Volume 9, Issue 5, May – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

Assessing Fine-Tuning Efficacy in LLMs: A Case Study with Learning Guidance Chatbots

Rabia Bayraktar[1]; Batuhan Sarıtürk[2]; Merve Elmas Erdem[3]
Huawei Turkey R&D Center Istanbul, Turkey

Abstract:- Training and accurately evaluating task-specific chatbots is an important research area for Large Language Models (LLMs). These models can be developed for general purposes with the ability to handle multiple tasks, or fine-tuned for specific applications such as education or customer support. In this study, Mistral 7B, Llama-2 and Phi-2 models are utilized which have proven success on various benchmarks, including question answering. The models were fine-tuned using QLoRa with limited information gathered from course catalogs. The fine-tuned models were evaluated using various metrics, with the responses from GPT-4 taken as the ground truth. The experiments revealed that Phi-2 slightly outperformed Mistral 7B, achieving scores of 0.012 BLEU, 0.184 METEOR, and 0.873 BERT. Considering the evaluation metrics obtained, the strengths and weaknesses of known LLM models, the amount of data required for fine-tuning, and the effect of the fine-tuning method on model performance are discussed.

Keywords:- *LLM, Mistral, Llama, Phi, Fine-Tune, QLoRa.*

I. INTRODUCTION

With the rise of open-source Large Language Models (LLMs) in recent years, customized chatbots have become prevalent in many applications. The fact that most LLM have been trained on huge data and have been developed for multiple tasks makes it easier to use these models for small-scope and custom use cases. Today, the most well-known LLMs with proven effectiveness are GPT [1] and its variations, which allow for the development of applications by directly using an API. However, the number of open-source high-performance large language models [2, 3, 4] continually increases. As a result, access to both pre-trained base models and fine-tuned models for specific use cases is becoming easier. [5]

There are numerous applications for customized chatbots for providing product recommendations on online shopping sites [6], enhancing customer relations in telecommunications [7], and creating question-answering systems in specific domains [8]. Adaptation of a language model for a specific use case can be achieved through various methods, including zero-shot learning, few-shot learning, parameter-efficient fine-tuning, or prompt tuning. During pre-training, an LLM will have already accessed the most publicly available information sources, enabling its use in various applications without requiring fine-tuning. But for a custom use case, for example, a domain-specific question-answering system, data usage and fine-tuning will be crucial. This process aims to ensure that the pre-trained model can learn as much information as possible with limited data. In addition, techniques such as Retrieval-Augmented Generation (RAG) can enable the model to obtain information from additional external sources during inference. [9]

This study involved the development of a chatbot that provides information about training contents and helps in planning career roadmaps in various software fields. To create a question-answering pair dataset, public data from Huawei's ICT Learning course catalog [10] was used, and state-of-the-art language models were fine-tuned using this dataset. The study focuses on analyzing the development of an effective fine-tuned model with a small amount of data within a limited scope. The remainder of the paper includes a literature review, method, and experimental results.

II. LITERATURE REVIEW

LLM studies have become increasingly popular, especially since the release of GPT-3 in 2020. Many different model architectures have been released, both open-source and commercial. The development of these large models has made it easy to adapt them to different use cases, leading to their widespread application in various business domains such as education, medicine, software, mathematics, law, finance, and coding [11, 12]. Most widely used open-source LLMs have performed well in various benchmarks are [2, 3, 4].

The generalization ability of these language models is very good, but fine-tuning is also necessary to handle cases such as enabling response safety in terms of ethics and toxicity, and understanding user intents for specific domains. An example of fine-tuning language models for textbook question answering is shown in [13], which investigates the effect of a Llama model fine-tuned with QLoRa [14] on the RAG output. The study used a total of 1076 course data with different contents. The model was trained with a total of 26,260 questions, with approximately 15,000 of the data used for training and 10,000 for testing and validation. The validation accuracy of the model trained with A100 GPU was obtained as 82.40, and the test accuracy as 84.24 [6].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

Integration of chatbots into educational frameworks has shown promising results. Some studies propose educational chatbots in specific domains [15] or for general purposes [16]. In a study conducted in 2024 [8], synthetic data was used for subsections determined using the dialogue dataset. Additionally, [7] proposes an educational chatbot as a mobile learning tool and provides implementation details. Furthermore, [18] offers a comprehensive overview of the stages involved in developing and implementing LLM-based systems. Another domain that can be supported by chatbots is health. [8] discusses how chatbots can provide assistance during health emergencies. These systems could evolve to triage patients, provide preliminary diagnoses, and offer medical advice, potentially saving lives in critical situations.

Evaluation of chatbot performance is challenging because LLMs are trained for multiple tasks and can be adapted to various subtasks. When ground truth data is available, similarity metrics such as BLEU, ROUGE, METEOR, and BERTScore can be used. For instance, in [17], the BLEU score is used to calculate the accuracy of generated responses in an educational chatbot. Another way to measure response accuracy is by using AI. In [1], the evaluation of a fine-tuned language model is done using GPT-4.

III. MATERIAL & METHODS

Our research aimed to compare the performance LLMs such as Mistral, Llama, and Phi by incorporating public data from Huawei ICT Academy courses catalog and demonstrate that a platform-specific chatbot model can achieve effective results when trained with a limited dataset. To achieve this, we used a fine-tuning methodology that exposed these models to question and answer (Q&A) pair dataset generated by using edicational materials from course catalog. By integrating information from these sources, we aimed to enrich the models' understanding and adaptation to industry-specific terminology, concepts, and contexts. This comprehensive approach allowed for a more robust enhancement of the models' proficiency in comprehending and responding to queries within the talent platform domain, aligning with our goal to strengthen their utility and effectiveness in real-world applications, particularly in their engagement with talent platforms.

A. Dataset

Data collection studies were carried out from the talent platform for a customized chatbot. The transaction was made using public data on the platform and kept limited with the ICT Academy 23 Courses Catalog. A dataset comprising 500 Q&A pairs was created using GPT-4 from this content to fine-tune the Mistral model. A sample question-answer pair can be seen in Fig. 1.

