*Fine-Tuning Large Language Model* (LLM) to Answer Basic Questions for Prospective New Students at Syiah Kuala University Using the *Retrieval-Augmented Generation* (RAG) Method

\*Note: Sub-titles are not captured in Xplore and should not be used

Hary Rachmat   
*Informatics Department*  
*Syiah Kuala University Faculty of Mathematics and Natural Sciences  
Banda Aceh, Indonesia*  
haryrachmat10@gmail.com

Hammam Riza  
*Informatics Department  
Syiah Kuala University Faculty of Mathematics and Natural Sciences  
Banda Aceh, Indonesia*  
hammamriza@usk.ac.id

Taufik Fuadi Abidin  
*Informatics Department  
Syiah Kuala University Faculty of Mathematics and Natural Sciences  
Banda Aceh, Indonesia*  
taufik.abidin@usk.ac.id

*Abstract*—USK Mistral 7B is a large language model designed to answer basic admission questions at Universitas Syiah Kuala (USK). The model was created using the open-source model of Mistral 7B using collected data from admissions and lec-tures at the university. The QLoRA and RAG techniques were used to train the model and retrieve relevant information from external data sources. The results were evaluated using the ROUGE score. Responses were generated with a score of >0.5, ten out of 46 questions with the RAG method and testing with the fine-tuning method was carried out on 20 questions and resulted in responses with a score of 1.0 from all questions asked. The performance of USK Mistral 7B shows its potential as an effective tool in helping students querying information about admission and lectures at USK.

Keywords— Large Language Model, Fine-tuning, RAG

# Introduction

With the rapid advancement of artificial intelligence (AI), large language models (LLMs) have revolutionized natural language processing (NLP) thus allowing computers to interact with text and language. One example of an LLM is Generative Pretrained Transformer4 (GPT-4) which has been extensively trained on large amounts of text data allowing GPT-4 to perform text analysis across multiple domains. LLM can be applied to various fields such as health, education, law, and other fields by conducting pre-training using large amounts of data in these fields.

One of the most popular chatbot applications today is ChatGPT, which was developed by OpenAI and released in late 2022 [1]. Various studies seek to explore ChatGPT's capabilities such as in the medical field developed to aid clinical documentation [2], Classifying texts in banking [3] and in the field of law to determine which laws have the potential to be violated [4].

In addition to closed-source LLMs such as GPT-4, some studies use open-source LLMs such as those conducted by Huang et al.[5] by adapting Llama's LLM to the legal domain to assist lawyers in preparing technical reports. Bhatti et al [6] fine-tuning the LLM Llama 70B named "SM70" to address a wide range of complex medical questions as well as clinical decision-making. Zhao et al [7] fine-tuning the LLM Llama 7B named "Ophtha-LLaMA2" to help diagnose eye diseases that will provide decision support for doctors. Barandoni et al [8] conducted a comparative analysis of LLMs to extract the needs of travel customers from *TripAdvisor* posts by Leveraging a variety of models, including open-source models such as Mistral 7B and closed-source models such as GPT-4 and Gemini. The results highlight the efficacy of open-source LLMs, especially the Mistral 7B, in achieving comparable performance to closed-source models. Research on chatbots as virtual assistants has been conducted by Jonatan & Igor [9] to improve the efficiency of service to customers. The research that has been presented has shown the potential of LLM to help work in various fields.

Based on data obtained from the data.usk.ac.id website [10] The number of students who register for Syiah Kuala University (USK) is increasing every year. To provide various information such as registration details, tuition fees, and others, USK provides a website with usk.ac.id address. So that students can find information about matters related to the university. When there is information that is not yet available on the website, prospective students can ask questions directly through social media such as *direct messages* (DM) through *the Instagram* application, and also meet university staff directly at the Integrated Service Unit (ULT) or Public Relations (PR) section. This study will examine the application of LLM as a virtual assistant in the form of a chatbot to provide interactive information related to academic administration at USK.

# Research Methods

## Dataset Conditions

The data set used is 231 data related to information about the lecture system and new student admissions at USK. From the dataset that has been obtained, the data is then pre-processed, for fine-tuning the Mistral 7B model here using as many as 20 Q&A datasets and converted to Alpaca style format and then stored in a format with .csv extension. A total of 231 datasets are also created in the form of statements and saved in a format with extensions .pdf Extensions are then saved into the *huggingface* repository which will later be used in the next stage, namely the fine-tuning model [11]. And for data with .pdf format will be used in the RAG method [12] so that the data can be managed more easily.

