Generating Question Using Retrieval Augmented Generation with Large Language Model: Comparative Study

Nur Aisyah
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6025241004@student.its.ac.id

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@its.ac.id

Abdullah Faqih Septiyanto
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025221022@student.its.ac.id

Agus Tri Haryono
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025231020@student.its.ac.id

Dwi Sunaryono
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
dwi@its.ac.id

2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C) | 979-8-3315-0857-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/BTS-I2C63534.2024.10941961

***Abstract*—Rapid technological developments affect all aspects of life, especially in the field of education. One application of technology in the field of education is the automatic creation of questions. Automatic creation of questions is proposed by the author using the Retrieval Augmented Generation (RAG) method with Large Language Model (LLM). The aim is to help teachers in creating automatic exam questions by generating questions in the question bank or question database with the difficulty level of the questions. The author uses 3 test scenarios, namely ideal conditions, non-ideal conditions and slightly ideal conditions using 3 models, such as Llama 3, Cohere LLM and GPT 2. The author's goal in using 3 models is to compare models seen from the results of their accuracy. The results of the experiments conducted are that in ideal conditions, Llama-3-8B and GPT 2 get a very good accuracy score of 100%, while Cohere LLM gets an accuracy score of 80%. In non-ideal conditions, Cohere LLM gets the best score among other models. Cohere LLM gets an accuracy score of 80%, while Llama-3-8B and GPT 2 get an accuracy score of 67%. In slightly ideal conditions, Llama-3-8B gets the highest accuracy score of 100%. Cohere LLM and GPT 2 get a poor score of 67%. From these results, it can be concluded that Llama-3-8B is the best model compared to the other 2 models in generating questions. Then followed by GPT 2 and Cohere LLM**

Keywords—RAG, LLM, generate, question, difficulty-level

I. INTRODUCTION

The rapid development of technology brings changes to all aspects of life. One aspect that is affected by this technological development is in the field of education. In education, technology can help facilitate school devices in developing planned programs, one of which is in the teaching and learning process. In the teaching and learning process in the classroom, the use of technology can help in understanding the concepts, theories and case studies that exist in learning [1]. One example is helping teachers in making exam questions. In general, teachers make exam questions every time an exam is held, for example, mid-semester exams and final semester exams with a manual system. This takes a very long time to make questions and sometimes the questions made by teachers do not consider a good level of difficulty.

The prior literature literally discuss about question answering, question reasoning, automated evaluation and chatbot. The example literature discuss about question reasoning by [2], the title is "BioRAG: A RAG-LLM Framework for Biological Question Reasoning". On this paper, the framework has ability to get relevant and current in formation form with a blend of traditional databases, toolkits and modern search engine. The result is a BIORAG can handle in complex biological queries. However, there are several papers that discuss about generating question. [3] the title is "Advanced Subjective Question Bank Generation Using Retrieval Augmented Generation Architecture" that discuss about how to generate an effective and informative question. The authors propose a novel approach that leverages Large Language Model (LLM) in conjunction with a custom knowledge base constructed from a specific text corpus. The aim of the method is to enhance the quality and relevance of the generated question by avoiding reliance on predefined templates, which can limit diversity. The result is the proposed method that authors' proposed show more quality and more relevance than existing method. The higher score of the method is Google Gemini Large Language Model (LLM).

Based on the problem, author proposes the Retrieval Augmented Generation (RAG) method using Large Language Model (LLM) which focuses on the model's ability to retrieve the right information according to the requested command. Our metric focus on factual accuracy, for example the ability to generating information/question matches with query. In addition, author uses 3 types of Large Language Model (LLM) like Llama-3-8B, Cohere LLM and GPT 2. Besides that, author uses 3 testing scenario for each model. The testing scenario is an ideal condition, a non-ideal condition and a slightly ideal condition.

