

A Simple Way to Achieve Dialogue Interviews and Interactive

Wenbo Zhong

*School of Computer Science, University of Western Australia, Perth, Australia
24247407@student.uwa.edu.au*

Abstract: Using text, large language models (LLMs), and retrieval-augmented generation (RAG), the modern spoken conversation model has higher intelligence than the conventional model. These sophisticated spoken conversation models can capture the timbre and stylistic aspects of speech in addition to comprehending audio, music, and other speech-related elements like GPT-4o. This research investigates a simple method for allowing dialogue-based interviews and interactive experiences. By combining LLMs with a retrieval mechanism, the suggested approach allows for contextual, precise, and interesting conversational exchanges by dynamically sourcing pertinent knowledge from an external knowledge repository or predetermined database. To enable meaningful interactions between people and AI, the system places a strong emphasis on scalability, ease of development, and natural language understanding. The framework is appropriate for a variety of applications, including customer service, healthcare, education, and recruiting, since it allows for domain-specific querying, real-time adaption, and conversation customization. Evaluation metrics show that integrating LLMs with RAG improves the depth and dependability of conversation systems by focusing on quality, coherence, and relevance. Additionally, the approach is designed to be scalable and cost-effective, requiring minimal infrastructure while achieving robust functionality. This work provides a practical and accessible pathway for creating interactive and knowledge-driven conversational systems that can meet the growing demand for intelligent, user-centric communication platforms across diverse domains.

Keywords: RAG, LLM, Chatbot.

1. Introduction

With the emergence of Large Language Models (LLMs), Natural Language Processing (NLP) has garnered more and more scholarly interest throughout the years. How people interact with machines has changed dramatically in recent years due to developments in artificial intelligence, especially in the area of NLP. LLMs, one of these developments, have become potent instruments that can produce text that is contextually aware, logical, and resembles human speech. These days, there are a lot of task-oriented LLMs available that can handle transcribed material, such as discussions and interviews. However, there are still misunderstandings and distortions in the LLMs' answers to user inquiries because of the intricate linguistic characteristics of that [1].

Although LLMs are excellent at comprehending and producing language, they can struggle to deliver precise and contextually relevant information, particularly when working with specialized fields or data that is changing quickly. Retrieval-Augmented Generation (RAG), a hybrid technique

that blends the accuracy of information retrieval systems with the creative powers of LLMs, has been presented to overcome this constraint. To create and execute conversation systems that take use of the synergy between LLMs and RAG, this research examines a simple yet efficient approach. The system may produce contextually rich and factually correct replies by dynamically accessing pertinent material from external repositories or specified knowledge sources by integrating retrieval methods into the conversational framework. This method improves the quality and dependability of conversation systems while also enabling them to be tailored to a variety of applications. From the first voice assistants like Alexa 1, Siri 2, and Google Assistant 3 to the newest intelligent conversation systems like GPT-4o, the spoken conversation model is one of the most straightforward forms of human-computer interaction. A conversation system that can produce intelligent verbal answers based on input speech is the fundamental definition of a spoken dialogue model. The voice mode serves as both the input and output interface for human-computer interaction in the spoken conversation paradigm. Conversely, the conversation system demands that the model have specific textual intelligence, such as the capacity to comprehend human social knowledge and produce thoughtful and professional answers.

This research primarily employs three novel techniques to enhance the capacity of big models to comprehend interview material and content development to accomplish these two goals: Innovative RAGs 2) Self-prompting LLM 3) Extraction of triple text.

The technical features of the project will be briefly discussed in this article. The following is a summary of our primary contributions:

The original text will be converted into context, question-answer pairs, and summaries for retrieval using self-prompt extraction, which will be used to extract possible semantics from the interview corpus. And use the LLM-based Reranker module to rank context, question-answer pairs, and summaries in a hierarchical fashion using LLM's semantic understanding capabilities. Then, combine these ranking similarities to get the final retrieval results. Lastly, provide a Cot-based conversation robot module that makes full use of Reranker's ranking outcomes and multi-level semantic advice for LLM to respond to queries.