## Quantization and Model Training

In Fig. 1, The fine-tuning process with the Mistral 7B model is divided into three stages, which can be seen as follows:

1. Model Quantization and Data Preparation



Fig. Fine-tuning *process* on the Mistral Model 7B

The fine-tuning process on the Mistral 7b model is trained using the help of the transformer library to quantize the Mistral 7B, the BitsAndBytesConfig library interface is used to quantize the Mistral 7B USK model. The purpose of this BitsAndBytes library is to reduce the precision value of the floating-point in the weight of the USK Mistral Model 7B which is processed by quantization from precision with high value to precision with lower value. In this process, the weights on the Mistral 7B USK model are converted into int4 format through a quantization layer and stored in the GPU. The main computing process is performed on the CUDA, which will reduce memory usage and will improve the efficiency of the model. This makes it possible to fine-tune larger LLMs on consumer-grade GPUs. The fine-tuned model is stored in the *huggingface repository.*

1. Pre-trained Model

Before the training, an analysis of well-known LLMs from the current LLM pool was carried out to select the most suitable model for the purpose of this research. Evaluation of LLMs mainly takes into account several key aspects such as adaptability to specific domains, compatibility in academic standards, bilingual language skills, availability of models on open source, efficiency in parameters, and by looking at cost and licensing considerations on the model. This consideration also looks at several evaluation metrics that have been tested across various benchmarks.

1. Parameter-Efficient Fine-Tuning (PEFT)

Taking into account the limited dataset and to reduce the cost of the training process and to reduce the potential for failure risk, the researchers chose the Progressive Layer Freezing and Fine-tuning (PEFT) method [13] to refine the Mistral 7B model. PEFT selectively reduces a small number of parameters in the additive model, in this way model training can be used to significantly reduce computational costs and memory storage costs, and will make it possible to efficiently adapt pre-trained LLMs for use in various application domains.

In this study, a low-level adjustment method (QLoRA) was specifically used [14] to improve large language models. Lora Method [15] involves freezing on pre-trained weights on the original model and creating a new version on the matrix with lower rank values for layers and query values. This lower-rated matrix has values on significantly fewer parameters than the original model, allowing for adjustment of memory usage on GPUs with smaller storage sizes. The advantage of this method can be seen in the ability in many LoRA adapters to reuse native LLMs, thereby reducing the overall memory usage required when providing answer text in use cases in tasks on specific domains as well as in various cases to be applied. Unlike LoRA, the QLoRA method represents an iteration of the LoRA method that will save more storage on memory. The QLoRA method takes LoRA one step further by measuring the value or weight on a LoRA adapter with a smaller matrix value to a lower precision value (e.g., the model weight value becomes 4-bit, and not the value on a model with an initial value of 8-bit). This can reduce the memory size and requirements on model storage. In the QLoRA method, the trained model is loaded into GPU memory with a quantized 4-bit weight value, in contrast to the 8-bit model used in the LoRA method. Despite the decrease in bit precision values, the QLoRA method can maintain a level of effectiveness comparable to that of the LoRA method. So this study uses a parameter-efficient improvement method with QLoRA. The fine-tuning method used for LLM training is by using the Unsloth library and Huggingface's TRL library[16]. The Unsloth library can make finetuning LLMs 2x faster.

The hyperparameter settings used in the fine-tuning process using the Unsloth library and Huggingface's TRL library are shown in the following table.

Table 1 shows the configuration for creating *an FastLanguageModel.from\_pretrained* instance, using *the unsloth* component, with specific configurations such as maximum sequence length, data type, and 4-bit loading [17].

Table FastLanguageModel.from\_pretrained Configuration

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| model\_name | mistralai/Mistral-7B-v0.1 |
| max\_seq\_length | 2048 |
| dtype | None |
| load\_in\_4bit | True |

In Table 2 the FastLanguageModel *object*  provides attributes *get\_peft\_model* a place to be able to configure various parameters for customization, such as the number of *attention heads*, target module, *dropout rate*, *LoRa alpha*, and more. The use of *checkpointing gradients* and other advanced techniques demonstrates *unslotting's* ability to optimize model performance [17].

Table FastLanguageModel.get\_peft\_model Configuration

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| r | 16 |
| target\_modules | ["q\_proj", "k\_proj", "v\_proj", "o\_proj", "gate\_proj", "up\_proj", "down\_proj",], |
| lora\_alpha | 16 |
| lora\_dropout | 0 |
| bias | None |
| use\_gradient\_checkpointing | "unsloth" |
| random\_state | 3407 |
| use\_rslora | False |
| loftq\_config | None |

Next steps in Table 3 and Table 4 is to initialize the *Supervised Fine-tuning Trainer* which helps the fine-tuning process. Starting with initializing the model, along with the dataset to be refined, tokenizer and all necessary Training Arguments (learning speed, maximum steps, weight reduction, optimization, etc) [17].