In this paper, author want to implement the Retrieval Augmented Generation (RAG) method with Large Language Model (LLM). The aim of this method is helping teacher to

make exam question automatically with difficulty level. The model can generate question based on base knowledge from existing file question in database. The questions have 3 difficulty level, "low, medium, high". So, we can summarize the author's contribution is creating dataset of questions with difficulty level (low, medium, high), generating questions using Retrieval Augmented Generation (RAG) with Large Language Model (LLM) based on difficulty level (low, medium, high) and comparing the most effective Large Language Model (LLM) in generating questions. Based on the author's contribution, we can conclude that the expected result can show the best effective model to generate the questions.

II. Literature Review

A. RAG (Retrieval Augmented Generation)

Retrieval Augmented Generation (RAG) emerged as a new paradigm using LLM and basic knowledge [4]. Retrieval Augmented Generation (RAG) is divided into 2 core components of NLP, namely Information Retrieval (IR) and Natural Language Generation (NLG). The framework of RAG is to combine retrieval methods with generative models on a large scale so that it can produce relevant and factual responses contextually and accurately. RAG extracts relevant information from a large corpus and takes information from the generation process so that it can improve factual results [5].

Retrieval Augmented Generation (RAG) is a model that integrates Large Language Model (LLM) in optimizing the output of given commands by taking information from trained data [6]. This model works by finding the proximity of the output to the training data combined with LLM [6].

B. LLM (Large Language Model)

Large Language Model (LLM) is a pre-trained model aimed at solving text or language problems (Vasmani, Shazeer, Parmar et al., 2017 in [7]). This model will produce output that is as similar as possible to the input data[7]. Based on [8], this LLM can handle tasks in natural language processing, including text generation, summarization and question answering. This LLM is trained on huge amounts of textual data and has the ability to understand and generate text.

Based on the paper written by [9], lately, LLM has open-source models such as Llama, Falcon, Mistral and GPT-J. LLM has innovative characteristics compared to language models that use traditional transformers. LLM has zero-shot and few-shot learning capabilities, meaning LLM can have extraordinarily deep capabilities with limited training and desired tasks. In this paper, the author use 3 model of LLM such as Llama-3-8B, Cohere LLM and GPT 2.

C. Transformer

Transformers are designed to handle sequential data inputs such as natural language and perform tasks such as text summarization and translation. Transformers are facilitated with attention mechanisms so that the model focuses on the input segments that match the desired output. In addition, transformers process all inputs simultaneously [3].

Transformers consist of 2 components, namely encoder and decoder. The encoder has a function to change words into numbers/numeric. While the decoder has the opposite function to the encoder. The decoder has a function to change the numbers that have been translated by the encoder into the possible presence of words on the output [3].

D. Llama-3-8B

The Llama-3-8B model is a model that has been previously trained and can be adjusted to the instructions given. This model uses a transformer to optimize the performance of this model. The Llama-3-8B model is usually used in chats such as chatbots and can also be used to complete tasks in creating natural language [10].

E. Cohere LLM

Cohere LLM is a model that has been trained using business languages and supports 100 languages. Rerank and embed system that helps provide reliable answers and encourages the use of RAG. RAG is used to connect and retrieve document information to produce sophisticated AI products. Thus, this Cohere LLM is considered the most effective and accurate in overcoming language problems in business [11].

F. GPT 2

GPT 2 is a type of LLM that has been trained without human assistance, meaning in terms of labeling. The data trained on GPT 2 is raw data that will be processed automatically so that input and labels will be produced from the trained text. This model is trained and will then extract features in completing its tasks. The output of this GPT 2 is that it can produce text based on the commands given [12].

G. Evaluation

Evaluation is a measurement activity to determine the value of a job/product produced. In this study, the author uses 3 types of evaluation, namely precision, recall and f1-score [13].