The content was extracted from 14 courses to create the question and answer pairs, with 7 courses focused on professional topics and 7 courses covering general subjects. Questions and answers of varying lengths were generated using GPT-4. When creating the questions, we followed a progression from general to specific to enable the chatbot to better generalize the information it learned.

B. Language Models


```
{"input": "How do self-driving cars perceive their surroundings?",
"output": "Self-driving cars use sensors and computer vision
algorithms to interpret the environment and make driving decisions."}
```


Fig 1: Example of a Q & A Pair

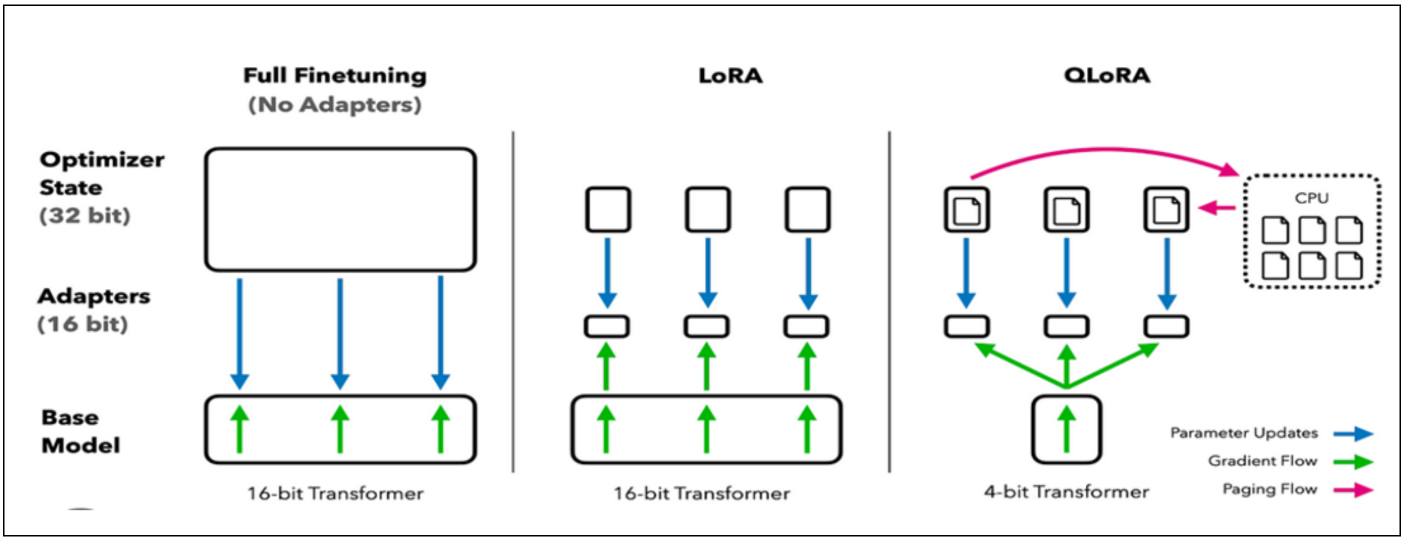

Fig 2: QLoRa

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

LLMs are artificial intelligence models with the ability to process and generate text, which falls under the category of Natural Language Processing (NLP). Depending on the characteristics of the problem to be solved, various models can be created. Most of these models are pretrained for multiple tasks and include transformer architectures, which allow for easy post-training adaptation to specific tasks. To effectively generalize the trained model, the dataset should contain as many different samples as possible. LLMs require a large dataset and a long training time due to their size, which also brings computational challenges.

Many LLMs have been updated since 2020 and have been proven to be effective. The superiority of the models offered by developers such as OpenAI, Google, Meta, and Stability AI against each other in some cases has been examined. When the size of the models used is examined, it is known that GPT uses 175 billion parameters, Gemini uses 3.25 billion, and Gemma uses 7 billion parameters. When fine-tuning LLMs with limited resources, managing the number of parameters is crucial. Fewer parameters mean lower computational and memory requirements, making the process more feasible at the cost of reduced performance.

LLMs can be easily adapted to the new domains and new tasks such as sentiment analysis, customer service, search, question answering etc. One of these models, Llama, is a model offered by Meta that uses approximately 70B parameters. The model called Llama2-chat was trained with the dialogue dataset provided for customized tasks. In cases where the resource needs of the 70B parameter model cannot be met, llama-2's 7B parameter chat model can be used.

The Mistral model, one of Llama's closest competitors, has two different uses: instruct and base. A significant increase in performance is observed with the decoder-only transformers used in the model trained using 7B parameters.

The Phi series has been a significant breakthrough in the field of language models. The original Phi model was trained with 1.3 billion parameters, and its successor, Phi-2, was trained with an even more extensive 2.7 billion parameters. Phi-2 has set new standards among base language models with less than 13 billion parameters, showcasing remarkable reasoning and language understanding capabilities.

C. QLoRa

QLoRa represents a significant advancement in the field of NLP, particularly in the realm of optimizing LLMs for efficiency and accessibility. Developed as an extension of the LoRa framework, QLoRa introduces a novel approach to parameter quantization, profoundly impacting memory utilization and hardware requirements (Fig. 2).