Table *SFTTrainer* Configuration

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| model | model |
| tokenizer | tokenizer |
| train\_dataset | dataset |
| dataset\_text\_field | "text" |
| max\_seq\_length | 2048 |
| dataset\_num\_proc | 2 |
| packing | False |

Table *TrainingArguments* Configuration

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| per\_device\_train\_batch\_size | 2 |
| gradient\_accumulation\_steps | 4 |
| warmup\_steps | 5 |
| max\_steps | 60 |
| learning\_rate | 2e-4 |
| fp16 | not is\_bfloat16\_supported() |
| bf16 | is\_bfloat16\_supported() |
| logging\_steps | 1 |
| optim | "adamw\_8bit" |
| weight\_decay | 0.01 |
| lr\_scheduler\_type | "linear" |
| seed | 3407 |
| output\_dir | "outputs" |

1. USK Mistral 7B Approach Model with RAG Method

At this stage, an approach was carried out on the Mistral 7B USK using the RAG method [18], This method aims to overcome the limitations of generative AI when it requires information that is outside the Mistral 7B USK training corpus, so that this method will avoid Mistral 7B USK which will produce inaccurate text, hallucinations, or distortions when providing answers to the given questions. in the RAG method, the data used is an external document that contains information on the USK academic system and is stored in an extended format .pdf. The data set of information is then embedding to convert text into vectors so that it can be stored into a database vector. The database vector used in this study is FAISS. The model to be used at this stage is a model that has been quantized into GPT-Generated Unified Format (GGUF) [19] so that it is possible to use the CPU when running the LLM by moving some of its layers to the GPU so that it can accelerate the model's performance in generating text. The query process on the model in retrieving the most relevant context of the user command with the RAG method is shown in Fig. 1.



Fig. . The RAG pipeline, during the query phase, takes the most relevant context from the user's commands, passing them to the Large Language Model (Source: docs.llamaindex.ai, 2023)

# Results and Discussion

The output produced in this study is a *web-based* chatbot application utilizing the *Large Language Model* (LLM) *Mistral 7B* which will be used as an alternative in helping prospective new students get information at Syiah Kuala University (USK). This *chatbot* was developed using the USK dataset by utilizing the ability to summarize LLM texts to generate information texts about the lecture system and new student admissions at USK. There are several stages that are carried out in utilizing and developing LLM, namely by collecting data related to lecture system information and new student admissions at USK, then the data collection will be carried out in *the preprocessing* stage by converting the raw data into .csv and .pdf formats. After the *preprocessing stage*, finetuning was carried out on the *Mistral 7B* model and also to make it easier to manage data*, Retrieval Augmented Generation* (RAG) was carried out with the concept of *embeddings*. The purpose of the RAG, which is to overcome the limitations of generative AI because whenever there is a question that requires information that is outside the LLM training corpus, it will result in hallucinations, inaccuracies, or distortions in the generated text. The next stage is to test and evaluate the generated text with the ROUGE method to compare the text generated by the model with the summary of the provided references. The last stage is to create a UI or web interface that is accessible to users.

## Testing Results and Evaluation of Inference Results

At this stage, the chatbot was tested by asking questions, the resulting text inference results were then calculated using the *Recall-Oriented Understudy for Gisting Evaluation (*ROUGE) score method. Using the ROUGE metric, responses from the chatbot can be evaluated to see similarities between the reference quantitatively and the answers generated by the chatbot. The results of the ROUGE calculation score on the model are shown in

Table Score value of ROUGE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Number of  Questions | ROUGE Score | | |
| R-1 | R-2 | R-L |
| *Fine-tuning* | 20/20 | 1.0 | 1.0 | 1.0 |
| RAG | 15/56 | >0.5 | >0.5 | >0.5 |

## ROUGE score categories

A good ROUGE score varies based on tasks on the summary and metrics. The ROUGE-1 score with the category is excellent with a score of around 0.5, and a score above 0.5 is considered good and 0.4 to 0.5 is moderate. For ROUGE-2, a score above 0.4 falls into the good category, and 0.2 to 0.4 falls into the moderate category.

The ROUGE-L score with a good category gets a score of around 0.4 and the low category ranges from a value of 0.3 to 0.4. While the ROUGE score is useful, it does not take into account semantic or syntactic qualities and should be complemented by other metrics and human evaluation for a complete assessment [21].

