

Article

Advanced Subjective Question Bank Generation Using Retrieval Augmented Generation Architecture

Amaan Sayed *, Mahendra Kanojia and Subhashish Nabajja

Department of Computer Science, Sheth L.U.J and Sir M.V College, Mumbai 400069, India;
kgkmahendra@gmail.com (M.K.); subhashishnabajja1575643@gmail.com (S.N.)

* Correspondence author: amaansayed1510231@gmail.com

Received date: 3 April 2024; Accepted date: 17 May 2024; Published online: 10 July 2024

Abstract: Large language models (LLMs) have shown great promise for various natural language processing tasks, including question generation. However, generating subjective questions that are relevant and informative remains a challenge. Existing approaches often rely on predefined question templates or manually crafted knowledge bases, which limit the diversity and quality of generated questions. In this paper, we propose a novel approach that leverages LLMs to generate subjective questions with the help of a custom knowledge base. Our knowledge base is constructed by extracting and embedding relevant information from a given text corpus. By combining the LLM's language generation capabilities with the domain-specific knowledge from the knowledge base, our system can generate more informative and contextually relevant subjective questions. Experimental results show that our approach beats the existing methods in terms of question quality and relevance. Specifically, the Google Gemini LLM achieved the highest score among the compared models, with an average rating of 3.92 out of 5 for question quality and an average relevance score of 0.90. Our approach has several advantages over existing methods. First, it does not rely on predefined question templates, which can limit the diversity of generated questions. Second, our custom knowledge base is constructed from a domain-specific text corpus, which ensures that the generated questions are relevant to the given domain. Third, our approach can be easily adapted to different domains by constructing a new knowledge base from the corresponding text corpus.

Keywords: large language models; generative pretrained transformers; question generation; natural language processing

1. Introduction

Large pretrained language models with greater efficiency on text generation tasks have been developed through recent artificial intelligence (AI) advancements, raising the question of whether we may use them to produce text completions that are valuable for education [1]. The establishment of AI as a subject in K–12 education puts additional demands on key players, particularly educators who direct the process of teaching and learning. Because of this, it's critical to know how prepared educators are to teach the new subject, as their level of preparedness will likely have a major impact on how well AI education proceeds [2]. Language learning frequently uses QA (Question and Answer) regarding a story's contents. Productive QA in a target language occurs when the student is asked to read a narrative, answer questions, and construct answer sentences using their plural language abilities. When doing this kind of quality assurance, the teacher will provide the learner with appropriate questions [3]. It takes a lot of time and effort to independently generate question banks for teaching purposes. It can be quite challenging for educational institutions, instructors, and material writers to create a comprehensive and varied set of questions. These challenges include the need for in-depth knowledge of the topic matter, the significant



time investment required, and the possibility of biases or inconsistent question designs. The growth of online learning platforms and the demand for personalized assessment materials have brought attention to the urgent need to find more efficient and scalable methods for question bank development. Generative Pretrained Transformer is one of the large language models, (GPT-3) have enabled tremendous advancements in natural language processing (NLP) in recent years [4–6]. Present day neural network-based learning models deploy incredibly large model architectures such as 175 billion parameters and train on colossal datasets such as that of nearly a terabyte of English text. This scalability makes it possible for LLM's to produce natural language more fluently and to be used for an extensive variety of additional activities without the need to manipulate their parameters [7]. Large Language Models have demonstrated their capacity to carry out novel tasks based on textual instructions after being trained on corpora of text [8]. The creation of an automated question bank generation system and learning support are the main objectives of this research. Because of the wide range of applications that large language models provide, their use in education has been suggested as a possible topic of study. Through the application of these models, people at all educational levels primary, secondary, tertiary, and professional development may be able to improve their experiences both teaching and learning. Furthermore, because every person has different learning needs, preferences, and skills, big language models present a special chance to deliver individualized and successful learning experiences.

Large language models can help primary school pupils strengthen their writing and reading comprehension (by offering grammatical and syntactic corrections, for example) as well as their writing style and critical thinking abilities. Using these models, educators can come up with questions and activities that challenge students to read and write critically as well as to evaluate and comprehend the material that is being presented to them [5]. By giving students recaps and clarifications of difficult literature, large language models can also help students build their reading comprehension abilities by making the information easier for them to read and comprehend [9,10].

Large language models may aid pupils in both lower secondary and higher secondary learn a language and writing styles for a variety of topics and courses, such as the field of physics literature and language acquisition, mathematics, and other subjects. These models can be used to create practice questions and tests that will aid students in comprehending, contextualizing, and remembering the information they are learning. Large language models additionally have the ability to assist students develop their problem-solving abilities by giving them explanations, specified solutions, and thought-provoking follow-up questions. This can enable students to comprehend the logic behind the solutions and foster analytical and creative thinking.

Large Language Models can help university students with their writing and research assignments as well as with the growth of their analytical and problem-solving abilities [11]. With the use of these models, students can quickly comprehend a text's essential ideas and arrange their thoughts for writing by creating highlights and sketches of the material. Large Language Models can also help students build their research skills by giving them resources and information on a certain subject while also making subtle allusions to undiscovered angles and hot subjects in the field. This can improve their comprehension and analytical abilities [10].

Large language models can be used in conjunction with text-to-speech or speech-to-text programs to empower learners with impairments, especially those who are visually impaired [12,13]. Language models can be utilized to construct comprehensive learning strategies with suitable help in tasks like flexible writing, interpreting and highlighting of important material in multiple formats, in addition to the previously identified group and remote counseling options. It is crucial to remember that using huge language models should be done so with the assistance of experts in the field, such as educators, speech therapists, and other specialists who can modify the technology to meet the unique requirements of students with disabilities.

ChatGPT and other Large Language Models have the potential to completely change education and support the teaching process. Some of the instances are provided below:

Teachers can utilize large language models to computerize the marking of student work for review and assessment purposes [5]. This involves highlighting the work's potential strengths and weaknesses, such as in papers, analysis work, and other written assignments. This can help teachers save a lot of time on activities like providing students with personalized feedback [14]. Moreover, plagiarism can be detected using LLMs, hence reducing the likelihood of cheating. LLMs can therefore assist teachers in pinpointing the areas in which pupils are having difficulty, leading to more precise evaluations of the problems and growth of student learning. The models' targeted training can be utilized to support students' success and present chances for ongoing growth.

