and physical information from large video files and using this to retrieve subtitle characters. Some state-of-the-art methods are also exploring the use of RAG for video recording, where video clips or clamps are used to capture the structure of short content that includes important events and actions in the original context.

1. **RAG for text:**

**Question Answering:**

The recent RAG QA approach focuses on improving retrieval accuracy and response accuracy by combining dense sentence retrieval with large pre-training samples such as T5 and GPT-4. Models such as REALM 2.0[[20](#bookmark=kix.hqdjizke0x8t)] and RETRO++ use end-to-end differential retrieval and listener-based fusion to match retrieval to queries. TOG 2.0 and KAPING++ improve LLM by integrating knowledge graphs and improving thinking about complex problems. Multilingual models such as mBERT-RAG leverage cross-lingual access to enable quality QA international datasets. Technologies such as Personal RAG++ also improve the ability of the model to be optimized to restore the real state.  
  
**Fact Verification:**

In fact verification, models such as VERIFI-RAG and FactRAG++ leverage the intensive retrieval of knowledge verification repositories and use entity-centric retrieval for instant claim validation. Recent developments in knowledge RAG have focused on multi-hop access and truth-aware transformers to cross-validate data from multiple sources to provide more reliable detection. Hybrid models such as FactGraphQA combine text and image recognition to improve the accuracy of fault detection.

1. **RAG for Image:**

**Image Generation:**

RAG has been used for image generation by enhancing the model's ability to retrieve and integrate relevant visual data. Models such as Imagen-RAG and DALL-E[[6](#bookmark=id.vvtmulhwjrzc)] RAG use retrieval technology to search for image-text pairs from large text-based databases. This model collects similar images and links and uses them to adjust the design model (such as a propagation model) to obtain a better image. Make-A-Scene RAG takes this a step further by combining user drawings or designs, taking similar elements, and combining them into designs to create complex and more context-sensitive images.

**Image Captioning:**

RAG allows the model to increase the accuracy of the description by preserving similar images and texts in the image collection. CLIP-RAG and ALIGN-RAG use different learning methods to extract similar images from large image datasets, and the resulting models help guide the naming process. The model can generate text that contains not only correct content but also other content by preserving the same content. This approach increases the efficiency of generating labels for complex images or abstract images where accurate description is important.

**Visual Question Answering (VQA**)**:**

Models such as VQA-RAG for VQA provide a backup process with LLM to improve the accuracy of responses to image queries. The model provides more informed answers by providing relevant images and image readouts from large databases. BLIP-RAG uses retrieval-enhanced image transformers to better match image content to questions, retrieve similar images, and highlight annotations to improve the final answer. These advances make query processing in the visual domain more contextual and accurate, especially when dealing with small differences or unknown image content.

The literature survey explores advancements in Retrieval-Augmented Generation (RAG) across audio, video, text, and image modalities. In audio generation, models like AudioGen 2.0 utilize transformer-based architectures and pre-trained dense retrievers like CLAP 2.0 to enhance semantic consistency in generated audio. Audio captioning models such as RECAP 2.0 leverage multimodal fusion and transformer-based architectures, incorporating self-supervised methods to generate diverse and contextually relevant captions. Similarly, in video generation, modern models like Make-A-Video integrate dense retrievers and cross-modal transformers to ensure temporal and semantic alignment, while GAN-based architectures improve video realism. Video captioning techniques, including VidCapGPT, employ retrieval-augmented methods with transformers to enhance the accuracy and detail of subtitles.

In text-based RAG, models like REALM 2.0 and RETRO++ focus on improving question-answering by combining dense sentence retrieval with large-scale pre-training, while fact verification models such as FactRAG++ emphasize multi-hop reasoning for reliable claim validation. For images, RAG facilitates image generation by retrieving relevant visual data for models like Imagen-RAG and DALL-E RAG, enhancing design consistency. Image captioning methods, such as CLIP-RAG, improve label generation by aligning image-text pairs, while visual question-answering models like BLIP-RAG leverage retrieval-enhanced transformers for context-aware and accurate query responses. Collectively, these advancements showcase the versatility and impact of RAG across various domains.

# Methodology:

# Video Retrieval

This section outlines the proposed methodology for enhancing learning using video retrieval through Retrieval Augmented Generation (RAG). The approach focuses on converting video content into text and utilizing models to improve access to key learning moments in educational videos.

**Video to Text:**

The process of converting video content into data files involves the use of automatic conversion tools or speech recognition. It starts with the extraction process, where the audio track is separated from the video file for analysis. Speech recognition is then used to use deep learning models trained on big data to convert spoken words into text. These advanced models can provide accurate text even in difficult situations where there are loud voices or multiple speakers. Advanced techniques such as noise reduction can be used to improve the recording quality, especially in videos with poor audio quality.