Table ROUGE Metric Value Category Table [21]

|  |  |  |  |
| --- | --- | --- | --- |
| ROUGE  Metric | Excellent | Good | Moderate |
| ROUGE-1 | 0.5+ | >0.5 | 0.4-0.5 |
| ROUGE-2 | - | >0.4 | 0.2-0.4 |
| ROUGE-L | - | ~0.4 | 0.3-0.4 |

## Calculating Resource Evaluation

In this study, the researcher conducted a time test on the Mistral 7B model, with *fine-tuning* and inference testing. Researchers used a single NVIDIA Tesla T4 GPU graphics card available in Google Colab for inference and also for *fine-tuning*. Based on the *fine-tuning* time test and using 20 datasets, as shown in Table 6, the researcher concluded that the *fine-tuning* approach with *the unsloth* library is effective and efficient, and requires minimal time. The inference time test shows that the USK Mistral 7B does not require excessive computing power during the inference process, making it an energy-efficient and efficient system to respond to user inquiries, only it takes 5-6 minutes to respond to user inquiries. It is hoped that this model will be an efficient system in handling information related to USK's academic administration in the future.

Table Time Count on model while *fine-tuning* and running RAG

|  |  |  |
| --- | --- | --- |
| Model | Fine-tuning *time* (hour) | RAG Time (minute) |
| USK Mistral 7B | 2 | 5-6 |

## Result Analysis

## Training Data Problems

An important factor contributing to LLM hallucinations is the nature of the training data. LLMs, such as Mistral 7B, GPT, Falcon, and Llama, undergo extensive unattended training with large and diverse datasets from various origins. Verifying the fairness, impartiality, and factual correctness of this data is a challenge. As the model learns to generate text, it can also find and replicate factual inaccuracies in the training data. This leads to scenarios where the model cannot distinguish between truth and fiction and can produce outputs that deviate from facts or logical reasoning. LLMs trained on datasets sourced from the internet can contain biased or incorrect information. This misinformation can spread to the model's output, as the model cannot distinguish between accurate and inaccurate data.

## Reduces Hallucinations

Efforts to reduce hallucinations are essential to maintain the credibility and functionality of LLMs. The primary method for identifying and mitigating these errors involves a combination of advanced metrics and critical human evaluation. These include.

* Linguistic quality metrics such as ROUGE and BLEU
* Content validity metrics, which are IE-based, QA-based, and NLI-based
* *FactScore* to check the accuracy of individual facts

## Retrieval-Augmented Generation (RAG) Method

Metode inovatif seperti *SelfCheckGPT* mendeteksi halusinasi dengan menilai konsistensi beberapa jawaban yang dihasilkan untuk pertanyaan yang sama. Selain itu, teknik seperti *chain-of-thought prompting* dan *Retrieval-Augmented Generation* (RAG) terus dieksplorasi untuk memperkuat kemampuan model dalam memberikan informasi yang tepat dan relevan.

## The Influence of GPUs in LLM Implementation

In running LLMs, GPUs play an important role. Dedicated GPUs with high VRAM can significantly accelerate the computation required by the model. In this study, the GPU used is the "NVIDIA Tesla T4 GPU" which is available on Google Colab for free, the results of testing with this GPU using the RAG method take 4-5 minutes to be able to generate responses to the questions asked.

# Conclusion

*Fine-tuning* the Mistral 7B model to USK Mistral 7B requires a lot of data so that it can produce a better LLM in answering questions related to the lecture system and new student admissions at USK. And it takes a long time approximately 2 hours to get a *fine-tuning* model with a dataset of 20 question and answer data.

*The Retrieval Augmented Generation* (RAG) method overcomes the limitations of generative AI when it requires information that is outside the LLM training corpus, so that this method will avoid LLMs that will generate inaccurate text, hallucinations, or distortions when providing answers to given questions. This RAG method can be used to generate answers faster because it uses external data. The RAG method allows the model to avoid limitations on generative AI models. The response generated by the RAG method is able to produce a fairly good answer by looking at the ROUGE score that has been tested.

Based on open-source *engineering metrics and assessment of computing resources*, it can be concluded that USK Mistral 7B has the potential to be applied because with low energy consumption it can produce a response that has a ROUGE score of >5.

LLM, especially Mistral 7B has extraordinary potential in its application in various fields such as in the field of academic and administrative services. This study shows that with the training of a small amount of information data related to academics and administration at USK, USK Mistral 7B is able to respond well to various questions. The impressive performance of LLMs such as the Mistral 7B highlights its ability as a powerful tool to assist students in obtaining information at USK.