Precision is the accuracy of information based on basic information with the answer information provided by the system [14]. Precision is dividing the true positive class by the total positive class [15]. The precision formula is shown in (1).

$$Precision = \frac{TP}{(TP + FP)} \quad (1)$$

Recall is a ranking of the system's success in finding information [14]. Recall is an evaluation that divides true positives into total true positives and false negatives [15]. The recall formula is shown in (2).

$$Recall = \frac{TP}{(TP + FN)} \quad (2)$$

F1-score is a calculation that shows the comparison between precision and recall based on true positive, false positive, true negative and false negative [16]. The F1-score formula is shown in (3).

$$F1 - Score = 2 \text{ x} \left(\frac{Precision \ x \ Recall}{Precision + Recall}\right) \quad (3)$$

Accuracy is the result of dividing true positives by total true positives, false positives and true positives. The accuracy formula is shown in (4).

$$Accuracy = \frac{TP}{(TP + FP + FN)} \quad (4)$$

Description:

- True Positive (TP) = generated text is the same as reference text
- False Positive (FP) = generated text is more than reference text
- False Negative (FN) = generated text is less than reference text

III. PROPOSED METHOD

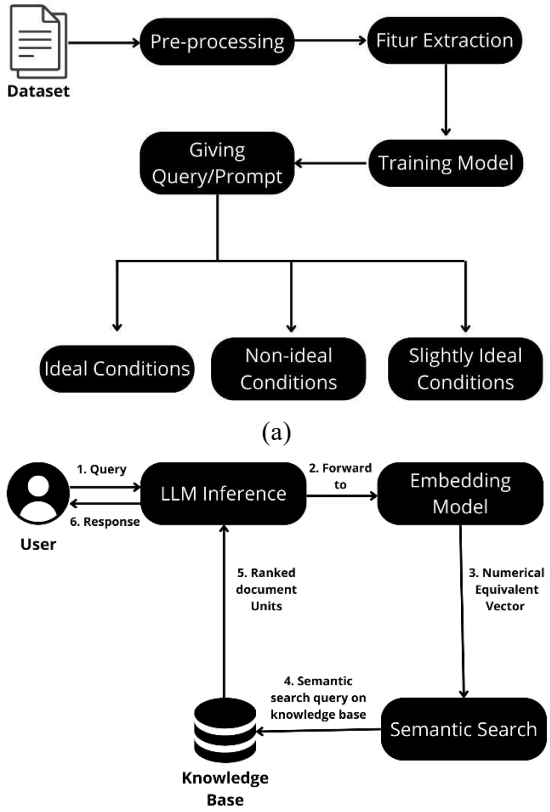

Fig. 1. Proposed method (a), Process of generating questions (b).

In this section, author will shows the proposed method that author creates. Author uses 5 step to do research as seen in Fig. 1.(a), such as dataset, pre-processing, feature extraction, training model and giving query or prompt.

A. Dataset

Based on Fig. 1.(a), the first step in this research is dataset, it means that, author must to prepare the dataset. The dataset that the author uses is a dataset of questions on simulation and digital communication subjects for class X vocational high schools. Dataset was carried out by collecting data from various sources. The questions are taken from several sources, namely from the question bank of SMK PGRI 1 Bangkalan, the Curriculum Communication Network in Bangkalan, the Simulation and Digital Communication School Examination (USP) of Bangkalan Regency, SMKN 1 Sorong Regency, LPMI Rokania, SMKN 1 Labang, Bangkalan, the School Examination (USP) of SMA/SMK Pontianak. From these questions, the author divided them into 15 question packages with a scheme of each question package consisting of 20 questions. So, the total amount of dataset is 300 data.

B. Pre-processing

After author get some of questions, based on Fig. 1.(a), author do pre-processing. Pre-processing is carried out by providing a label for the difficulty level of each question. The labelling process use manually method. The label of difficulty level is divided to be 3 level, such as low, medium and high. Output from pre-processing step is dataset which contain question with difficulty level.

C. Feature Extraction

After that, in the Fig. 1.(a), the dataset will be tokenized. The tokenizer process aims to divide sentences into several groups of small units so that they can be recognized by the system. After the tokenizer is done, the next step is the embedding process. Embedding is a process of changing letters/words into numbers so that they can be recognized by the computer. This embedding process is based on the tokenizer that has been done. This means changing letters to numbers through groups of small units resulting from the tokenizer. After that, the results of the number vector from embedding are stored as basic knowledge.