Large language models may aid teachers create inclusive plans of action and lessons for their classes [15]. Faculty members can feed the models the body of documents that they wish to use as the foundation for a course. A course curriculum with a brief synopsis of each topic can be the result. LLM networks may

additionally offer questions and prompts that stimulate problem-solving and reflective thinking while encouraging participation from individuals with varying backgrounds and skill sets. Additionally, they can be utilized to create customized and targeted practice questions and tests, which can guarantee that students are grasping the subject matter.

2. Literature Review

The application of technology and artificial intelligence (AI) in education has been the subject of numerous studies in recent years. The study [1] highlighted recent AI advancements, particularly in pre-trained language models like GPT-3, for generating educational quizzes. It underscored the potential of AI-generated quizzes to enhance formative feedback and engagement in learning, despite challenges in generating high-quality distractor answers. In [2], researchers investigated teachers' preparedness and willingness to teach artificial intelligence in educational settings, as published in *Computers and Education: Artificial Intelligence*. The study [3] introduced automated methods for generating adaptive questions from English story content, encompassing various question types to cater to learners' diverse understanding levels, covering 80% of questions found in novice-level problem collection books, with a high proportion of semantically competent question sentences generated from Japanese junior high school textbooks. The study [4] investigated fine-tuning stability in biomedical NLP, revealing sensitivity to pretraining settings and proposing techniques to address instability, such as freezing lower layers for BERT-BASE models and layer wise decay for BERT-LARGE and ELECTRA models, ultimately achieving state-of-the-art performance across various biomedical NLP tasks. In [5], the article discussed the benefits and challenges of incorporating large language models in education, emphasizing the need for competencies to understand their limitations. It highlighted opportunities for personalized learning but cautioned about potential biases and the necessity of responsible integration. The article [6] proposed enhancing in-context learning for multi-span question answering with answer feedback, as presented in the arXiv preprint arXiv:2306.04508. The [7] paper exposed a vulnerability in large language models trained on private data, demonstrating a training data extraction attack capable of retrieving individual examples. Using GPT-2 as an example, the attack successfully extracted sensitive information, emphasizing the need for safeguards during training. The paper [8] introduced LLaMA, a series of foundation language models ranging. In [9], researchers evaluated the quality of student-generated short answer questions using GPT-3, presented at the European Conference on Technology Enhanced Learning. In [10], an Intelligent Question Bank System for English Grammar was announced, demonstrating how custom question banks and technology may work together. Prominent grammatical reference books like Baidu Encyclopedia provided the authors with English grammar data. The setup employed a five-layer structure. Ref. [11] carried out a thorough analysis of automated question creation for educational purposes, highlighting its applicability in contemporary classrooms. Ref. [12] explored enhancing large language models (LLMs) with speech. The paper [13] presented the Seq2SQL framework, that generated structured SQL queries via natural language using reinforcement learning. According to the study [14], a dashboard was created to help teachers keep an eye on student participation and provide timely interventions. These models have the ability to enhance student performance by assessing engagement with various activities and materials through integration within virtual learning environment (VLE) systems. In the study provided by [15], the impact of teachers' adaptations to formative classroom instruction in response to professional development feedback on students' academic success was examined by the writers of this research report. In [16], ChatGPT's performance on the US medical license exam was evaluated, bringing attention to the significance of big language models for medical knowledge assessment and education. Ref. [17] examined Large Language Models (LLMs) in the medical domain, such ChatGPT. Based on the available data, Ref. [18] looked at ChatGPT as an ideal illustration of LLMs in academic/scientific writing, medical care, and education. The study conducted by [19] examined the prospective fields for future research to comprehend the emergence and scalability of the emergent talents that have recently been found in language models as a result of scaling up. Using the VNHSGE English dataset, Ref. [20] assessed how well three well-known LLMs performed on challenges incorporating natural language processing, helping programmers and academics to gain a better understanding of each LLM's applicability in practical situations. Ref. [21] provided a thorough analysis of the assessment of big language models and stressed the importance of ethical frameworks and comprehensive evaluation. The paper [22] addressed accessibility issues with large language models (LLMs) despite their recent performance advancements. Ref. [23] investigated open-source Transformer techniques for question-answering systems in the cloud domain, providing a comparison with ChatGPT3.5-Turbo. In [24], Two new chatbot models were released: GPT4All-J v1.3 Groovy, licensed under Apache-2, and GPT4All-13B-snoozy, licensed under GPL. Trained on a diverse corpus, they improved upon previous releases by using larger and cleaner datasets. Ref. [25] proposed the Transformer, a novel network architecture based solely on attention mechanisms, eliminating the need for recurrent or

convolutional layers. The paper [26] introduced Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging from 7 billion to 70 billion parameters. In the study provided by [27], a comparison was made between Gemini Pro and GPT-4V in educational settings using visual question answering (VQA). GPT-4V showed significantly higher accuracy and performance in scoring student-drawn scientific models compared to Gemini Pro, suggesting its superior capability in handling complex multimodal educational tasks. In [28], Orca, a 13-billion parameter model trained to imitate LLMs' reasoning processes was introduced. The paper [29] introduced a personalized online practice system utilizing ChatGPT for question generation and feedback in education. With a prompt generator and text analyzer, the system processed student responses and integrated adaptive feedback to assess mastery. Ref. [30] evaluated the performance of three large language models (LLMs)—GPT-3.5, GPT-4, and Google Bard - on a neurosurgery oral boards preparation exam. The study of [31] analyzed the influence of Information Retrieval (IR) components on Retrieval-Augmented Generation (RAG) systems, representing a significant advancement over traditional Large Language Models (LLMs). Contrary to initial assumptions, their findings revealed that including irrelevant documents in the retrieval phase unexpectedly enhanced performance by over 30% in accuracy. This underscored the importance of developing specialized strategies to integrate retrieval with language generation models, laying the groundwork for future research in this field.