2. Method

GPT-4 was assessed using a range of tests that were initially created for humans in order to gauge its performance in these kinds of situations. It does rather well on these tests and frequently outperforms the great majority of human test takers. As an illustration, GPT-4 scores in the top 10% of test takers on a mock bar exam. GPT-3.5, on the other hand, is in the lowest 10% [2].

For hardware, this paper needs (1) Processor (CPU): The speed of database operations depends on the CPU's performance. For improved performance, a multi-core CPU with a clock speed of 2.0 GHz or higher is advised. (2) Memory (RAM): The size, complexity, and anticipated demand of the database all affect how much RAM is needed. A minimum of 2 GB of RAM is typically advised, but larger databases or high-traffic situations can necessitate much more. (3) Storage: To store the program and any database backups or work with data files, this needs enough storage space. The size of the databases accessed determines how much storage is needed. (4) Network: Make sure the machine has a reliable network connection. A dependable internet connection is required.

2.1. Introduction of LLM

The notable examples are BERT, GPT, and T5. Large-scale datasets, which frequently include text from books, journals, webpages, and other publicly accessible sources, are used to train LLMs. They may do tasks including answering questions, translating languages, summarizing material,

creating original text, and even programming code by mastering patterns, syntax, context, and semantic linkages, as Figure 1 shows.

Before transformers, NLP depended on models that had trouble with long-term dependencies, such as MLP, CNN, RNN, LSTM, and GRU. Transformers are the basis for contemporary big language models since they addressed these problems using a self-attention mechanism. And the transfer Learning method enables AI models to adapt their expertise from one related activity to another. It enhances performance without requiring the model to be trained from the beginning. Scaling can train data, a larger model size, and more processing power are the three ways that AI models perform better when scaled. This scaling aids in the creation of AI systems with greater capability [3].

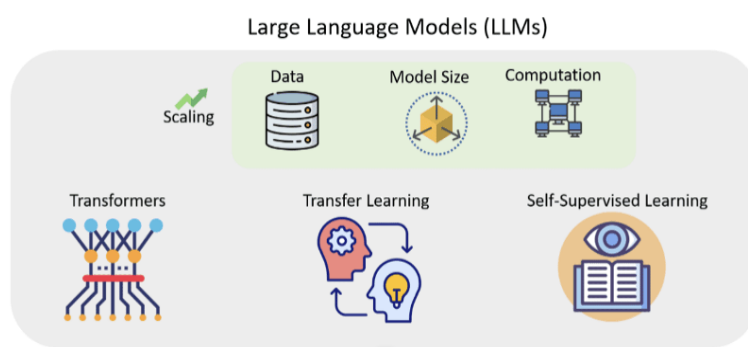


Figure 1: Image is from the survey paper on OpenAI LLMs [3].

2.2. RAG

An artificial intelligence (AI) system called a Large Language Model (LLM) is made to comprehend and produce human language. Natural language processing (NLP) has been transformed in recent years by a family of models known as transformers, which constitute the foundation of LLMs [4].

An AI architecture called Retrieval-Augmented Generation (RAG) combines text generation and information retrieval to improve the calibre, precision, and applicability of generated replies. Facebook AI (Meta) created it to overcome the drawbacks of conventional language models that only use previously learnt information.

According to Shuster, RAG incorporates two essential elements: Retriever: Looks for pertinent facts from an outside knowledge source, such as the internet, databases, or publications (Figure 2). Generator: Produces replies based on both retrieved and pre-existing information using a language model. The models frequently generate hallucinations during conversation generation, which results in erroneous responses. These hallucinated facts are locked at the moment of model training, making it impossible to retrain the model, even if these massive language models store amazing knowledge within their neural network parameters. The Retrieval Augmented Generator is a strong tool that may still be used if the model is properly bootstrapped. An AI system called retrieval-augmented generation was created to improve the calibre of replies produced by large language models (LLM). To enhance the generating capabilities of these models, it uses external information sources, usually knowledge graphs or other knowledge bases, like Wikipedia. Through this integration, the model may integrate more external data during text production, increasing its accuracy and applicability while guaranteeing access to the most recent, trustworthy data [5].