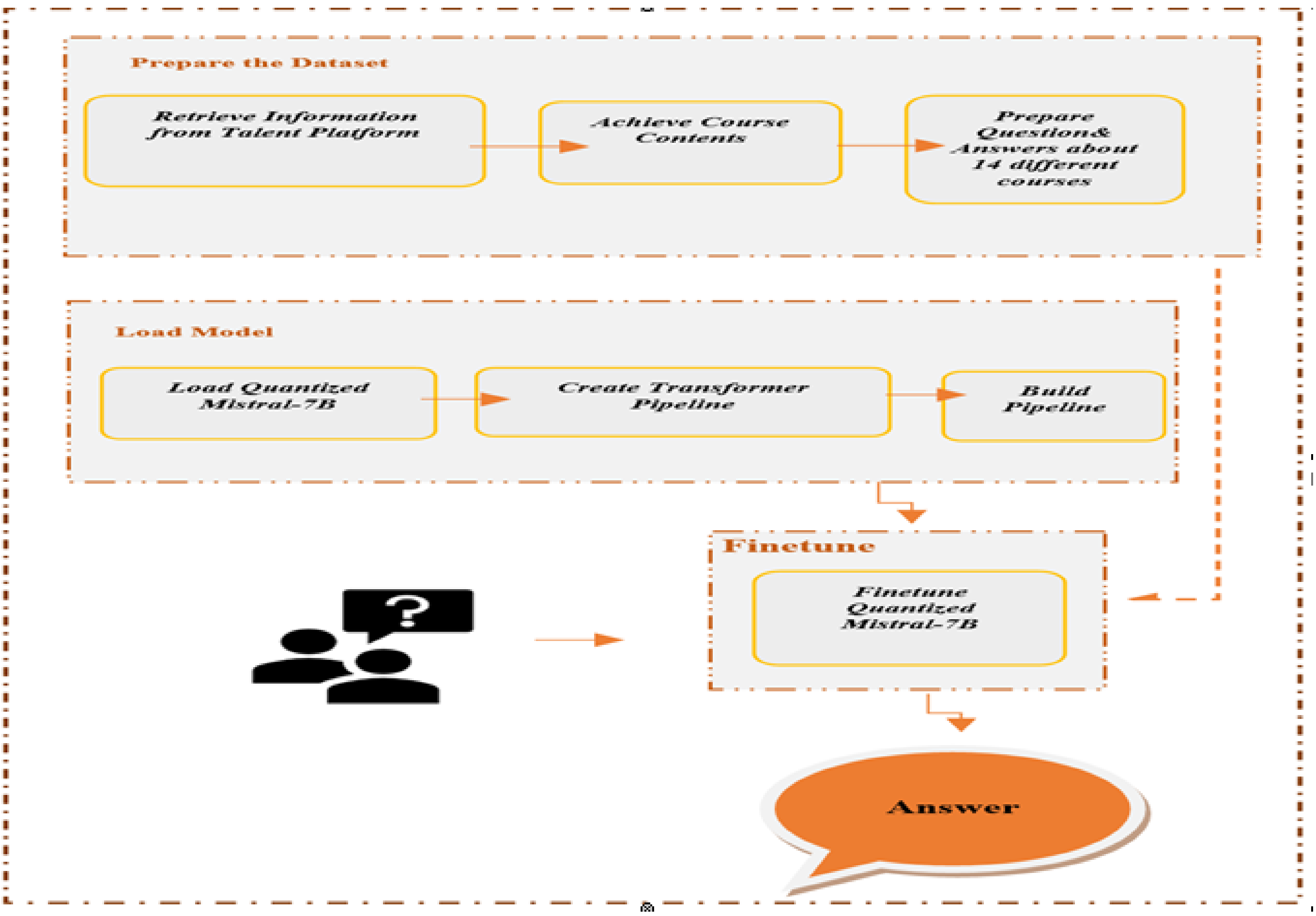

Fig 3: Model Fine-Tune Diagram

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

At its core, QLoRa leverages quantization techniques to reduce the precision of weight parameters within pre-trained LLMs to an unprecedented 4-bit format. Unlike traditional models that store parameters in a 32-bit format, QLoRa's compression significantly diminishes the memory footprint of LLMs, unlocking the potential for fine-tuning operations on a single GPU. This breakthrough not only streamlines memory consumption but also democratizes access to cutting-edge language models by enabling their execution on consumer-grade hardware [14].

D. Fine-Tuning

The main goal of fine-tuning is to achieve the highest accuracy with minimal customization requirements for a model. In many cases, good results can be obtained by retraining a well-trained model with your own data. This approach requires much fewer resources compared to training a model from scratch.

There are many PEFT (Parameter Efficient Fine-tuning) methods for fine-tuning LLMs. LoRa is the most popular of these methods. It can be integrated into the fully connected layers of the existing model. It aims at optimization through the parameters it adds, without increasing the complexity of the model. QLoRa works on the original network to further increase the effectiveness of LoRa.

It scales the weight values of the original network from Float32, which is a high-resolution data type, to Int4. By rescaling, the memory needs of the trainee and the training time are reduced. It is very suitable to use in situations where there is a lot of data and limited hardware resources. In this study, the Mistral-7B model was fine-tuned with QLoRa.

In this study, we proposed a method for fine-tuning the base models using 500 Q&As. The diagram in Fig. 3 illustrates the data exchange between the main and helper units that make up the system. Fine-tuning processes requires the base model pipeline to be loaded, and transform matrices must be created. The Mistral-7B base model was used in this study, and we drew from the hugging face of the model pipeline to define transform and tokenizer rules.

After building the pipeline, we conducted new training using the QLoRa model selected for fine-tuning. QLoRa was chosen for fine-tuning in this study because it can more effectively prevent overfitting and underfitting when retraining large models with small datasets.

In the last phase, with the use of the fine-tuned model, users are expected to find answers to their questions regarding training content, course format, duration, and suitable courses based on their expertise level in a specific domain.

IV. EXPERIMENTS & RESULTS

A. Fine-Tuning with QLoRa

Various experiments were carried out to fine-tune models frequently mentioned in the literature with QLoRa. In all experiments, 500 Q&As generated from the catalog data were used for model training. Q&A pairs contain general and course-specific questions and answers about the platform. Care was also taken to ensure that the questions and answers were of different lengths. Different step and batch sizes were used for the Phi-2, Llama-2 and Mistral 7B models depending on the model requirements.

During the fine-tuning process of the Phi-2 model, a 5k step training was performed with A100 GPU and 8 batch size. The training took approximately 2 hours. Train and evaluate losses of the training are shown separately in Fig. 4. 10% of the training data was used for evaluation.