3. Research Methodology

3.1. RAG

Retrieval augmented generation (RAG) is a method aimed at enhancing retrieval accuracy and improving the quality of responses generated by large language models (LLMs) through the integration of data sourced from external repositories. Also known as in-context learning [31], RAG offers several notable advantages. Firstly, it ensures the delivery of up-to-date and precise responses by integrating real-time external data sources, thus mitigating dependence solely on static training data. Secondly, it aids in diminishing inaccurate responses or hallucinations by anchoring the model's output on pertinent external knowledge, potentially including citations for validation. Thirdly, RAG facilitates the creation of contextually relevant and domain-specific responses tailored to an organization's proprietary or specialized data. Finally, it proves to be an efficient and cost-effective approach compared to alternative methods of customizing language models with domain-specific data, as it does not necessitate extensive model customization, rendering it particularly advantageous for frequent updates with new data [32]. The architecture of the RAG is schematically depicted in Figure 1. The architecture comprises four primary phases:

- **Data Preparation:** This stage focuses on getting the data ready for the next phases.
- **Index Relevant Data:** This section is about organizing and pinpointing the most relevant pieces of information within the prepared data.
- **Information Retrieval:** This is where the diagram showcases how the key information is extracted for use.
- **LLM Inference:** The final stage demonstrates how a Large Language Model (LLM) draws conclusions and generates responses using the retrieved information.

The data flow starts with raw data files, such as PDFs, Word documents, and PowerPoint presentations, which undergo data cleaning to remove irrelevant information, fix errors, and format the data properly. Next, the cleaned dataset is segmented into smaller chunks for easier processing before being converted into embeddings, which are numerical representations capturing the meaning and context of the text. These embeddings are stored in a specialized Vector Search index, facilitating quick searches for similar or relevant data. Finally, the Large Language Model uses both the user query and the relevant data chunks to generate a tailored answer through LLM inference.

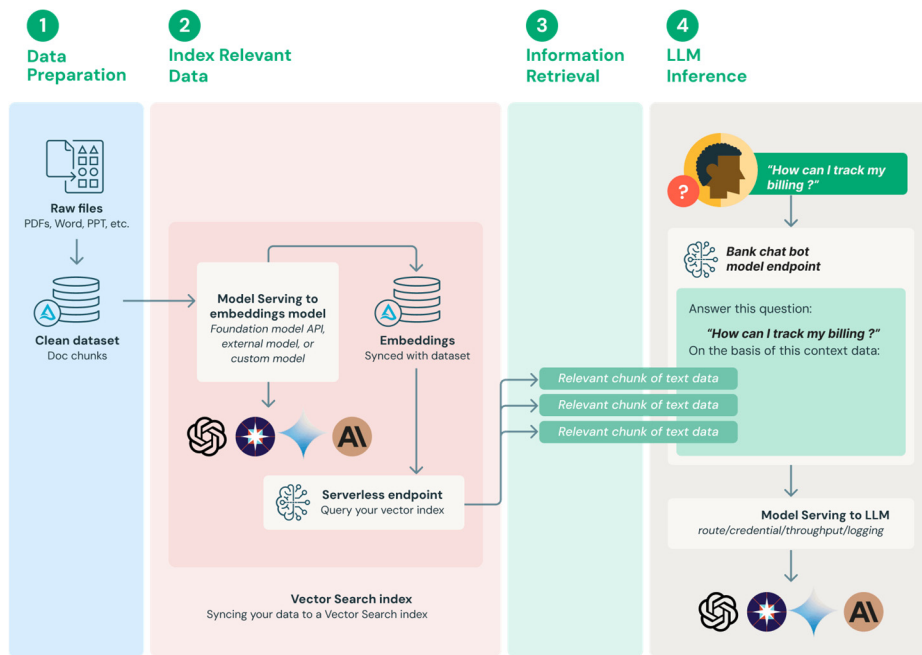


Figure 1. Reference architecture for RAG applications.

3.2. Transformer Architecture

The Transformer represents a deep learning (DL) model founded on a self-attention mechanism, which assesses the significance of individual segments within the input data. Its primary applications are observed in computer vision (CV) and natural language processing (NLP).

Similar to recurrent neural networks (RNNs), Transformers are engineered to handle sequential input data, such as natural language, and undertake tasks like text summarization and translation. However, unlike RNNs, Transformers process the entire input simultaneously. This is facilitated by the attention mechanism, which enables the model to concentrate on the most pertinent segments of the input for each output.

The Transformer architecture (see Figure 2) comprises two primary components: the Encoder and the Decoder.

1. **Encoder:** Initially, the input data, typically English text, undergoes processing within the Transformer. As the model inherently does not comprehend English language, each word in the text is converted into a unique numeric ID. This conversion is accomplished utilizing a predefined vocabulary dictionary, derived from the training data, which maps each word to a corresponding numeric index.
2. **Decoder:** Contrasting the Encoder, the Decoder operates with two inputs and implements multi-head attention twice, with one instance incorporating a “masked” mechanism. Moreover, the final linear layer within the Decoder is dimensioned to match the size (i.e., the number of units) of the target dictionary, such as the French language dictionary in this instance. Each unit in this layer is assigned a score, with the SoftMax function applied subsequently to convert these scores into probabilities, indicating the likelihood of each word's presence in the output.

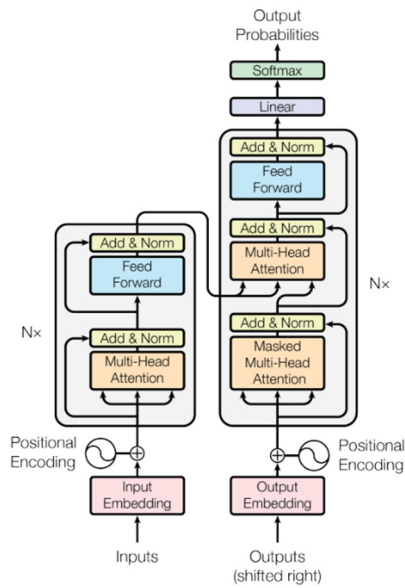


Figure 2. Transformer architecture [25].