D. Training Model

Training model is important part of this research. Training model is done to train the model to understand the contents and patterns of the inputted dataset. In this research, there are 3 models used, namely Llama-3-8B, Cohere LLM and GPT 2. The flow of model training is that the user inputs a dataset, then the model will identify, read and process the content and patterns of the dataset.

E. Giving Query/Prompt

The last step in the Fig. 1.(a) is giving query/prompt. Giving query or can be said as generating question starts from the user writing a query or prompt as seen in Fig. 1.(b). Then the query that has been written by the user is received by LLM Inference and LLM Inference identifies, reads and processes the query. Then LLM Inference will forward the query to the embedding model to be understood by the computer. After embedding is done, it enters the semantic search section where it will understand the meaning of the embedding results. Then semantic search will look for the best, suitable and appropriate response in the knowledge base. Then the knowledge base will create a ranking order of the best answers and approach the requested query and send it to LLM Inference. Then, based on Fig. 1.(b), LLM Inference will provide a response to the user according to the requested query.

From the given query or prompts, the trained models produce output. From the three models, the models can display different outputs based on how well the models learn and store the trained question data. The way the model generates new models is by taking questions with criteria based on the word length which is generally taken in each

question category. The following are the criteria for sentence length.

1. For easy questions, questions that have a word length of less than 150 words.
2. For medium questions, questions that have a word length between 150 and 200 words.
3. For difficult questions, questions that have a word length of more than 200 words.

In addition, the author proposes to conduct 3 test scenarios for each model. Below is an explanation of the 3 test scenarios.

1. Ideal conditions, namely conditions where the question bank matches the given prompt.
2. Non-ideal conditions, namely conditions where the question bank does not match the given prompt.
3. Slightly ideal conditions, namely conditions where the question bank matches but slightly with the given prompt.

In this proposed method, there are several differences compared to current technology. This proposed method focuses on one subject, namely Computer and Digital Systems (Siskomdig) for grade X SMK, so that the questions and answers produced are relevant to the context of the subject.

IV. Results and Discussion

In this section, the author presents the experimental results that have been carried out. This experiment was carried out on 3 LLM, namely Llama-3-8B, Cohere LLM and GPT 2. In addition, in this experiment, the author used 3 scenarios for each model, namely ideal conditions, non-ideal conditions and slightly ideal conditions.

A. Result

Ideal conditions is a conditions where the question bank matches the given prompt or query. For example, can be seen at the table I., the prompt asks for 10 questions with the criteria of 3 low questions, 4 medium questions and 3 high questions, then the model will provide output that matches the prompt. Table IV. will present the results of the ideal test scenario.

Non-ideal conditions is a conditions where the question bank does not match the given prompt or query. For example, can be seen at the table II., the prompt asks for 10 questions with the criteria of 3 low questions, 2 medium questions and 5 high questions, but the question bank doesn't have the high question. In this situation, author wants to know how the models solve this problem. Table V. will present the results of the non-ideal test scenario.

Slightly Ideal conditions is a conditions where the question bank matches but slightly with the given prompt or query. For example, can be seen at the table III., the prompt asks for 10 questions with the criteria of 0 low questions, 5 medium questions and 5 high questions, but the question bank have only 3 high question. From this case, the author wants to evaluate how well the model works in solving this problem. Table VI. will present the results of the slightly ideal test scenario.

B. Discussion

In ideal conditions, it can be seen at the table I and IV. that the Llama-3-8B and GPT 2 models show good results with an accuracy of 100%. This means that the Llama 3-8B and GPT 2 models can generate questions well, so that both models can meet the demands of the prompt. On the other hand, the Cohere LLM model shows poor results with an accuracy of 60%. It can be concluded that the Cohere LLM model cannot generate questions well, so the Cohere LLM model cannot meet the demands of the prompt.

While in non-ideal conditions, it can be seen at the table II and V. that Cohere LLM shows a fairly good accuracy score of 80%, while for Llama-3-8B and GPT 2 shows a poor accuracy value of 67%. From these results, it can be concluded that the Cohere LLM model can provide a fairly good solution where the model can generate new questions with existing basic knowledge, so that it can provide output from requests from prompts even though they are not fully fulfilled. On the other hand, the Llama-3-8B and GPT 2 models cannot provide a solution to this case. Both models can only display questions available in the question bank.