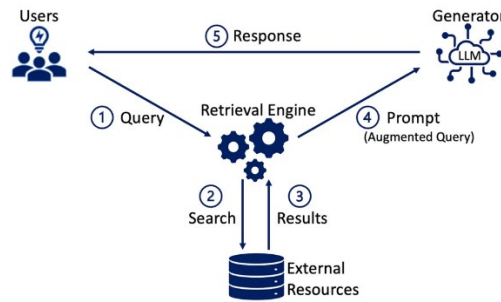


Figure 2: Introduction of RAG [4]

The retrieval-augmented generator. Guu first presented the retrieval-augmented generator (RAG) in their paper "REALM: Retrieval-Augmented Language Model Pretraining," in which they address the use of dense retrievers to improve language model pre-training (Figure 3). In REALM, the generator's signal backpropagates through the retriever after the retriever has accessed information from a textual knowledge corpus, such as the entirety of Wikipedia. They tested the suggested Retrieval-Augmented Language Model pre-training (REALM) on the difficult Open-domain Question Answering (Open-QA) challenge to confirm its efficacy. The results showed increases in interpretability and modularity, as well as a significant advantage over three benchmarks (4–16% absolute accuracy) [6].

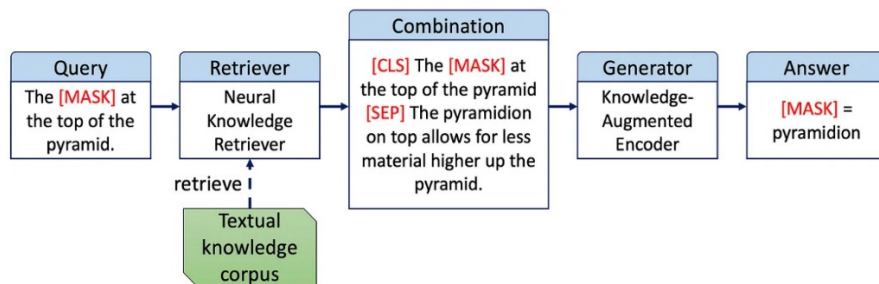


Figure 3: Introduction of RAG [4]

More researchers are now studying RAG as a result of the two papers mentioned above. Other works display particular implementation techniques based on the fundamental framework depicted in the first illustration. All of these previous studies show how RAG contributes to LLM in the following ways:

2.3. Structure of LLM Framework Architecture

Figure 4 displays, the information flow and interactions between the StructLLM Framework's components are highlighted in this figure. Here is a thorough explanation: A convolutional neural network called Backbone extracts visual characteristics at various granularities and concatenates them. Before transferring image characteristics to the prediction layer, Neck is a network layer stack that aggregates and mixes those variables, which is often FPN or PANET. The head is used to construct bounding boxes, identify categories, and predict image features. For example, the input sources, interview Text: This serves as the system's primary input. It stands for unstructured, raw data that needs processing. A user-provided query that the system analyses to provide a response. And essential Elements Self-Inspiring Extract which is important components from the interview text, such as summaries, question-answer pairings, and context. Provides the Chroma DB with these structured data items. The vector database, or Chroma DB, and context and QA pairs are stored in a vector database [7].

Serves as the method for obtaining pertinent facts in response to a query. A component for addressing queries that is Analyses user enquiries. Pulls pertinent information from Chroma DB, including summaries, QA pairings, and context. Then, create precise answers by using the Chain of Thought, which is a methodical approach to thinking. Reranker Based on LLM. Make sure the most pertinent information is sent to the QA bot by refining the obtained data. Improves accuracy and gives priority to the most pertinent information for the final response.