3.3. Falcon

Introducing the Falcon series: 7B, 40B, and 180B parameters causal decoder-only models trained on a diverse, high-quality corpus primarily compiled from web data. The largest model, Falcon-180B, underwent training with over 3.5 trillion tokens of text - marking the most extensive openly documented pretraining run. Falcon-180B notably surpasses models like PaLM or Chinchilla and enhances upon concurrently developed models such as LLaMA2 [33].

The authors provide detailed evaluations [34], along with an in-depth exploration of the methods and custom tooling utilized to pretrain Falcon. Particularly, they elaborate on their custom distributed training codebase, enabling efficient pretraining of these models on up to 4,096 A100s on cloud AWS infrastructure with limited interconnect. Additionally, they release a 600B token extract of their web dataset, as well as the Falcon-7/40/180B models under a permissive license, aiming to promote open source eco-system of LLMs and accelerate the development of a collaborative ecosystem of large language models.

Both Falcon-7B and Falcon-40B models use a modified decoder-only transformer architecture. Modifications made to this architecture include[34]:

- Flash Attention
- RoPE embeddings
- Multi-Query Attention
- Parallel Attention and Feed-Forward Layers

Flash Attention is an attention algorithm used to reduce this problem and scale transformer-based models more efficiently, enabling faster training and inference. Flash Attention is an efficient and precise Transformer model acceleration technique proposed in 2022. By perceiving memory read and write operations, Flash Attention achieves a running speed 2–4 times faster than the standard Attention implemented in PyTorch, requiring only 5–20% of the memory.

Rotary Position Embedding, or RoPE, is a type of position embedding which encodes absolute positional information with rotation matrix and naturally incorporates explicit relative position dependency in self-attention formulation. Notably, RoPE comes with valuable properties such as flexibility of being expand to any sequence lengths, decaying inter-token dependency with increasing relative distances, and capability of equipping the linear self-attention with relative position encoding.

Multi-query attention offers a streamlined approach to multi-head attention by sharing keys and values across different attention "heads". This optimization significantly reduces the memory bandwidth requirements of incremental decoding compared to standard multi-head attention, where each attention head has its own set of keys and values. By sharing these parameters, multi-query attention minimizes both memory and computational costs, making it particularly advantageous for large models or scenarios involving incremental decoding. Despite its efficiency gains, multi-query attention remains effective across various tasks such as machine translation, question answering, and summarization, highlighting its potential to enhance the efficiency of Transformer models in diverse applications.

Parallel Attention In the context of neural machine translation, parallel attention refers to the use of multiple attention mechanisms operating in parallel on different parts of the input sequence. Instead of having a single attention layer that processes the entire input sequence sequentially, parallel attention divides the input into segments and assigns a separate attention mechanism to each segment. This allows the model to focus on different parts of the input simultaneously, potentially capturing more complex relationships and dependencies.

Feed-forward layers are a type of neural network layer that applies a non-linear transformation to its input. They consist of a linear transformation (e.g., a matrix multiplication) followed by a non-linear activation function (e.g., ReLU). Feed-forward layers are often used in neural networks to learn complex relationships between input features and to extract higher-level representations.

3.4. Llama

Llama LLM, a series of large language models (LLMs) developed by Meta AI, made its debut in February 2023 [8]. Renowned for its proficiency in comprehending and generating human language across multiple tasks, including translation, summarization, and creative writing, Llama leverages advanced techniques such as mixed precision training and checkpoint ensembling. These methodologies aim to yield precise, fluent, and adaptable outputs, thereby opening avenues for groundbreaking applications in natural language processing (NLP) across research, education, and beyond. The transformer architecture of Llama family is depicted in Figure 3. The LLaMA family encompasses a spectrum of foundation language models, ranging from 7B to 65B parameters, all based on the transformer architecture with several enhancements. Key deviations from the original architectures include [35]:

- The RMSNorm normalizing function is implemented to enhance training stability by normalizing the input of each transformer sub-layer, rather than normalizing the output.
- The traditional ReLU non-linearity is substituted with the SwiGLU (Swish-Gated Linear Unit) activation function to enhance overall performance.
- Absolute positional embeddings are eliminated, and rotary positional embeddings (RoPE) are introduced at every layer of the network to improve positional encoding efficiency and effectiveness.

LLaMA Models: These models are large language models (LLMs) designed for efficient performance and accessibility.

Normalization (Norm): In machine learning, normalization helps stabilize training and improve model performance by scaling data into a common range. There are different techniques:

- LayerNorm: Normalization within individual layers of a model.
- RMSNorm: Root Mean Square normalization, a specialized variant.

MLP: Multilayer Perceptron, a basic type of neural network often found in language models.

Attention: A crucial mechanism in LLaM architectures that focuses the model on the most important aspects of the input.

Llama 2 presents several advantages over the original LLaMa models. Firstly, Llama 2 models feature an extended context length of 4,096 tokens, twice that of LLaMa 1, enhancing their capacity to retain and comprehend larger text volumes during inference. This leads to more coherent and fluent natural language interactions. Additionally, while LLaMa 1 was limited to research applications, Llama 2 is accessible to any organization with fewer than 700 million active users, broadening its accessibility and utility. Moreover, Llama 2 underwent more rigorous training, being pre-trained on 40% more data to enrich its knowledge base and contextual comprehension. Notably, Llama 2 chat models were fine-tuned using reinforcement learning from human feedback (RLHF), deviating from LLaMa 1, resulting in responses better aligned with human expectations [36].

While Llama2 LLM offers several advancements, it's crucial to recognize potential drawbacks. Firstly, akin to any large language model, there's a risk of biased or inaccurate content generation, highlighting the importance of thorough scrutiny and human oversight to ensure precise information dissemination. Secondly, despite its capabilities, Llama2 LLM cannot substitute specialized medical expertise and should instead complement healthcare professionals. Additionally, concerns arise regarding data privacy, especially when employing Llama2 LLM for sensitive patient information, underscoring the necessity for stringent privacy protocols to avert breaches and unauthorized access. Notwithstanding these challenges, with proper management and oversight, the advantages of Llama2 LLM can still be effectively leveraged.