**Mapping Sentences with Timestamp:**

To facilitate the effective use of specific sections from the video, each transcribed sentence is followed by the corresponding time. The process begins with the paragraph section, where the text is divided into main sections such as sentences or paragraphs. Following this, time alignment is performed, mapping each text segment to the specific time frame of the video it corresponds to. This alignment allows access to specific moments in the video based on the text. Some systems can also use models that include word or sentence level, allowing for better control of the recovery process. This ability is particularly useful in educational settings where detailed knowledge of specific topics is important for effective learning.

**Storing text data:**

Store recorded and time-mapped text for easy and efficient querying and retrieval. This includes creating a database where text and time are organized in formats such as JSON, XML, or relational data. This organization allows for rapid and efficient recovery. The study also used a scale that evaluates the sentence according to time setting, making it easier to determine the impact of the text and the impact of the video. Technologies such as full-text searches or retrospective evaluations can further enhance the research. As learning systems evolve to include larger video archives, data processing capabilities, and cloud storage solutions like Elasticsearch, they make it efficient to add more data while maintaining functionality and usability.

**Applying prompts on Stored Text:**

Once the text is stored, natural language processing (NLP) is used to interpret and respond to user prompts. The process begins with query-based access, where the user enters a query and the system uses search algorithms such as GPT, BERT, or RAG to search the relevant data collection. Advanced NLP technology then generates context-aware responses by analyzing the meaning of the question and determining the best sentences that provide the response or message that was decided to be worked on. This approach improves overall learning by ensuring that users receive accurate and relevant information.

**Video Retrieval using generated answer:**

Once the response is generated, the system matches the generated text with the stored text to determine the relevant time. These time logs are used to store video links, allowing users to see the time most relevant to their query. This system enhances learning by providing access to valuable video content. The system can also highlight key moments in the video, focusing on key learning points related to the user’s questions.

1. **Video Summarization**

The proposed method focuses on generating text from video content using the Retrieval Augmented Generation (RAG) method. The process of converting video to text, storing it for easy access, and using feedback from users to create concise content.

**Video to Text:**

The process of converting video content into data files involves the use of automatic conversion tools or speech recognition. It starts with the extraction process, where the audio track is separated from the video file for analysis. Speech recognition is then used to use deep learning models trained on big data to convert spoken words into text. These advanced models can provide accurate text even in difficult situations where there are loud voices or multiple speakers. Advanced techniques such as noise reduction can be used to improve the recording quality, especially in videos with poor audio quality.

**Storing Text Data:**

Once the transcription is complete, the text is stored in a format such as JSON or a database. The text may also include timestamps to maintain compatibility with the video. This step ensures that the output is properly accessible and can be retrieved when needed. This way of composing the text is important to quickly and accurately provide the rest of the summary or other questions that the user has decided on.

**Applying prompt on Stored Data:** Once the text is stored, users can enter specific questions or queries, such as requesting a quote or finding answers to specific questions from the video. The prompt is processed by a retrieval system, which scans the stored text and retrieves the most relevant part based on the query. This retrieval process allows users to find information that directly addresses their needs, whether it is a specific topic, detailed information, or answers to questions.

**Summary Generation:** Finally, a text summarization model generates a concise summary of the retrieved content. The compilation process can be either extractive (key sentences are selected directly from the text) or abstract (the structure of a new sentence that shows the main meaning). The resulting summary provides a condensed summary of the content by summarizing the important points and key ideas in the film. This makes the video information more understandable and accessible, especially for educational purposes.

This approach provides a way to summarize video content through transcription, prompt-based retrieval, and text generation, thus promoting effective learning by providing concise and relevant content based on the user's needs.

# Challenges and Limitations:

**Challenges:**

Despite the promising advancements in RAG, several challenges

remain to be addressed:

* Data Quality and Quantity: The effectiveness of RAG depends on the quality and quantity of the underlying data. Ensuring the accuracy, correctness and diversity of the data collected is important to produce accurate answers and information
* Computational Cost: RAG models can be computationally expensive, especially when dealing with large datasets and complex queries. This can limit the scalability and availability of application programs.
* Retrieval Efficiency: Efficient retrieval of relevant data from large data sources is crucial to RAG’s operation. It is quite difficult to develop fast and accurate search algorithms.
* Bias and Fairness: RAG models can introduce biases into training data, which can lead to bias or discriminatory outputs. Addressing bias and ensuring fairness in RAG systems is critical to ethical behavior.
* Explainability and Interpretability: Understanding how the RAG model reaches its conclusions is crucial to building trust and ensuring accountability. Developing strategies to explain the logic behind RAG-driven responses is a challenging area of research.
* Privacy and Security: Protecting user privacy and data security is an important consideration when implementing the RAG model. Ensuring that sensitive information is handled appropriately is crucial to responsible use.
* Domain Specificity: RAG models will struggle to succeed in certain areas where domain-specific knowledge is essential. Customizing a RAG model for a specific site requires careful consideration of materials and methods.