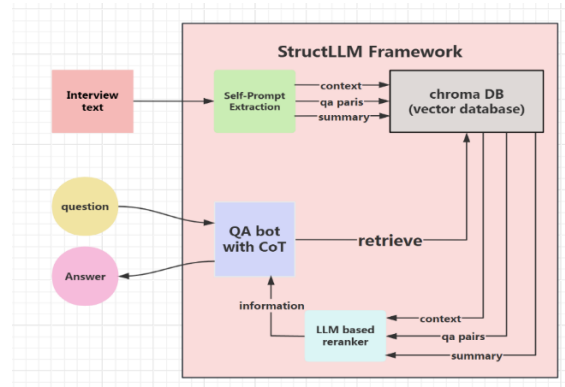


Figure 4: Illustration of the architecture of the LLM Framework (Photo/Picture credit: Original).

2.4. Self-Prompt Extraction, LLM-based Reranker and Chain of Thinking

To convert the source text into context, question-answer pairs, and summaries for retrieval, it employs Self-Prompt Extraction to extract possible semantics from the interview corpus.

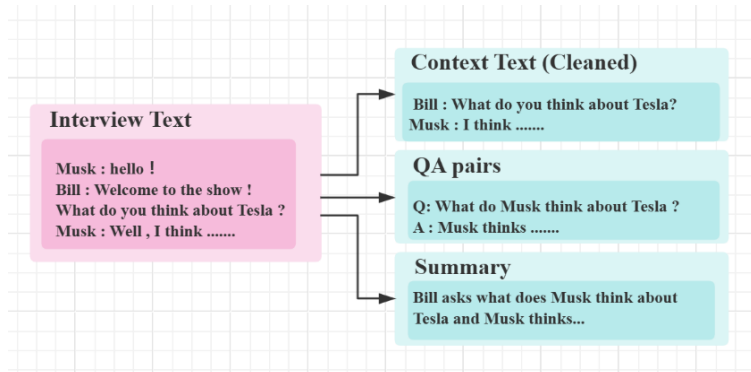


Figure 5: Illustrations of the architecture of the Self-Prompt Extraction. (Photo/Picture credit: Original).

To increase clarity and usability, an interview text can be processed into various formats, as Figure 5 shows. It begins with an unprocessed interview transcript and transforms it into a summary, a cleaned conversation, and a question-answer (QA) format. From the graph,

Interview Text (Raw Conversation) includes all of the interview's raw discourse, including introductions and conversational components. The context Text (Cleaned Version) preserves just the main discussion about the subject and eliminates extraneous elements like greetings. QA Pairs (Question-Answer Format) facilitate the retrieval of particular information; the discourse is converted into structured questions and answers. In general, it is a succinct synopsis of the main points of the discussion, excluding any extraneous information.

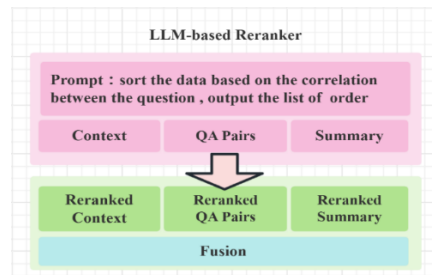


Figure 6: Illustrations of the architecture of the LLM-based Reranker (Photo/Picture credit: Original).

As in Figure 6, at the top of the graphic, the system employs a prompt that directs the LLM to **sort data** based on the connection between a query and certain components of the data. The purpose is to output a **list in order** of relevance or significance.

There are three primary parts to the input. The context is the more general text or data pertaining to the inquiry.

QA Pairs: queries and responses that could be connected to the question. A succinct synopsis of the information presented.

Reranking Procedure: After analyzing the input prompt, the LLM ranks each element according to how pertinent it is to the query. This step's output consists of reranked Context: The context is arranged according to how pertinent it is to the question. And reranked QA Pairs: These pairs were arranged according to how well they answered the question. The summaries were rearranged according to how closely they answered the question. The final output is produced by fusing the results after they have been reranked. To provide a thorough, well-organized response or summary that best fits the inquiry, this fusion may entail combining the reranked elements.