The loss graphs illustrate how well the model is learning the provided material. As the training progresses, the loss consistently decreases, reaching levels as low as 0.3 to 0.2. This indicates that the model is effectively understanding and absorbing the content. Due to the limited amount of available training data, we decided to limit the number of training steps to prevent overfitting. Overfitting can occur when the model becomes too specialized in the training data and doesn't perform well on new, unseen data. Despite this cautious approach, the model still managed to learn the content quite well. The decreasing loss values demonstrate that the model is improving at minimizing errors and enhancing its accuracy in predicting answers. By being cautious with the number of training steps, we aimed to strike a balance between allowing the model to learn as much as possible and ensuring it doesn't become too narrowly focused on the training data. Despite these limitations, the model exhibits strong proficiency in understanding and retaining the course material, which validates the effectiveness of our training approach.

The recommended step size is 300 to prevent overfitting when fine-tuning the Llama model with small datasets like ours. When Llama examples are examined in the literature, it is seen that long training is carried out with sufficient hardware resources, but 300 steps for fine-tuning with QLoRa are enough to compare with other models. For this study, training was conducted with 8 batch sizes and 300 steps, but the expected curves in the loss graphs were not achieved as shown in Fig. 5. For this reason, a new training was conducted and evaluated with 600 step size.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

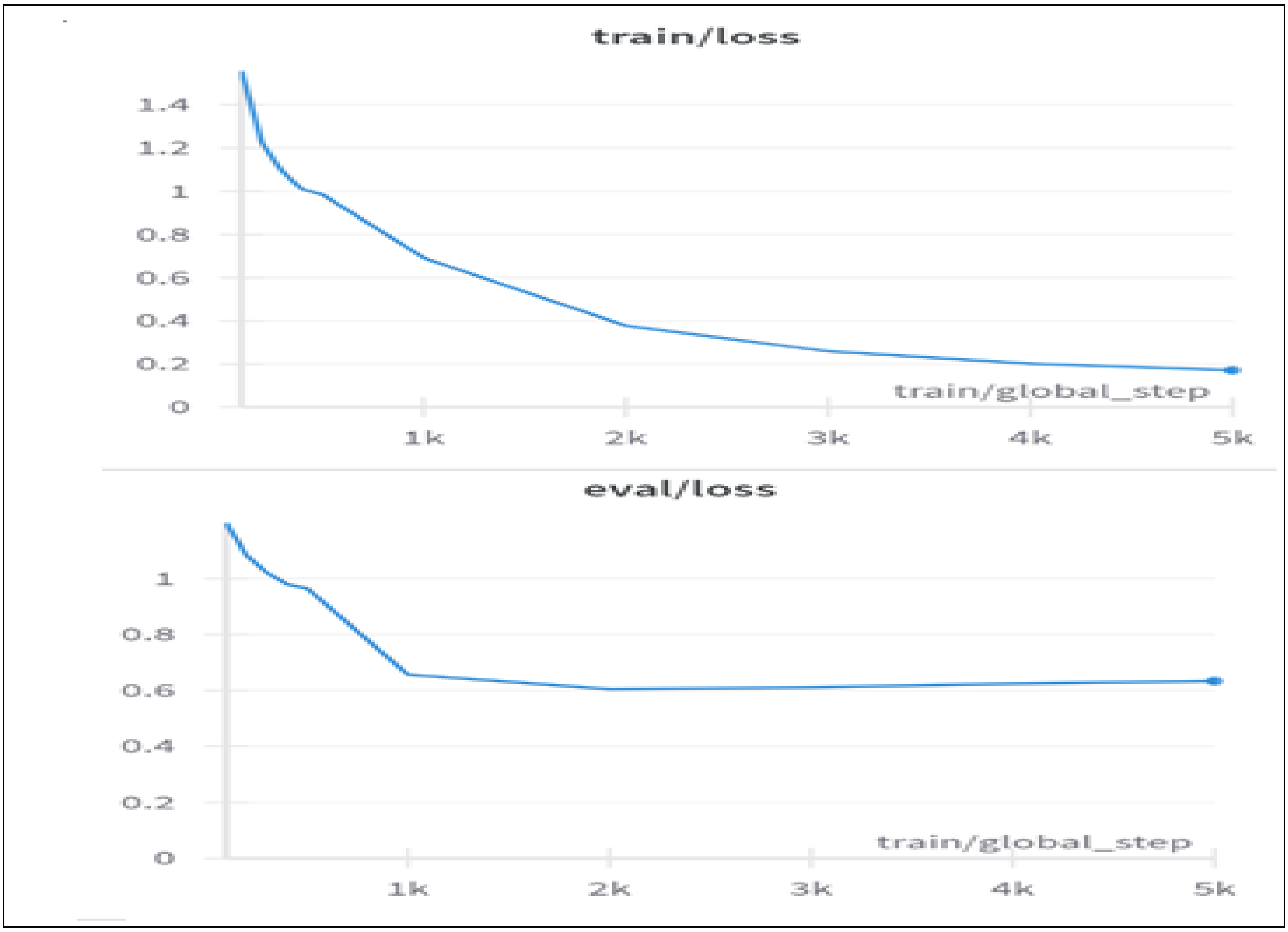

Fig 4: Phi-2 Loss Graphs

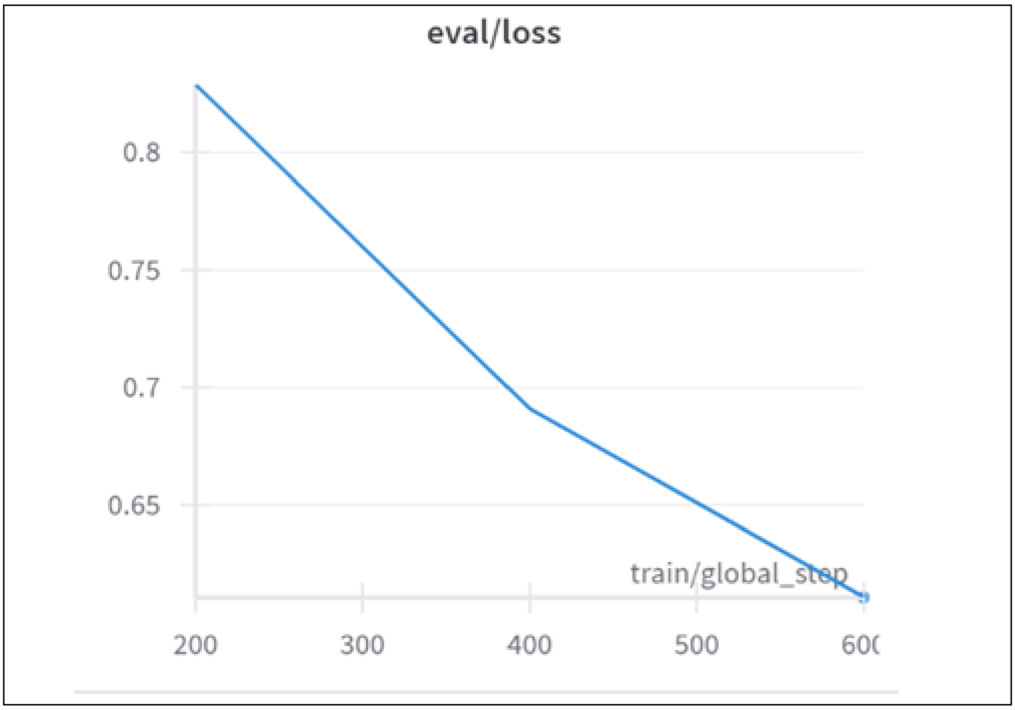

Fig 5: Llama-2 Loss Graph

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