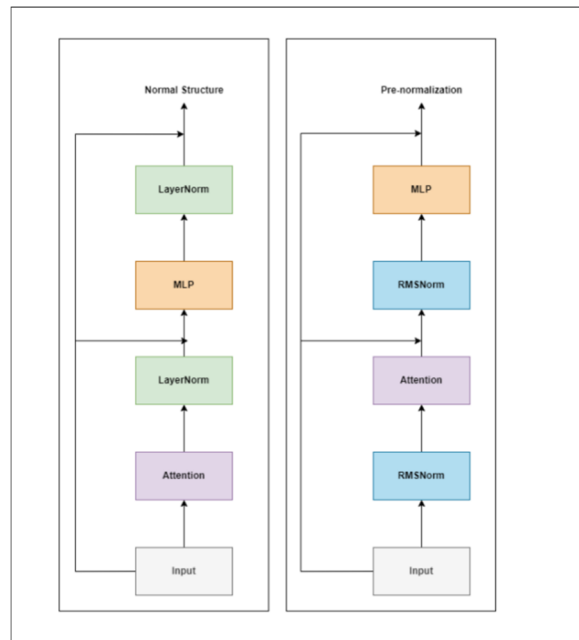


Figure 3. LLaMA's Transformer Block architecture [35].

3.5. Orca

Orca LLM, developed by Microsoft, aims to overcome the constraints of ChatGPT by introducing a logic-based framework that emulates human-like reasoning [28]. The name "Orca" stands for "Logical and Linguistic Model," underscoring its dual emphasis on both logical reasoning and linguistic comprehension. Diverging from conventional language models that primarily rely on statistical patterns in data, Orca incorporates logical reasoning mechanisms to augment its output.

Orca 2, an extension of the LLaMA 2 model family [37]. Orca 2, like other large language models, confronts inherent challenges and potential drawbacks in its deployment and utilization. One notable concern is the perpetuation of data biases, which can inadvertently influence the model's outputs and contribute to biased or unfair content. Additionally, despite its remarkable language capabilities, Orca 2 may struggle with contextual understanding, resulting in inaccuracies or nonsensical responses. The model's inherent complexity also poses challenges regarding transparency, as it operates as a "black box," making it difficult to interpret the rationale behind its outputs. Moreover, large language models like Orca 2 carry the risk of causing various content harms, necessitating vigilance and proactive measures to mitigate potential negative impacts. The phenomenon of hallucination further complicates matters, highlighting the need for caution in relying on model-generated content for critical decisions or information.

On the other hand, Orca's training methodology brings several significant advantages to the table. By addressing the capacity gap challenge through the use of an intermediate teacher model, Orca can effectively learn from a more proficient source, leading to improved performance, particularly for smaller student models. The model's progressive learning approach enables it to incrementally build upon its knowledge, starting from simpler examples and gradually incorporating more complex ones. This incremental learning process enhances Orca's capacity for reasoning and explanation generation, contributing to its overall effectiveness. Furthermore, Orca's ability to emulate the reasoning process of LLMs such as GPT-4 offers promising opportunities for enhanced performance across a diverse range of tasks. Leveraging the insights gleaned from GPT-4's explanation traces and step-by-step thought processes, Orca can further refine its capabilities and deliver more robust outcomes.

3.6. Google Gemini

Google's Gemini family comprises highly proficient multimodal models, jointly trained across image, audio, video, and text data to establish a model with robust generalist capabilities across modalities, coupled with advanced understanding and reasoning proficiency in each domain [27]. Gemini 1.0 is available in three sizes: Ultra, Pro, and Nano, each meticulously crafted to cater to diverse computational constraints and application needs. This advancement significantly pushes the boundaries in large-scale language modeling, image comprehension, audio processing, and video understanding, building upon the foundations of sequence models, neural network-based deep learning, and distributed machine learning

systems enabling extensive training.

Gemini models are built upon Transformer decoders [25], which have been enhanced with architectural improvements and model optimization to facilitate stable training at scale and optimized inference on Google’s Tensor Processing Units. These models are trained to support a context length of 32 k and employ efficient attention mechanisms, such as multi-query attention [38]. They are specifically trained to handle textual input interspersed with a wide array of audio and visual inputs, including natural images, charts, screenshots, PDFs, and videos, and have the capability to generate both text and image outputs. The visual encoding in Gemini models draws inspiration from Google’s prior work on Flamingo [39], with the significant distinction that these models are multimodal from inception and can directly output images using discrete image tokens [40]. Video understanding is achieved by encoding the video as a sequence of frames within the large context window, allowing for the seamless interleaving of video frames or images with text or audio as part of the model input. Moreover, Gemini models can accommodate variable input resolutions to allocate more computational resources to tasks requiring detailed comprehension. Additionally, they have the capability to directly process audio signals at 16 kHz from the Universal Speech Model (USM) [41] features, enabling the capture of nuances that may be lost when audio is naively mapped to a text input.

Gemini is a multi-modal LLM that accepts text, audio, video and images as it’s input [42], Figure 4 describes a high-level architecture of the model. These inputs are tokenized in the token convertor and passed on to the transformer architecture model. The output of the transformers is then passed on to their respective decoders which generates the successive answers. As of now the model is capable of generating text and image outputs.

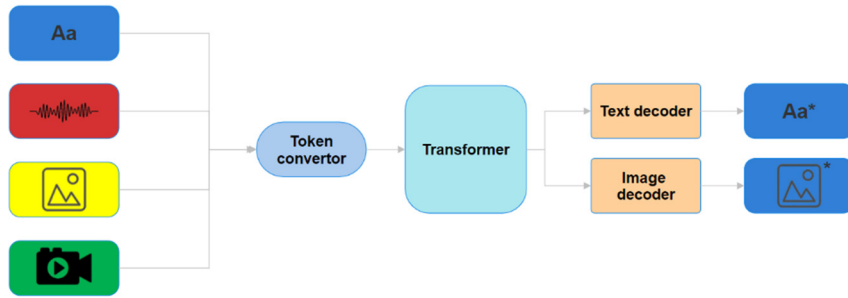


Figure 4. Model architecture of google gemini.