**Scope:**

Despite these challenges, the future of RAG is promising, with numerous opportunities for innovation and improvement:

* Multimodal Integration: Expanding RAG to incorporate multiple modalities, such as images, audio, and video, increase its functionality and be suitable for many tasks.
* Continuous Learning: Developing RAG models that can continuously learn and adapt to new information, allowing them to stay up-to-date with evolving knowledge and trends.
* Explainable AI: Advancing techniques for explaining RAG model decisions, making them transparent and accountable.
* Ethical Considerations: Addressing bias and fairness in RAG models to ensure their ethical use and promote social good.
* Real-world Applications: Discover and develop new applications of RAG in education, healthcare, customer service, scientific research and other activities.
* Integration with Other Technologies: Combine RAG with other AI techniques, such as reinforcement learning and generative adversarial networks, to create even more powerful and versatile systems.
* By addressing these challenges and seizing the opportunity, RAG has the potential to lead to significant advances in many areas by revolutionizing the way we capture, process and produce data.

# Conclusion:

This survey demonstrates the potential of Retrieval Augmented Generation (RAG) models to improve student learning by extracting high-quality video content and documenting the content. The RAG system makes difficult learning content more accessible by providing clear and concise content using multimodal information. The combination of retrieval mechanisms and generative models leads to accurate and timely indexing of video lectures, encouraging students to quickly and effectively retrieve important information. The research results show that the RAG model can play an important role in today's education by improving learning, especially in the case of large non-standard data such as video lectures. However, challenges such as optimizing accuracy and controlling various devices remain for future research. Further advances in knowledge-based learning tools, particularly improvements in search and design strategies, will be key to unlocking the potential of RAG systems for paper learning. As the industry matures, these models are likely to become the basis for AI-powered learning, paving the way for more personalized and effective learning.

**References**

1. Chakraborty, T., KS, U. R., Naik, S. M., Panja, M., & Manvitha, B. (2024). Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art. *Machine Learning: Science and Technology,* 5(1), 011001.
2. Dakshit, S. (2024, October). Faculty Perspectives on the Potential of RAG in Computer Science Higher Education. In Proceedings of the 25th Annual Conference on Information Technology Education (pp. 19-24).
3. De Souza, V. L. T., Marques, B. A. D., Batagelo, H. C., & Gois, J. P. (2023). A review on generative adversarial networks for image generation. *Computers & Graphics*, 114, 13-25.
4. Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. (2023, June). Clap learning audio concepts from natural language supervision. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
5. Elmessiry, A., & Elmessiry, M. (2024). NAVIGATING THE EVOLUTION OF ARTIFICIAL INTELLIGENCE: TOWARDS EDUCATION-SPECIFIC RETRIEVAL AUGMENTED GENERATIVE AI (ES-RAG-AI). In INTED2024 Proceedings (pp. 7692-7697). IATED.
6. Ge, Y., Xu, J., Zhao, B. N., Joshi, N., Itti, L., & Vineet, V. (2022). DALL-E for Detection: Language-driven Compositional Image Synthesis for Object Detection.

1. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... & Hassan Awadalla, H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. arXiv. *arXiv preprint arXiv:2302.09210*.
2. Kalyan, K. S. (2023). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 100048.
3. Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., ... & Adi, Y. (2022). Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352.*
4. Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.
5. Liu, J., Dong, Y., Li, S., Li, Z., & Mo, Y. (2024). Unraveling large language models: From evolution to ethical implications-introduction to large language models. *World Scientific Research Journal*, *10*(5), 97-102.
6. Liu, S., Cho, H. J., Freedman, M., Ma, X., & May, J. (2023). Recap: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation.
7. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., ... & Sun, L. (2024). Sora: A review on background, technology, limitations, and opportunities of large vision models.
8. Manzoor, M. A., Albarri, S., Xian, Z., Meng, Z., Nakov, P., & Liang, S. (2023). Multimodality representation learning: A survey on evolution, pretraining and its
9. Modran, H., Bogdan, I. C., Ursuțiu, D., Samoila, C., & Modran, P. L. (2024). LLM intelligent agent tutoring in higher education courses using a RAG approach. July. https://doi. org/10.20944/preprints202407, 519, v1.
   1. applications. *ACM Transactions on Multimedia Computing, Communications and Applications, 20(3), 1-34.*
10. Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121.
11. Thüs, D., Malone, S., & Brünken, R. (2024). Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature. Frontiers in Psychology, 15, 1474892.
12. Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X. L., Qin, C., ... & Joty, S. (2023). Retrieving multimodal information for augmented generation: A survey*. arXiv preprint arXiv:2303.10868*
13. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
14. Zhu, Y., Ren, C., Xie, S., Liu, S., Ji, H., Wang, Z., ... & Pan, C. (2024). REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models..