In the experiments where 600 steps and 8 batch sizes were used, it was observed that the Llama model requires more memory compared to other models. Upon reviewing Table 1, it was noted that the model's generated answers often do not correlate with the questions. This indicates a weaker content relationship in the Q&A compared to the Phi-2 model. Analysis of the training and evaluation loss graphs shows a consistent decrease. It is anticipated that with longer training sessions using more powerful GPUs, the train and test losses may drop below ~0.4.

It has been reported that Mistral 7B performed better than the Phi-2 and Llama-2 models in the training conducted with benchmark datasets. The loss graphs obtained from our study using Mistral 7B are displayed in Fig. 6. The consistent decrease in the loss graphs of the model trained with 500 Q&As is evident. It is observed that the training conducted with a batch size of 8 and 300 steps outperforms the Llama-2 model trained with 600 steps. The results we obtained align with the benchmark results.

B. QA Comparison

In this section, a Q&A comparison of 3 different models that we fine-tuned is presented. Based on the data on which the chatbot was trained, the same questions were asked to three different models and their answers were evaluated. It was monitored how well the content obtained from the Talent Platform matches the answers given by the bot.

Evaluating Q&A models presents unique challenges as it requires not only the accuracy of information but also the relevance and naturalness of the responses [19]. Traditional metrics used in other forms of NLP tasks, such as translation or summarization, also apply here but with additional considerations due to the dialogic and informational nature of Q&A systems [19].Q&A models are best evaluated through a combination of metrics that can capture both the precision of the answers in terms of content and their linguistic quality, and the metrics selected to be used in this study are particularly suitable for this task.