The capabilities of the system encompass various functionalities, including seeking information by amalgamating global knowledge with data extracted from images and videos, addressing queries related to detailed object identification in visual media, comprehending digital content by extracting information from a wide array of sources such as infographics, charts, figures, tables, and web pages, generating structured content in formats like HTML and JSON based on provided instructions, crafting descriptions for images and videos with varying levels of detail, and offering extrapolations by suggesting additional insights based on location, predicting potential events before, after, or between visual content, and enabling imaginative applications such as crafting narratives inspired by visual inputs.

Gemini exhibits various capabilities and limitations across different tasks. In terms of spatial reasoning, it may face challenges in accurately localizing objects or text within images and may demonstrate reduced accuracy with rotated images. Additionally, its counting capability is limited to providing rough estimates, especially for obscured objects. While capable of handling longer videos as a distinct modality, Gemini processes data from non-contiguous image frames and does not analyze information beyond the initial 2 minutes of the video. To improve performance with dense video content, shortening the video length is recommended. The model may encounter difficulties with tasks requiring multiple reasoning steps and may occasionally produce hallucinations by extrapolating beyond actual content or generating inaccurate output. It's not suitable for medical image interpretation or providing medical advice and is not trained for multi-turn chatbot functionality.

4. Proposed Methods

The dataset originates from a file that users upload to the model, which then utilizes this dataset to produce the intended output. In this particular investigation, the file utilized is the PDF version of the book “Linux Server Administration” by Wale Soyinka [43]. This book encompasses a total of 665 pages and is segmented into 30 chapters. The entirety of the embedding procedure for each chapter requires approximately 2 minutes across all models, with the exception of Google Gemini. Google Gemini

completes the embedding process for a specific chapter in approximately 30 seconds. The topics for which questions were generated stem from a computer science course at the university level.

This text describes a system that utilizes a knowledge base to enhance a large language model's (LLM) capability for generating subjective questions.

The process begins with a user uploading a document through a dedicated web interface. The system then extracts and processes the text content. This involves breaking the text down into smaller units called tokens (words, sentences, or phrases) and converting each token into a numerical vector using an embedding model. These vectors capture the meaning and relationships between words, allowing the system to work efficiently with numerical data. Finally, these vectors are stored in a Vector Database, forming the knowledge base. The overall process is shown Figure 5a. This process is a one-time process and the knowledge base formed, is used to extend the LLM's context window.

When a user wants to generate subjective questions, they submit a query through the interface. This query is also converted into a numerical vector using the same embedding model. The system then performs a semantic search on the knowledge base, comparing the query vector to the stored information. This search retrieves relevant documents or units (sections, paragraphs) that are semantically similar to the user's query.

The retrieved information, along with the original query, is then combined to form a "prompt" that provides richer context for the LLM. Finally, the LLM utilizes this enhanced context to generate subjective questions that are both relevant to the user's query and informed by the knowledge base's content.

In essence, this system leverages the knowledge base to provide the LLM with broader context, enabling it to generate more meaningful and insightful subjective questions. This process is depicted in Figure 5b.

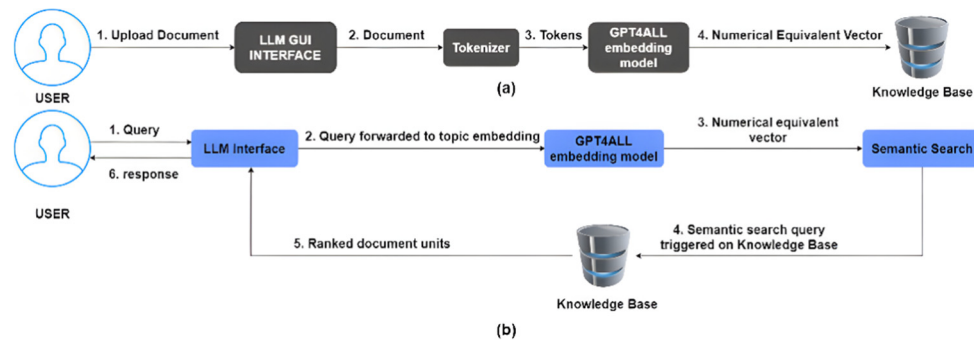


Figure 5. (a) Knowledge Base uploading process. (b) Flowgraph of Proposed Model [44].

5. Results

Figure 6 shows two graphs, "a" and "b". One named model rating and the other named model relevancy. The ratings were documented on a scale ranging from 0 to 5, with 5 denoting the highest rating and 0 indicating the lowest.

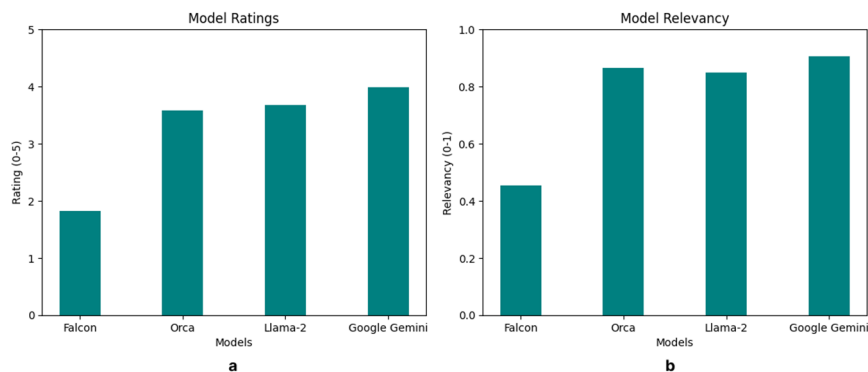


Figure 6. Results of all the four LLMs; (a) shows the ratings of all four models in a graphical representation. (b) shows the relevancy of all four models in graphical representation.

Table 1 presents a quantitative comparison of four different models, in the context of generating subjective questions. It compares two key metrics:

Table 1. Quantitative comparison of the models.

Models	Average Ratings	Average Relevancy
Falcon	1.82	0.45
Orca	3.58	0.87
Llama-2	3.68	0.85
Google Gemini	3.92	0.90

Average Ratings: This column represents the average score users assigned to the questions generated by each model. A higher score indicates that users generally found the questions to be better quality.