The last, in slightly ideal conditions, it can be concluded from tables III and VI. that the Llama-3-8B model can meet the question request by taking questions from other question categories with an accuracy score of 100%. This is a solution even though it is not as expected. The author expects the model to be able to generate new questions from existing difficult question category references. For Cohere LLM and GPT 2 showed less than optimal results, namely 67%. This means that the Cohere LLM and GPT 2 models have not been able to generate new questions to complete the question request.

From a comparison of the 3 models with the 3 scenarios above, the author can summarize it at table VII. At table VII. shows accuracy from 3 models in 3 scenario testing. From the comparison of 3 models with 3 scenarios above, the author can summarize it in table VII. Table VII shows the accuracy of 3 models in testing 3 scenarios. The results show that Llama-3-8B is the best model for generating questions compared to other models. Llama-3-8B works by generating questions if one or all of the questions are available in reference. For Cohere LLM, it works by generating questions even though the questions are not available in reference. While for GPT 2, it works by generating questions that are available in the reference. Of these three models, to overcome the existing shortcomings, efforts are needed to increase the dataset and conduct repeated training, so that the model can be better at learning relevant question patterns. The inability of the model to generate varied questions is one of the main challenges faced by the author in this study.

TABLE I. Output Ideal Conditions Using Llama-3-8B

Output Ideal Conditions Using Llama-3-8B
Evaluasi Kesesuaian Jumlah Soal yang Diminta dan yang Tersedia: Jumlah soal kategori rendah yang diminta: 3, yang diambil: 3 Jumlah soal kategori sedang yang diminta: 4, yang diambil: 4 Jumlah soal kategori tinggi yang diminta: 3, yang diambil: 3 Jumlah soal sesuai dengan yang diminta. Evaluasi Kesesuaian Tingkat Kesulitan Soal: Kategori kesulitan soal yang diambil: [0 1 2] Tingkat kesulitan soal sesuai dengan yang diminta. === Evaluation Metrics ===

Output Ideal Conditions Using Llama-3-8B
True Positives (TP): 3 False Positives (FP): 0 False Negatives (FN): 0 Accuracy: 1.00 Precision: 1.00 Recall: 1.00 F1-Score: 1.00 Soal dan Kategori yang Dipilih: - Soal: Sistem perangkat lunak biasa disebut dengan a. Computer b. Hardware c. Laptop d. Software e. Mouse | Kategori: 0 - Soal: Melaksanakan perintah pada command line adalah fungsi dari a. Port b. Terminal c. File d. Folder e. OpenOffice.org | Kategori: 0 - Soal: Berikut ini langkah awal dalam menyusun program adalah a. Membeli buku b. Membuat algoritma c. Membeli komputer d. Menulis e. Membuat program | Kategori: 0 - Soal: Diketahui P=10; P=P+5; Q=P. Maka nilai P & Q adalah a. 15 & 0 b. 0 & 10 c. 0 & 15 d. 10 & 15 e. 15 & 15 | Kategori: 1 - Soal: Parameter pada perintah prompt yang digunakan untuk menghapus karakter sebelumnya dan fungsinya sama dengan backspace adalah a. $g b. $h c. $n d. $p e. $q | Kategori: 1 - Soal: Guna menyelesaikan setiap masalah, algoritma berkaitan dengan a. Pola pikir manusia b. Pengejaan huruf c. Huruf yang sulit d. Angka normal e. Pengurutan angka | Kategori: 1 - Soal: Langkah-langkah dalam memecahkan masalah adalah a. Algoritma-Looping-Eksekusi b. Model-Eksekusi-Input c. Output-Masalah-Algoritma d. Flowchart-Masalah-Hasil e. Algoritma-Program-Hasil | Kategori: 1 - Soal: Bertugas untuk menangkap objek pada layar monitor adalah tugas dari a. Take Screenswarm b. Kabel c. Port d. Copy e. Picture | Kategori: 2 - Soal: 1. Menyalakan sesuai prosedur, 2. Bukalah jendela command prompt, 3. Ketik perintah tasklist pada tampilan command prompt, 4. Mengamati hasil yang ditampilkan, urutan tersebut adalah langkah-langkah untuk a. Mengecek harddisk b. Menunjukkan pengguna bekerja direktori utama c. Mengecek aktivitas windows d. Menampilkan program apa saja yang sedang berjalan e. Menutup program secara paksa | Kategori: 2 - Soal: Dibawah ini yang termasuk fungsi command prompt adalah a. Memperbaiki bad sector pada harddisk b. Menghentikan aplikasi secara paksa c. Mengecek status aktivitas windows d. Mengetahui mac address komputer e. Tidak ada yang benar | Kategori: 2