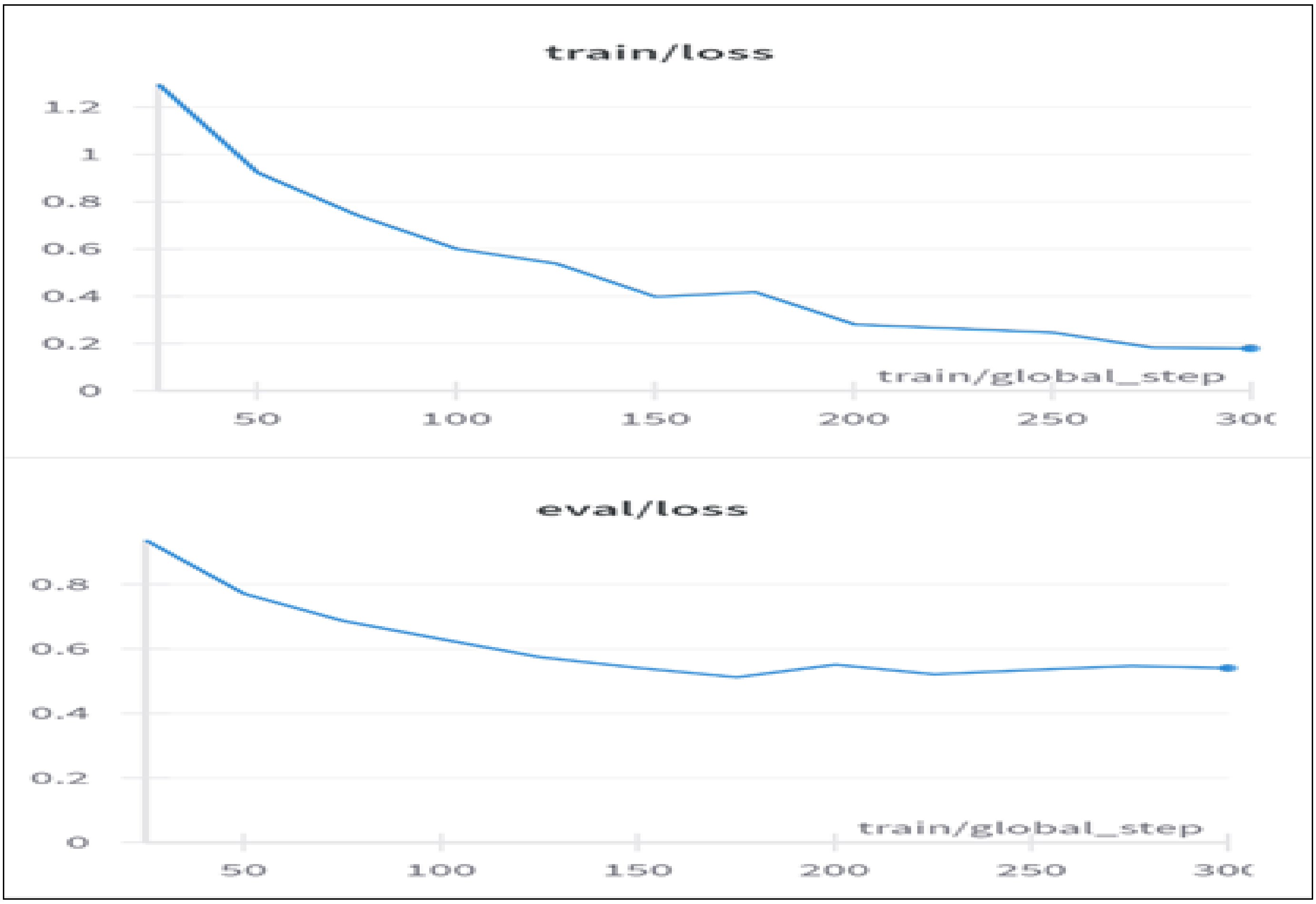

Fig 6: Mistral 7B Loss Graph

To assess the performance of the models, the chosen metrics for this study are BLEU, ROUGE, METEOR, and BERTScore, each offering unique insights into different aspects of text generation quality. BLEU (Bilingual Evaluation Understudy) is used to measure the similarity between machine-generated text and human reference translations [20]. It is particularly useful in Q&A settings where accurate terminology is crucial. BLEU evaluates the precision of n-grams in the generated text that match the reference text, applying a brevity penalty for overly short responses. It helps quantify the match between the model's answers and the expected responses. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used to evaluate the recall of essential content in model responses,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

ensuring no critical information is omitted [21]. It is used in text generation tasks and evaluates the recall of n-grams and word sequences between generated and reference texts. It includes measures such as ROUGE-N, ROUGE-L, and ROUGE-Lsum among others. Each of these measures captures different aspects of the text, such as n-gram overlap or longest common subsequence. METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric that balances precision and recall, providing a nuanced evaluation of linguistic and semantic quality [22]. It uses synonymy and paraphrase matching to align generated text with reference text, allowing for accurate, paraphrased responses. This makes it ideal for dynamic Q&A interactions and tasks requiring semantic accuracy and paraphrasing abilities. This metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features that are not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. BERTScore uses BERT's contextual embeddings to evaluate the semantic similarity between generated and reference texts [23]. This is crucial in Q&A tasks, as the correctness of an answer often depends on its semantic rather than surface-form similarity to the reference answers. It goes beyond word overlap, computing cosine similarity between token embeddings for a nuanced quality assessment. This is key for tasks requiring understanding of context, nuance, and paraphrasing. It has been shown to correlate with human judgment on sentence-level and system-level evaluation.

In the context of our Q&A models, these metrics collectively offer a robust framework for evaluation. The combined use of these metrics allows us to comprehensively evaluate the models not only for accuracy but also for their ability to interact naturally and informatively, mimicking human-like performance in answering queries. Evaluation results of the models are shown in Fig. 7.

It has been observed that all Chatbots can provide accurate answers to general information about the platform. However, the answers given by the Mistral model confirm the platform's contents. All chatbots were able to produce understandable answers to questions requiring course-specific details, but Phi-2 and Mistral 7B models were able to come closer to the most accurate answer than Llama-2. Three chatbots could not reach the expected success in the questions about the number of courses and durations asked about the courses, but the closest answers were obtained with the Mistral model. It seems that it is necessary to improve the training data for the questions in which all chatbots fail.

In Table 1, we asked the same questions to the 3 models we trained with the same dataset and ChatGPT3.5. Among the answers of the models we trained, the correct answers were marked green, those close to the correct answer were marked yellow, and the wrong answers were marked orange. ChatGPT's answers wer e used as an evaluation criterion. As can be seen from the table, while the most accurate answers were obtained with Mistral 7B, even close answers could not be obtained with Llama-2.

When BLEU scores are evaluated, it has been observed that Phi-2 model has the highest score among all three with 0.012. But all scores are very low across all models. This suggests that the generated answers have limited overlap with the reference answers at the word level. Low BLEU scores in a Q&A context might indicate that while the answers are semantically correct, they use different wording or structure compared to the reference answers. Additionally, when the BLEU precision values are examined, it has been observed that, Unigram Precision is substantially higher compared to other metrics, the highest is 0.555 from Phi-2 model. These scores suggest that at the unigram (single word) level, there's a decent match between the generated and reference answers. High unigram precision indicates that key terms are being correctly used in the answers. Therewithal, there is a significant drop in precision values from Unigram to 4-gram, indicating a significant divergence from the reference text as the n-gram size increases. It suggests that while the model can recall relevant terms, it struggles with maintaining coherent longer phrases that match the reference. For the ROUGE Scores, Phi-2 model provided the highest scores on all ROUGE scores. 0.326 for ROUGE-1, 0.118 for ROUGE-2, 0.214 for ROUGE-L, and 0.271 for ROUGE-Lsum. These scores indicate better recall of important words and phrases.

Table 1: Evaluation Metric Results

Metrics	Mistral-7B	Phi-2	Llama 2
BLEU	0.010	0.012	0.007
Unigram Precision	0.525	0.555	0.547
Bigram Precision	0.131	0.159	0.143
Trigram Precision	0.042	0.040	0.057
4-gram Precision	0.010	0.015	0.028
ROUGE-1	0.303	0.326	0.280
ROUGE-2	0.107	0.118	0.091
ROUGE-L	0.200	0.214	0.194
Rouge-Lsum	0.243	0.271	0.243
METEOR	0.162	0.184	0.140
BERTScore F1	0.870	0.873	0.869