Average Relevancy: This column shows the average level of relevance between the generated questions and the user's query. A higher score indicates that the questions were more pertinent to the user's intended topic.

6. Result Discussion

To assess the models, automatic evaluation metrics are insufficient, prompting a manual human review of each question to ascertain the outcomes. The ratings and relevance were assessed by a Linux expert who meticulously reviewed each question generated by the model. Throughout this research, the expert meticulously examined approximately 1,400 questions to assign ratings and evaluate their relevance. The results were derived based on the following criteria for rating and relevance expressions.

$$Y_{rating} = \frac{\sum_{i=1}^n A_i}{n}$$

$$Y_{relevancy} = \frac{\sum_{i=1}^n B_i}{n}$$

Y_{rating} represents the mean rating, while $Y_{relevancy}$ represents the mean relevance factor, with A_i denoting the rating and B_i representing the average relevancy of question i , respectively, and n indicating the total number of generated questions. Due to the inadequacy of automatic evaluation metrics, specific assessment metrics were employed to evaluate the proposed model. All models within the proposed systems were evaluated and run on an MSI laptop equipped with an Intel i5-11400H processor. Falcon, Llama2, Orca, and Google gemini all four were executed on the CPU, out of which Falcon, Llama2, and Orca resulted in an average response time of 150.95 seconds, while Google Gemini exhibited an average response time of 12.5 seconds.

It can be observed that falcon achieved the lowest average rating (1.82) and relevancy score (0.45), suggesting users found its questions to be of lower quality and less relevant to their queries. Orca Performed moderately better than Falcon, with a higher average rating (3.58) and relevancy score (0.87). Llama2 showed even better performance, with the second-highest average rating (3.68) and a slightly lower relevancy score (0.85) compared to Google Gemini. Google Gemini Achieved the highest average rating (3.92) and the highest relevancy score (0.90), indicating that users judged its questions to be of the best quality and most relevant to their queries amongst the compared models.

7. Conclusions

This study introduces a system leveraging Large Language Models (LLMs) to generate Subjective Questions aimed at enhancing academic performance for both students and teachers. The proposed model utilizes various LLMs, with a comparison of these models provided in Table 1.

Table 2 presents a comparative analysis of research articles focusing on Large Language Models (LLMs) and their contributions within the field. Each model is paired with a dataset utilized for fine-tuning or expanding the context of the LLM. Notably, these systems undergo evaluation through human review, as automatic evaluation methods are deemed unsuitable. This research was undertaken at no cost, signifying that no financial resources were allocated to its execution. The primary objective of this study was to evaluate and compare the performance of leading LLMs accessible online, which are freely available to the public.

Table 2. Comparison of the proposed systems with recent models [44].

LLM	Data	Proposed System	Results
[1] model	EQG-RACE	Fine-tuned GPT-3 on macaw I I-b and EEQG	Manual human evaluations were done which suggests that the model generates high quality questions.
[6] gpt-3.5	MultispanQA, QUOREF, DROP	Developed a prompting system named FBPrompt.	Improved the performance of the LLMs
[29] ChatGPT	None	Personalized adaptive practicing system using ChatGPT	Correct feedback by ChatGPT 99%, rest 1% human intervention needed.
[30] GPT-3.5, GPT-4, and Google Bard	149-question Self-Assessment Neurosurgery Exam (SANS) Indications Exam was used.	Performance analysis of LLMs	correct feedback percentage GPT-3.5 = 62.4%, GPT4 = 82.6%, Google Bard = 44.2%
[44] Falcon, LLAMA2, Orca	University Textbook	LLM with extended context using custom knowledge based	Manual Human evaluation was performed on the generated questions and received average rating of 3.03.

8. Future Discussion

In the discussion, we explore the implications and future directions arising from the comparative analysis of four prominent LLM models within the domain of subjective question bank generation systems.

Firstly, examine more deeply the fine-tuning and optimization techniques employed for each LLM model to enhance their performance in generating subjective questions. Investigate how different fine-tuning strategies impact the quality and diversity of the generated questions. Additionally, assess the scalability and efficiency of the LLM models, particularly in handling larger datasets and generating questions in real-time. Identify potential strategies for improving efficiency without compromising quality.

Secondly, evaluate the robustness and generalization capabilities of the LLM models across various domains and datasets. Explore how well each model adapts to different contexts and whether there are any biases or limitations that need to be addressed. Consider incorporating human evaluation and feedback mechanisms to assess the quality and relevance of the generated subjective questions.

Thirdly, discuss the potential integration of the LLM-based question generation system with existing educational platforms or learning management systems. Explore how such integration can streamline the process of creating personalized assessments and supporting adaptive learning experiences. Address the ethical and societal implications of deploying LLM models for subjective question generation, including concerns related to bias, fairness, and privacy.

Finally, outline potential future research directions and areas of exploration in the field of LLM-based question generation systems. This could include investigating novel architectures, exploring multimodal approaches, or advancing natural language understanding capabilities to further enhance the quality and diversity of generated questions. By addressing these points in the future discussion section, you can provide valuable insights and guidance for researchers, educators, and practitioners interested in leveraging LLM models for subjective question generation in educational contexts.

Author Contributions

A.S. and S.N. designed and performed the experiments, derived the models and analysed the data. A.S. wrote the manuscript in consultation with M.K. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflict of Interest Statement


The authors declare no conflicts of interest.

Data Availability Statement

No new data were created or analyzed in this study. Data sharing is not applicable to this article as no new data were created.