This system leverages the power of LLMs to prioritize and organize information, making it easier for users to find the most relevant data in response to a question. It can be particularly useful in applications like search engines, question-answering systems, and knowledge retrieval tasks.

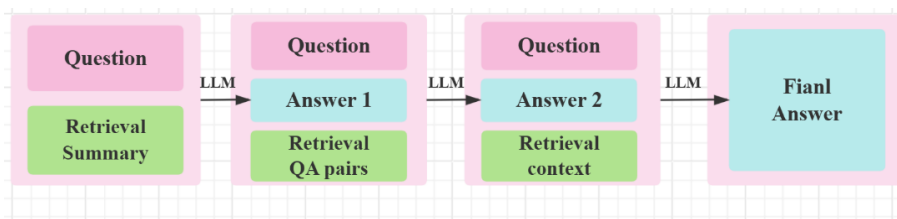


Figure 7: Illustrations of the architecture of the chatbot Chain of Thought (CoT) (Photo/Picture credit: Original).

CoT reasoning is depicted in Figure 7. It improves chatbot replies by segmenting the replying process into many parts and making incremental improvements.

The first is the Question and Summary of Retrieval, where the user asks a question to the chatbot. From a knowledge base, it obtains a summary of pertinent data (Retrieval Summary). To produce a first response, the retrieval summary and the query are processed by the Large Language Model (LLM). Obtaining the First Response and Further Context. To improve its comprehension, the chatbot uses the first response to extract relevant QA pairings. The LLM refines the answer using this further context. Retrieving the Second Answer and Additional Context. The chatbot improves its response even further by obtaining more context pertinent to the query. In addition, an enhanced answer is produced by a second repetition of the LLM processing. Final Answer Generation: To generate the Final Answer, the chatbot compiles all of the earlier answers and edits. By segmenting

the response process into several steps, Chain of Thought (CoT) reasoning enhances chatbot accuracy [8].

Retrieval-based improvement guarantees that responses are supported by pertinent data. The chatbot's response is improved through several iterations, becoming more precise and contextually rich.

2.5. Prompt

The prompt is a request delivered to the chatbot to acquire the required output. Here, the request is delivered in natural language, i.e. as text input. Here, the text input might be a command telling the chatbot to do a task or a question requesting information. In RAG-based AI, the prompt is carefully crafted to incorporate retrieved information from external sources, guaranteeing accurate, context-aware, and current outputs. The prompt is the structured input provided to the model to generate high-quality responses when building an AI system using a Large Language Model (LLM) with Retrieval-Augmented Generation (RAG).

First of all, the meaning of LLM and RAG prompt. An LLM's reaction is guided by the input text known as a prompt. Both user questions and knowledge that has been obtained from databases, search engines, or knowledge graphs are included into the prompts of a RAG-based AI.

RAG Prompt Structure:

A standard RAG prompt includes User Question: The initial question posed by the user. The context that was acquired from an external source includes pertinent publications, frequently asked questions, or structured data.

There are a few steps of the prompt. Step 1 is The Prompt User's First Question In order to retrieve pertinent information, the retrieval system receives the first cue, which is a direct user query. And step 2 Contextualization and Retrieval• From a database, API, or search engine, the AI extracts pertinent information, which is then put to the LLM prompt. Step 3, Iterative Prompting for Answer Refinement: The chatbot retrieves relevant QA pairs or further context to help refine the response. Step 4: Final Answer Generation with Summarization The best response is produced by structuring the final prompt with all the fine-tuned elements [3].

2.6. Tables of Detection Data

Table 1: Illustration of the Experiment

Model	Musk1 Accuracy	Musk2 Accuracy
GBT-4	2/10	3/20
Deepseek v3	3/20	1/20
GBT-4 + The paper model	7/10	13/20
Deepseek v3 + The paper model	7/10	15/20

Table 1 shows the comparison of different numbers of update variance. Part of the data is retrieved from the paper. The smaller the variance is, the better the performance. To prevent conflicts with the model training set, it has created two specific test sets to assess the language model's understanding of the two conversation corpora, Mask1 and Mask2. These test sets collect certain elements that are unique to the corpus.