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

Table 2: Sample Results

Question	ChatGPT 3.5	Mistral	Llama	Phi-2
What are the distinguishing characteristics of HCIP, HCIA, and HCIE certifications within Huawei's certification system, and how do they cater to different levels of expertise in Huawei technologies?	Huawei's certification system offers various levels of certification to cater to different levels of expertise in Huawei technologies. The three main certifications are HCIP (Huawei Certified ICT Professional), HCIA (Huawei Certified ICT Associate), and HCIE (Huawei Certified ICT Expert).	HCIP focuses on advanced skills, HCIA covers fundamental concepts, and HCIE demonstrates expert-level proficiency. They cater to varying levels of expertise, providing individuals with tailored learning opportunities based on their career goals and current knowledge base.	The HCIP level is the most advanced level of Huawei certification and is designed for advanced-level specialists. The HCIP certification is for advanced-level specialists with significant experience in Huawei technologies and is considered the pinnacle of Huawei certification.	HCIP offers advanced specialization, HCIA focuses on foundational knowledge, and HCIE demonstrates expert-level mastery in Huawei technologies. These certifications cater to diverse skill levels, ensuring accessibility for beginners while providing rigorous standards for experienced professionals.
Can you provide an overview of the role of the course satisfaction survey in the "Introduction to 5G Knowledge," "5G Network and Applications," and "Introduction to Computer Networks and the Internet" courses?	Certainly! The course satisfaction survey plays a crucial role in evaluating the effectiveness of the courses "Introduction to 5G Knowledge," "5G Network and Applications," and "Introduction to Computer Networks and the Internet."	Yes, I can. In the 'Introduction to 5G Knowledge', '5G Network and Applications', and 'Introduction to Computer Networks and the Internet' courses, the course satisfaction survey serves as a feedback mechanism for students to evaluate their learning experience. It allows instructors to gather insights into students' perceptions of the course content, delivery methodology, and overall satisfaction with the course structure.	An overview of the exam structures for HCIP, HCIA, and HCIE certifications is provided, including differences in exam formats and content coverage. This helps students understand the different levels of certification and the requirements for each.	In the 'Introduction to 5G Knowledge' course, the satisfaction survey helps instructors understand students' understanding of key concepts, their level of engagement with the material, and suggestions for improving the course content. In the '5G Network and Applications' course, the survey assists instructors in evaluating the effectiveness of the curriculum and identifying areas for improvement. In the 'Introduction to Computer Networks and the Internet' course, the survey provides insights into students' comprehension of network principles and practical skills acquired throughout the course.
Describe the primary focus of the "5G Network and Applications" course.	The "5G Network and Applications" course focuses on understanding 5G technology, its architecture, protocols, standards, and applications.	The primary focus of the '5G Network and Applications' course is to explore the architecture, features, and applications of fifth-generation (5G) wireless networks, along with their impact on various industries. It provides insights into the advancements in telecommunications	The primary focus of the '5G Network and Applications' course is to provide participants with a comprehensive understanding of the key technologies, architectures, and use cases driving the evolution of 5G networks, including	The primary focus of the '5G Network and Applications' course is to provide an in-depth understanding of the technical aspects, standards, and applications related to 5G networks. It covers topics such as network architecture, deployment models, spectrum bands, core network

		technology and how 5G networks are revolutionizing different sectors.	5G core, edge computing, network slicing, and IoT integration. By the end of the course, participants will be equipped with the knowledge and skills necessary to design, deploy, and manage 5G networks effectively.	technologies, edge computing, and use cases for 5G technology. Through comprehensive theoretical sessions and hands-on exercises, students gain insights into designing, deploying, and managing 5G systems.
How many experimental modules are included in the "Search and Artificial Intelligence" course structure, and what do they involve?	The "Search and Artificial Intelligence" course structure typically includes several experimental modules. These modules are designed to provide hands-on experience and practical application of concepts related to search algorithms, artificial intelligence (AI), and related technologies.	There are a total of 4 experimental modules included in the 'Search and Artificial Intelligence' course structure. These modules provide hands-on experiments or simulations aimed at reinforcing theoretical concepts and developing practical skills related to search algorithms and artificial intelligence applications. Participants engage in these experiential activities to deepen their understanding and apply knowledge gained throughout the course.	The 'Search and Artificial Intelligence' course structure includes one experimental module on search algorithms and another on AI basics. These modules provide hands-on experience with key concepts and techniques in search and AI.	The 'Search and Artificial Intelligence' course includes 2 experimental modules designed for students interested in delving deeper into specific aspects of search algorithms or AI technologies. These modules provide hands-on activities or projects to reinforce learning and explore advanced topics related to search and artificial intelligence.
Can you explain the balance between theory and hands-on practice in the "Internet of Things and Applications" course?	In an "Internet of Things and Applications" course, theory provides foundational knowledge of IoT concepts, while hands-on practice allows students to apply this knowledge practically. The balance ensures students understand both the theoretical underpinnings and practical implementation of IoT systems.	Yes, absolutely! In the 'Internet of Things and Applications' course, we strike a balance between theoretical sessions and hands-on practice to ensure students gain comprehensive knowledge and practical skills. The theoretical sessions cover IoT fundamentals, networking protocols, embedded systems, cloud computing, and IoT applications. Hands-on practice includes lab sessions where students engage in activities such as device-cloud integration, sensor networks, edge	The 'Internet of Things and Applications' course balances theory and hands-on practice by providing a theoretical foundation in IoT concepts and technologies, along with practical exercises and projects that allow students to apply their knowledge in real-world scenarios. This approach ensures students gain both theoretical understanding and practical skills necessary for	The balance between theory and hands-on practice in the 'Internet of Things and Applications' course ensures that students gain a solid theoretical understanding of IoT concepts and technologies while also developing practical skills through lab assignments, projects, and simulations. This approach prepares students for real-world challenges by equipping them with both knowledge and practical abilities.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

		computing deployments, and IoT solution development using development.	developing and deploying IoT solutions.	
What is the duration and structure of the "Internet of Things and Applications" course, and are there any recommended prerequisites?	The "Internet of Things and Applications" course typically covers IoT concepts, architecture, connectivity technologies, data analytics, security, applications, development, and a final project. Prerequisites may include basic knowledge of computer networks, programming languages, electronics, and data analytics.	The 'Internet of Things and Applications' course spans a total of 64 lessons, with 32 theory sessions and 32 hands-on practice sessions. It is designed for students majoring in computer applications and related majors, providing them with essential skills and knowledge in IoT development.	The duration of the 'Internet of Things and Applications' course may vary depending on the institution or platform offering it. It may span several weeks to a semester, covering fundamental concepts, practical skills, and advanced topics related to IoT development and deployment.	The 'Internet of Things and Applications' course spans a specified duration with a combination of theory sessions and hands-on practice modules. It may be offered as a traditional classroom-based course or delivered online with interactive components. Recommended prerequisites may include a basic understanding of networking concepts, programming skills.