TABLE II. OUTPUT NON-IDEAL CONDITIONS USING COHERE LLM

Output Non-ideal Conditions Using Cohere LLM
True Positives, False Positives, and False Negatives by Difficulty Level: Low: TP = 3, FP = 1, FN = 0 Medium: TP = 1, FP = 0, FN = 1 High: TP = 0, FP = 0, FN = 0 Evaluation Metrics: Precision: 0.85 Recall: 0.80 F1 Score: 0.78 Accuracy: 0.80 1. Low Difficulty: What is the correct order of steps to display the DOS on the screen according to the given instructions? A. 4, 2, 1, and 3 B. 3, 2, 1, and 4 C. 3, 2, 4, and 1 D. 4, 1, 2, and 3 E. 1, 4, 3, and 2 2. Low Difficulty: What is the purpose of running the tasklist command on the command prompt? A. To check the status of windows B. To close an application forcibly C. To check the hard drive D. To display all currently running programs E. To show the user's home directory 3. Medium Difficulty: What is the function of the command prompt that is not among the following?

Output Non-ideal Conditions Using Cohere LLM
A. Fixing bad sectors on the hard drive B. Forcibly stopping an application C. Checking the status of the operating system D. Finding the computer's MAC address E. None of the above 4. Low Difficulty: Question: What is the correct order of steps to display the DOS screen according to the given instructions? A. 3, 4, 2, and 1 B. 1, 2, 4, and 3 C. 2, 3, 1, and 4 D. 4, 1, 3, and 2 E. 1, 2, 4, and 3 5. Low Difficulty: Question: When activating the computer, which option should you choose to display the command prompt? A. Start B. Accessories C. Activate Computer D. Command Prompt E. Close