3. Results and Discussion

The existence of these nonverbal cues during interviews may make it more difficult to comprehend answers. A little shift in tone, a moment of delay, or an unexpected facial expression can all have a

subtle impact on how someone interprets a message. Additionally, certain written representations could not fully convey the subtleties of an interviewee's intended meaning, which could result in misunderstandings. This disparity might be the result of formatting restrictions in the text or the lack of visual and aural clues that give the spoken word context [9].

This describes GPT-4, a huge multimodal model that performs at a human level on some challenging academic and professional standards. OpenAI created the sophisticated AI language model known as GPT-4 (Generative Pre-trained Transformer 4). It has improved thinking, creativity, and language comprehension over its predecessor, GPT-3.5. GPT-4 can perform more complicated queries, process and produce text more accurately, and even read pictures (in some versions). GPT-4 performs better than the great majority of documented state-of-the-art systems (which frequently incorporate task-specific fine-tuning) and surpasses current big language models on a set of NLP tasks. It discovered that enhanced talents can be seen in a wide range of languages, even though they are often tested in English. Demonstrate how our accurate predictions of GPT-4's loss and capabilities were made possible by predictable scaling [10].

GPT-4 poses additional hazards owing to greater capabilities and highlights some of the approaches and outcomes used to analyse and improve its safety and alignment. Though there is more work to be done, GPT-4 marks a big breakthrough towards generally useable and reliably deployed AI systems [11]. From the previous research, it found out that rich, unsupervised pre-training is a crucial component of many language comprehension systems, as empirical advancements brought about by transfer learning with language models have shown. These findings specifically make it possible for deep unidirectional architectures to be advantageous for applications with limited resources. Our main contribution is the extension of these results to deep bidirectional architectures, which enables a wide range of NLP tasks to be effectively handled by the same pre-trained model [12]. It describes GPT-4, a large multimodal model that performs at a human level on some challenging academic and professional standards. GPT-4 performs better than the great majority of documented state-of-the-art systems (which frequently incorporate task-specific fine-tuning) and surpasses current big language models on a set of NLP tasks. Discover that enhanced talents can be seen in a wide range of languages, even though they are often tested in English. Demonstrate how our precise forecasts of GPT-4's loss and capabilities were made possible by predictable scaling [13].

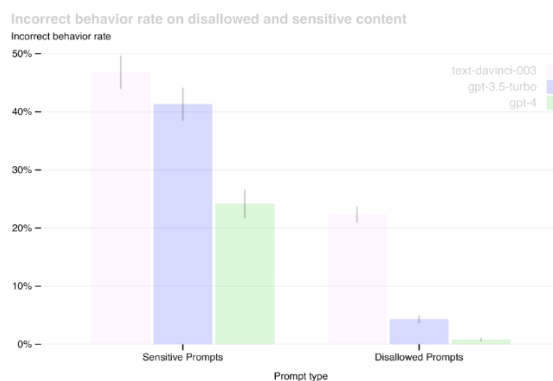


Figure 8: Inaccurate behaviour rate on sensitive and prohibited cues. Lower numbers are preferable. The rate of improper behaviour is significantly lower with GPT-4 RLHF than with previous models [13].

Figure 8 model-level interventions make it harder to elicit negative behaviour, but it is still feasible. For instance, "jailbreaks" are still used to create material that deviates from our usage policies. It's crucial to supplement these constraints with deployment-time safety measures like abuse monitoring and a pipeline for quick iterative model development as long as they persist.