METEOR scores indicate moderate performance. This metric accounts for synonymy and paraphrasing, suggesting your model does fairly well at capturing meaning, even if exact wording does not always match. Phi-2 model comes on top with 0.184 among all three.

F1 scores from BERTScore are relatively high compared to other metrics, with all models performing closely, and the highest is 0.873 from Phi-2. BERTScore evaluates the contextual embedding similarity between the generated and reference texts, indicating that semantically, the model's answers are quite close to the desired answers. This indicates that the semantic quality of the model outputs is relatively strong, with good contextual alignment with the reference answers.

V. CONCLUSION

In this paper, we discuss the development and implementation of a small-scale learning and career guidance chatbot using public data from Huawei's ICT Learning Catalog Courses. For the fine-tuning process of the selected models, a q&a pair dataset was generated using GPT-3 and external information from catalog documents regarding course details and software-related categories. We used small-scale models such as 2.7B parameters for phi, 7B for Mistral, and 8B for Llama to tackle computational challenges, and employed QLoRa to reduce trainable parameters for the same reasons. Our study demonstrates the analysis of the usage of s.o.t.a. language models with limited data in a specific domain. For the evaluation part of our study, human evaluation was conducted for the generated results and also compared with the GPT-4 responses using metrics such as BLEU, ROUGE, METEOR, and BERTScore. The sample results for the responses show that acceptable accuracies could be achieved with limited processing power and a small scale of data for customized chatbots. This highlights the potential of small-sized models like Phi-2 compared to other larger models. However, a detailed analysis of the safety, toxicity, or validity of the responses of fine-tuned models is not within the scope of this study. These analyses and post-training alignment of the models could be considered as future work.

REFERENCES

[1]. T.F. Tan, K. Elangovan, L. Jin, Y. Jie, L. Yong, J. Lim, S. Poh, W.Y. Ng, D. Lim, Y. Ke, N. Liu, D.S.W. Ting, "Fine-tuning Large Language Model (LLM) Artificial Intelligence Chatbots in Ophthalmology and LLM-based evaluation using GPT-4," arXiv preprint arXiv:2402.10083, 2024.

[2]. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, ... and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.

[3]. Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y.T. Lee, "Textbooks are all you need ii: phi-1.5 technical report," arXiv preprint arXiv:2309.05463, 2023.

[4]. A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.D.L. Casas, ... and W.E. Sayed, "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY1600

[5]. Hugging Face – The AI community building the future. (n.d.). https://huggingface.co/

[6]. F. Khennouche, Y. Elmir,, Y. Himeur, N. Djebari, A. Amira, "Revolutionizing generative pre-traineds: Insights and challenges in deploying ChatGPT and generative chatbots for FAQs." Expert Systems with Applications, 246, 123224, 2024.

[7]. M. Jovanović, K. Kuk, V. Stojanović, and E. Mehić, "Chatbot Application as Support Tool for the Learning Process of Basic Concepts of Telecommunications and Wireless Networks," Facta Universitatis, Series: Automatic Control and Robotics, 22(2), 2024, pp. 057-073.

[8]. S. Balakrishnan, P. Jayanth, S. Parvathynathan, and R. Sivashankar, "Artificial intelligence-based vociferation chatbot for emergency health assistant," In AIP Conference Proceedings (Vol. 2742, No. 1). AIP Publishing, 2024.

[9]. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, ... and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," Advances in Neural Information Processing Systems, 33, 2020, pp. 9459-9474.

[10]. Huawei Talent. (n.d.). https://e.huawei.com/en/talent/portal/#/

[11]. Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, ... and D. Xiong, "Evaluating large language models: A comprehensive survey," arXiv preprint arXiv:2310.19736, 2023.

[12]. H. Naveed, A.U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, ... and A. Mian, "A comprehensive overview of large language models," arXiv preprint arXiv:2307.06435, 2023.

[13]. H.A. Alawwad, A. Alhothali, U. Naseem, A. Alkhathlan, and A. Jamal, "Enhancing Textbook Question Answering Task with Large Language Models and Retrieval Augmented Generation," arXiv preprint arXiv:2402.05128, 2024.

[14]. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRa: Efficient finetuning of quantized llms," Advances in Neural Information Processing Systems, 36, 2024.

[15]. J.C. Chow, L. Sanders, and K. Li, "Design of an educational chatbot using artificial intelligence in radiotherapy," AI, 4(1), 2023, pp. 319-332.

[16]. N. Ghorashi, A. Ismail, P. Ghosh, A. Sidawy, R. Javan, and N.S. Ghorashi, "AI-powered chatbots in medical education: potential applications and implications," Cureus, 15(8), 2023.

[17]. J. Wang, J. Macina, N. Daheim, S.P. Chowdhury, and M. Sachan, "Book2Dial: Generating Teacher-Student Interactions from Textbooks for Cost-Effective Development of Educational Chatbots," arXiv preprint arXiv:2403.03307, 2024.

[18]. A. Bandi, and H. Kagitha, "A Case Study on the Generative AI Project Life Cycle Using Large Language Models," Proceedings of 39th International Confer, 98, 2024, pp. 189-199.

[19]. A. Chen, G. Stanovsky, S. Singh, and M. Gardner, “Evaluating question answering evaluation,” in Proceedings of the 2nd workshop on machine reading for question answering, 2019, pp. 119–124.

[20]. K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. “BLEU: a method for automatic evaluation of machine translation,” In Proc. 40th Annual Meeting of the Association for Computational Linguistics (Association of Computational Machinery), 2002, pp. 311–318.

[21]. L. Chin-Yew, “ROUGE: A package for automatic evaluation of summaries,” In Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81.

[22]. S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005.

[23]. T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” In International Conference on Learning Representations, 2020.