References

1. Dijkstra, R., Genç, Z., Kayal, S. and Kamps, J., 2022. Reading Comprehension Quiz Generation using Generative Pre-trained Transformers.
2. Ayanwale, M.A., Sanusi, I.T., Adelana, O.P., Aruleba, K.D. and Oyelere, S.S., 2022. Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence*, 3, p.100099.
3. Kunichika, H., Katayama, T., Hirashima, T. and Takeuchi, A., 2004. Automated question generation methods for intelligent English learning systems and its evaluation. In *Proc. of ICCE (Vol. 670)*.
4. Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J. and Poon, H., 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4).
5. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102274.
6. Huang, Z., Zhou, J., Xiao, G. and Cheng, G., 2023. Enhancing In-Context Learning with Answer Feedback for Multi-Span Question Answering. *arXiv preprint arXiv:2306.04508*.
7. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U. and Oprea, A., 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633-2650).
8. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A., LLaMA: open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.
9. Moore, S., Nguyen, H.A., Bier, N., Domadia, T. and Stamper, J., 2022, September. Assessing the quality of student-generated short answer questions using GPT-3. In *European conference on technology enhanced learning* (pp. 243-257). Cham: Springer International Publishing.
10. Liu, X. and Liu, H., 2021. Design and Application of English Grammar Intelligent Question Bank System. *Scientific Programming*, 2021, pp.1-10. Vancouver
11. Kurdi, G., Leo, J., Parsia, B., Sattler, U. and Al-Emari, S., 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30, pp.121-204.
12. Hao, H., Zhou, L., Liu, S., Li, J., Hu, S., Wang, R. and Wei, F., 2023. Boosting large language model for speech synthesis: An empirical study. *arXiv preprint arXiv:2401.00246*.
13. Zhong, V., Xiong, C. and Socher, R., 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.
14. Hussain, M., Zhu, W., Zhang, W. and Abidi, S.M.R., 2018. Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational intelligence and neuroscience*, 2018.
15. Andersson, C. and Palm, T., 2017. The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and instruction*, 49, pp.92-102.
16. Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A. and Chartash, D., 2023. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1), p.e45312.
17. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F. and Ting, D.S.W., 2023. Large language models in medicine. *Nature medicine*, pp.1-11.
18. Sallam, M., 2023. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv*, pp.2023-02.
19. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
20. Dao, X.Q., 2023. Performance comparison of large language models on vnhsge english dataset: Openai chatgpt, microsoft bing chat, and google bard. *arXiv preprint arXiv:2307.02288*.
21. Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y. and Ye, W., 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*
22. Anand, Y., Nussbaum, Z., Treat, A., Miller, A., Guo, R., Schmidt, B., ... & Mulyar, A. (2023). GPT4All: An Ecosystem of Open Source Compressed Language Models. *arXiv preprint arXiv:2311.04931*.

23. Leal Castillo, E. (2023). Investigating Open Source Transformer Techniques for Question Answering Systems on Cloud Domain: A Comparison with ChatGPT3. 5-Turbo (Master's thesis).
24. Anand, Y., Nussbaum, Z., Treat, A., Miller, A., Guo, R., Schmidt, B., Community, G., Duderstadt, B. and Mulyar, A., 2023. GPT4All: An Ecosystem of Open Source Compressed Language Models. arXiv preprint arXiv:2311.04931.
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
26. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
27. Lee, G.G., Latif, E., Shi, L. and Zhai, X., 2023. Gemini pro defeated by gpt-4v: Evidence from education. arXiv preprint arXiv:2401.08660.
28. Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., & Awadallah, A. (2023). Orca: Progressive learning from complex explanation traces of gpt-4. arXiv preprint arXiv:2306.02707.
29. Kabir, M.R. and Lin, F., 2023. An LLM-Powered Adaptive Practicing System.
30. Ali, R., Tang, O.Y., Connolly, I.D., Fridley, J.S., Shin, J.H., Sullivan, P.L.Z., Cielo, D., Oyelese, A.A., Doberstein, C.E., Telfeian, A.E. and Gokaslan, Z.L., 2022. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*, pp.10-1227.
31. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., ... & Silvestri, F. (2024). The power of noise: Redefining retrieval for rag systems. arXiv preprint arXiv:2401.14887.
32. Retrieval augmented generation (no date) Databricks. Available at: <https://www.databricks.com/glossary/retrieval-augmented-generation-rag>.
33. Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q. and Mazzotta, D., 2023. The falcon series of open language models. arXiv preprint arXiv:2311.16867.
34. Wolfe, C. R. (2023) Falcon: The pinnacle of open-source LLMs, Towards Data Science. Available at: <https://towardsdatascience.com/falcon-the-pinnacle-of-open-source-llms-600de69c333c> (Accessed: March 1, 2024).
35. (Chingis), C. (2023) LLMs explained: LLaMA and its architecture (part 1), Medium. Available at: <https://chingisoinar.medium.com/llms-explained-llama-and-its-architecture-part-1-716627e7754c> (Accessed: March 1, 2024).
36. Pokhrel, K. (2023) Analysis on the power of Llama2 LLM: Pros, cons, and best practices in health & medicine, Medium. Available at: <https://medium.com/@kushalpokhrel/unveiling-the-power-of-llama2-llm-pros-cons-and-best-practices-in-health-medicine-d849a39953ba> (Accessed: March 1, 2024).
37. Singh, P. (2024) Unlocking the power of Orca LLM, Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2024/01/orca-llm-simulating-the-reasoning-processes-of-chatgpt/> (Accessed: March 1, 2024).
38. Shazeer, N., 2019. Fast transformer decoding: One write-head is all you need. arXiv preprint arXiv:1911.02150.
39. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyler, L. and Kolesnikov, A., 2022. Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794.
40. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I., 2021, July. Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821-8831). PMLR.
41. Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G. and Meng, Z., 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. arXiv preprint arXiv:2303.01037.
42. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A. and Millican, K., 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.
43. Soyinka, W. (2020) *Linux administration: A beginner's guide*. New York: McGraw-Hill Education.
44. Sayed A, Kanojia M, Nabajja S (2023) Subjective question bank generation using large language models with custom knowledge base. In: *Bio-Inspired Computing and Applications*. Springer
45. Medina, J.R. and Kalita, J., 2018, December. Parallel attention mechanisms in neural machine translation. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 547-552). IEEE.
46. Strengths and limitations (no date) Google Cloud. Available at: <https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/strengths-limits> (Accessed: March 1, 2024).
47. Recursively split by character  Langchain. Available at: https://python.langchain.com/docs/modules/data_connection/document_transformers/text_splitters/recursive_text_splitter (Accessed: 11 October 2023).
48. Anonymous4chan/LLAMA-2-70B hugging face anonymous4chan/llama-2-70b Hugging Face. Available at: <https://huggingface.co/anonymous4chan/llama-2-70b>.