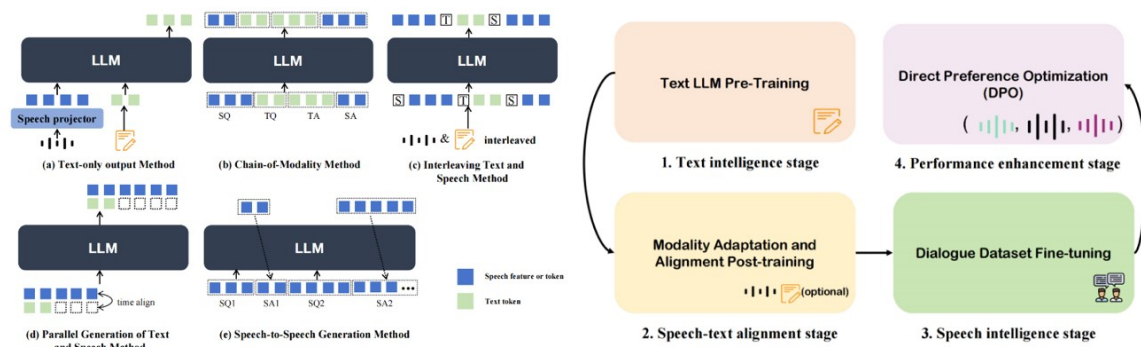


Figure 9: Diagram of Multi-stage Training Steps (right) and Categorisation Diagram of Spoken Dialogue Model Architectural Paradigms [13].

From Figure 9, an adaptation of text-based large language models (LLMs) into conversation systems with voice processing capabilities is the main topic of the Training Paradigm of Spoken Conversation Model section. The choice and layout of training paradigms directly impact the model's performance, multimodal alignment, and real-time performance.

4. Conclusion

Despite the importance of these factors, there is still a significant lack of rigorous academic research in this area. No thorough or established studies have conclusively examined the extent to which non-verbal factors affect the accuracy and interpretation of transcribed interviews; therefore, this remains a topic of interest for future research, with the potential to yield ground-breaking insights into the complex relationship between verbal and non-verbal communication. Additional research could ultimately improve our ability to accurately convey and interpret human interactions across various communication mediums by contributing to fields like linguistics, psychology, artificial intelligence, and media studies. This study provides a thorough examination and knowledge of human detection, gender detection, object detection, additional object detection, and the use of deep learning networks in human detection applications. With the emergence of Large Language Models (LLMs), Natural Language Processing (NLP) has garnered more and more scholarly interest throughout the years. With the paper read, here are the conclusions and summary from the previous studies.

References

- [1] Ji, S., Chen, Y., Fang, M., Zuo, J., Lu, J., Wang, H., Jiang, Z., Zhou, L., Liu, S., Cheng, X., & others. (2024). WavChat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.
- [2] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., & others. (2023). MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.
- [3] LLMNanban. (n.d.). What is a prompt. Retrieved from <https://llmnanban.akmmusai.pro/Introductory/What-is-a-prompt/>
- [4] Aryani, A. (2023, January 31). A brief introduction to retrieval-augmented generation (RAG). Medium. Retrieved from <https://medium.com/@amiraryani/a-brief-introduction-to-retrieval-augmented-generation-rag-4bd6e50da532>
- [5] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *arXiv*.
- [6] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-augmented language model pre-training. *arXiv*.
- [7] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., & Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

- [8] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. *arXiv preprint arXiv:1810.04805*.
- [9] Tay, Y., Wei, J., Chung, H. W., Tran, V. Q., So, D. R., Shakeri, S., Garcia, X., Zheng, H. S., Rao, J., Chowdhery, A., & others. (2022). *Transcending scaling laws with 0.1% extra compute*. *arXiv preprint arXiv:2210.11399*.
- [10] Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., & Gao, J. (2020). *Adversarial training for large neural language models*. *arXiv preprint arXiv:2004.08994*.
- [11] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., & others. (2022). *PaLM: Scaling language modeling with pathways*. *arXiv preprint arXiv:2204.02311*.
- [12] Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer normalization*. *arXiv preprint arXiv:1607.06450*.
- [13] Wang, B., & Komatsuzaki, A. (2021). *GPT-J-6B: A 6 billion parameter autoregressive language model*. *arXiv*.