TABLE III. OUTPUT SLIGHTLY IDEAL CONDITIONS USING GPT 2

Output Slightly Ideal Conditions Using GPT 2
Evaluasi Kesesuaian Jumlah Soal yang Diminta dan yang Tersedia: Jumlah soal kategori rendah yang diminta: 0, yang diambil: 0 Jumlah soal kategori sedang yang diminta: 5, yang diambil: 5 Jumlah soal kategori tinggi yang diminta: 5, yang diambil: 3 Jumlah soal tidak sepenuhnya sesuai dengan yang diminta. Evaluasi Kesesuaian Tingkat Kesulitan Soal: Kategori kesulitan soal yang diambil: [1 2] Tingkat kesulitan soal tidak sesuai dengan yang diminta. === Evaluation Metrics === True Positives (TP): 2 False Positives (FP): 0 False Negatives (FN): 1 Accuracy: 0.67 Precision: 1.00 Recall: 0.67 F1-Score: 0.80 Soal dan Kategori yang Dipilih: - Soal: Berikut adalah tahap penyelesaian masalah yaitu a. Input-Proses-Output b. Masalah-Input-Output c. Model-Proses-Input d. Masalah-Program-Hasil e. Algoritma-Model-Eksekusi | Kategori: 1 - Soal: Asal kata dari logika adalah a. Logis b. Logic c. Logos d. Logika e. Intelligence | Kategori: 1 - Soal: Parameter pada perintah prompt yang digunakan untuk menghapus karakter sebelumnya dan fungsinya sama dengan backspace adalah a. $g b. $h c. $n d. $p e. $q | Kategori: 1 - Soal: Perhatikan perintah berikut: 1. Klik start, 2. Pilih accessories, 3. Aktifkan komputer, 4. Pilih command prompt, Urutan yang benar untuk menampilkan DOS di layar adalah a. 1, 2, 3 dan 4 b. 2, 1, 3 dan 4 c. 3, 1, 2, dan 4 d. 3, 2, 1 dan 4 e. 3, 4, 2 dan 1 | Kategori: 1 - Soal: Diketahui algoritma P=10; sedangkan P=P+5; nilai Q=P. Oleh karena itu, nilai P dan Q masing-masing adalah a. 15, 0 b. 0, 10 c. 0, 15 d. 10, 15 e. 15 , 15 | Kategori: 1 - Soal: Untuk menangkap objek pada monitor adalah fungsi dari a. New b. RJ-45 c. Spam d. Keyboard e. Take Screenswarm | Kategori: 2 - Soal: 1. Menyalakan sesuai prosedur, 2. Buka command prompt, 3. Kemudian ketik perintah tasklist, 4. Menganalisis hasil yang ditampilkan, merupakan algoritma untuk a. Memantau aktivitas windows b. Menutup program c. Mengecek harddisk d. Menampilkan program apa saja yang sedang berjalan e. Menunjukkan pengguna bekerja direktori utama | Kategori: 2 - Soal: Dibawah ini yang termasuk fungsi command prompt adalah a. Memperbaiki bad sector pada harddisk b. Menghentikan aplikasi secara paksa c. Mengecek status aktivitas windows d. Mengetahui mac address komputer e. Tidak ada yang benar | Kategori: 2

TABLE IV. COMPARISON MODELS IN IDEAL CONDITIONS

Models	Evaluation Metric			
	Precision	*Recall*	*F1-Score*	*Accuracy*
Llama-3-8B	**1.00**	**1.00**	**1.00**	**1.00**
Cohere LLM	0.76	0.60	0.45	0.60
GPT 2	**1.00**	**1.00**	**1.00**	**1.00**

TABLE V. COMPARISON MODELS IN NON-IDEAL CONDITIONS

Models	Evaluation Metric			
	Precision	*Recall*	*F1-Score*	*Accuracy*
Llama-3-8B	**1.00**	0.67	**0.80**	0.67
Cohere LLM	0.85	**0.80**	0.78	**0.80**
GPT 2	**1.00**	0.67	**0.80**	0.67

TABLE VI. COMPARISON MODELS IN SLIGHTLY IDEAL CONDITIONS

Models	Evaluation Metric			
	Precision	*Recall*	*F1-Score*	*Accuracy*
Llama-3-8B	**1.00**	**1.00**	**1.00**	**1.00**
Cohere LLM	0.60	0.67	0.62	0.67
GPT 2	**1.00**	0.67	0.80	0.67

TABLE VII. MODELS COMPARISON IN GENERATING QUESTION

Models	Accuracy		
	Ideal	*Non-ideal*	*Slightly Ideal*
Llama-3-8B	**1.00**	0.67	**1.00**
Cohere LLM	0.60	**0.80**	0.67
GPT 2	**1.00**	0.67	0.67

V. CONCLUSION

The rapid development of technology has an impact on all aspects of life, one of which is in the world of education. In the world of education, technology can help teachers in creating exam questions. However, in general, teachers create questions manually every time an exam is held and the unstructured level of difficulty of the questions that have been created.

To solve this problem, author proposes the RAG method with LLM in creating exam questions. The aim is to help teachers in creating automatic exam questions by generating questions in the question bank or question database. The author uses 3 test scenarios, namely ideal conditions, non-ideal conditions and slightly ideal conditions using 3 models, such as Llama 3, Cohere LLM and GPT 2. The results of each experimental scenario for each model can be seen in table I-III. Meanwhile, to see a comparison of the three models in each test scenario, you can see table IV. From table I-IV it can be concluded that Llama is the best model, then followed by GPT 2 and finally Cohere LLM.