##### References

[1] S. Mohamadi, G. Mujtaba, N. Le, G. Doretto, and D. A. Adjeroh, “ChatGPT in the Age of Generative AI and Large Language Models: A Concise Survey,” pp. 1–60, 2023, [Online]. Available: http://arxiv.org/abs/2307.04251

[2] H. P. Baker, E. Dwyer, S. Kalidoss, K. Hynes, J. Wolf, and J. A. Strelzow, “ChatGPT’s Ability to Assist with Clinical Documentation: A Randomized Controlled Trial,” *J. Am. Acad. Orthop. Surg.*, vol. 32, no. 3, pp. 123–129, Feb. 2024, doi: 10.5435/JAAOS-D-23-00474.

[3] L. Loukas, I. Stogiannidis, P. Malakasiotis, and S. Vassos, “Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance,” *FinNLP-Muffin 2023 - Jt. Work. 5th Financ. Technol. Nat. Lang. Process. 2nd Multimodal AI Financ. Forecast. conjunction with IJCAI 2023 - Proc.*, pp. 74–80, 2023.

[4] A. Trozze, T. Davies, and B. Kleinberg, “Large Language Models in Cryptocurrency Securities Cases: Can ChatGPT Replace Lawyers?,” pp. 1–49, 2023, [Online]. Available: http://arxiv.org/abs/2308.06032

[5] Q. Huang, M. Tao, C. Zhang, and Z. An, “Lawyer LLaMA: Enhancing LLMs with Legal Knowledge,” 2023.

[6] A. Bhatti, S. Parmar, and S. Lee, “SM70: A Large Language Model for Medical Devices,” no. 1, pp. 1–5, 2023.

[7] H. Zhao *et al.*, “Ophtha-LLaMA2: A Large Language Model for Ophthalmology,” *ArXiv [Preprint]*, pp. 1–19, 2023, [Online]. Available: http://arxiv.org/abs/2312.04906

[8] S. Barandoni, F. Chiarello, L. Cascone, and S. Puccio, “Automating Customer Needs Analysis : A Comparative Study of Large Language Models in the Travel Industry,” 2024.

[9] V. Jonatan and and A.-A. Igor, “Creation of a Chatbot Based on Natural Language Processing for Whatsapp,” *J. Database Manag.*, vol. 3, no. 4, pp. 39–53, 2023, doi: 10.14810/elelij.2023.12402.

[10] data.usk.ac.id, “Data Mahasiswa Daftar Ulang,” data.usk.ac.id. Accessed: Apr. 13, 2024. [Online]. Available: https://data.usk.ac.id/mahasiswa-daftar

[11] T. Fatyanosa, “Fine-Tuning Pre-Trained Transformer-based Language Model,” medium.com. Accessed: Jan. 25, 2024. [Online]. Available: https://fatyanosa.medium.com/fine-tuning-pre-trained-transformer-based-language-model-c542af0e7fc1

[12] K. T. Wicaksono, “Memahami Retrieval Augmented Generation (RAG) Melalui Rencana Pembangunan LaporMacet.com,” medium.com. Accessed: Jan. 25, 2024. [Online]. Available: https://kukuhtw.medium.com/memahami-retrieval-augmented-generation-rag-melalui-rencana-pembangunan-lapormacet-com-5990a76edda7

[13] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, “Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment,” pp. 1–20, 2023, [Online]. Available: http://arxiv.org/abs/2312.12148

[14] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLORA: Efficient Finetuning of Quantized LLMs,” *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.

[15] E. Hu *et al.*, “Lora: Low-Rank Adaptation of Large Language Models,” *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, pp. 1–26, 2022.

[16] D. Han-Chen, “Make LLM Fine-tuning 2x faster with Unsloth and ? TRL,” https://huggingface.co/blog/unsloth-trl. Accessed: Jun. 30, 2024. [Online]. Available: https://huggingface.co/blog/unsloth-trl

[17] K. V., “Fine-Tuning Large Language Models with Unsloth,” medium.com. Accessed: Jul. 30, 2024. [Online]. Available: https://medium.com/@kushalvala/fine-tuning-large-language-models-with-unsloth-380216a76108

[18] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, 2020.

[19] M. Grootendorst, “Which Quantization Method is Right for You? (GPTQ vs. GGUF vs. AWQ),” maartengrootendorst.com. Accessed: Feb. 29, 2024. [Online]. Available: https://www.maartengrootendorst.com/blog/quantization/

[20] docs.llamaindex.ai, “High-Level Concepts,” docs.llamaindex.ai.

[21] S. M. Walker II, “What is the ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)?,” klu.ai. Accessed: Jul. 30, 2024. [Online]. Available: https://klu.ai/glossary/rouge-score