Future work in this research is to add more datasets with different subjects and school levels. In addition, it is recommended to use more models and more test scenarios so that it will be more aware of the accuracy and validity of the question generation.

ACKNOWLEDGMENT

This research was funded by Institut Teknologi Sepuluh Nopember (ITS) under Penelitian Kolaborasi Pusat and under the Project Scheme of the Publication Writing and Intellectual Property Rights (IPR) Incentive (Penulisan Publikasi dan Hak Kekayaan Intelektual/PPHKI) Program 2024.

REFERENCES

[1] S. Chauhan, "A Meta-Analysis of The Impact of Technology on Learning Effectiveness of Elementary Students," *Comput Educ*, vol. 105, no. Pedagogical, pp. 14–30, Nov. 2016.

[2] C. Wang *et al.*, "BioRAG: A RAG-LLM Framework for Biological Question Reasoning," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.01107

[3] A. Sayed, M. Kanojia, and S. Nabajja, "Advanced Subjective Question Bank Generation Using Retrieval Augmented Generation Architecture," in *International Journal of Computer Information Systems and Industrial Management Applications*, Cerebration Science Publishing, Jul. 2024, pp. 264–277.

[4] S. Muller, A. Loison, B. Omrani, and G. Viaud, "GroUSE: A Benchmark to Evaluate Evaluators in Grounded Question Answering," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.06595

[5] A. A. Khan, T. Hasan, K. K. Kemell, J. Rasku, and P. Abrahamsson, "Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report."

[6] G. Guinet, B. Omidvar-Tehrani, A. Deoras, and L. Callot, "Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation," May 2024, [Online]. Available: http://arxiv.org/abs/2405.13622

[7] J. W. Pontes Miranda, H. Bruneliere, M. Tisi, and G. Sunyé, "Towards an In-Context LLM-Based Approach for Automating the Definition of Model Views," in *Proceedings of the 17th ACM SIGPLAN International Conference on Software Language Engineering*, New York, NY, USA: ACM, Oct. 2024, pp. 29–42. doi: 10.1145/3687997.3695650.

[8] H. Jung *et al.*, "Enhancing Clinical Efficiency through LLM: Discharge Note Generation for Cardiac Patients," 2024.

[9] H. Samo, K. Ali, M. Memon, A. Abbasi, M. Yaqoob Koondhar, and K. Dahri, "Fine-Tuning Mistral 7B Large Language Model for Python Query Response and Code Generation: A Parameter Efficient Approach," *VFAST Transactions on Computer Sciences*, vol. 12, no. 1, pp. 205–217, 2024, doi: 10.21015/vtcs.v12i1.1885.

[10] "Llama-3-8B." Accessed: Nov. 09, 2024. [Online]. Available: https://huggingface.co/meta-llama/Meta-Llama-3-8B

[11] "Cohere LLM." Accessed: Nov. 09, 2024. [Online]. Available: https://cohere.com/

[12] "GPT 2." Accessed: Nov. 09, 2024. [Online]. Available: https://huggingface.co/openai-community/gpt2).

[13] R. Abdullah, Suhariyanto, and R. Sarno, "Aspect Based Sentiment Analysis for Explicit and Implicit Aspects in Restaurant Review using Grammatical Rules, Hybrid Approach, and SentiCircle," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 5, pp. 294–305, 2021, doi: 10.22266/ijies2021.1031.27.

[14] Salsabila, R. Sarno, I. Ghozali, and K. R. Sungkono, "Improving cyberbullying detection through multi-level machine learning," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, pp. 1779–1787, 2024, doi: 10.11591/ijece.v14i2.pp1779-1787.

[15] I. Ghozali, K. R. Sungkono, R. Sarno, and R. Abdullah, "Synonym based feature expansion for Indonesian hate speech detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 1–1, 2023, doi: 10.11591/ijece.v13i1.pp1-1x.

[16] A. T. Haryono, R. Sarno, and K. R. Sungkono, "Stock price forecasting in Indonesia stock exchange using deep learning: A comparative study," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 861–869, Feb. 2024, doi: 10.11591/ijece.v14i1.pp861-869.