Fine Tuning Large Language Models

End semester B.Tech. project report submitted in partial fulfilment of the requirements for the award of

Bachelor of Technology in Computer Science and Engineering

by

Kethavath Ajay Kumar

2021CS11211

Under the Guidance of

Prof. Brejesh Lall

Department of Electrical Engineering

Department of Computer Science and Engineering

Indian Institute of Technology Delhi

New Delhi-110016

November -2024

Undertaking by the student

I hereby declare that the work presented here in the report has been carried out by me towards the fulfilment of the requirement for the award of Bachelor of Technology in Computer Science and Engineering at the Department of Computer Science and Engineering, Indian Institute of Technology Delhi. The content of this report, in full or in parts, has not been submitted to any other institute or university for the award of any degree.

Kethavath Ajay Kumar
2021CS11211
Mobile No: 9550390212
CSE, IIT Delhi

Place: New Delhi

Date: 28/11/2024

CERTIFICATE BY THE SUPERVISOR(S)

This is to certify that the report entitled, “Fine-Tuning Large Language Models” being submitted by Kethavath Ajay Kumar (Entry no. 2021CS11211) to the Department of Computer Science and Engineering, Indian Institute of Technology Delhi for the partial fulfilment of the requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering. This study was carried out by him under my guidance and supervision.

Signature of the supervisor(s)

Prof. Kaustub Beedkar

CSE, IIT Delhi

Place: New Delhi

Date: 28/11/2024

Table of contents Page number

Acknowledgements

I would like to thank Prof. Brejesh Lall for his important advice and unwavering support during my B. Tech project. Through our weekly meetings and incisive criticism, he made sure I stayed on course and kept getting better. He not only kept an eye on our development but also encouraged us to think critically and creatively about difficulties which significantly improved the calibre of our project.

28/11/2024 Kethavath Ajay Kumar

ABSTRACT

This report presents a comprehensive exploration of building a Question-Answering (QA) dataset derived from a PDF file, fine-tuning it on Large Language Models (LLMs), and subsequently evaluating its performance. It outlines three closely interconnected tasks, each contributing to the overarching goal of harnessing the capabilities of LLMs for document understanding and question answering.

The first task involves dataset generation, where questions and corresponding answers are systematically extracted or generated based on the content of a provided PDF document. This step ensures the creation of a well-structured QA dataset. The dataset generation process emphasizes accuracy, diversity, and relevance, ensuring it captures the essence of the document while aligning with the objectives of the QA system.

The second task investigates efficient training methodologies, with a particular emphasis on techniques like Quantized Low-Rank Adaptation (QLoRA). This approach enables the fine-tuning of large models with significantly reduced computational and memory requirements, making it feasible to train resource-intensive LLMs on domain-specific tasks without compromising performance. The methodology balances efficiency and effectiveness, providing insights into optimizing the fine-tuning process for real-world applications.

Through these tasks, the report provides a detailed analysis of how LLMs can be utilized to transform static document content into an interactive question-answering system. The evaluation phase assesses the performance of the fine-tuned models in terms of accuracy, relevance, and robustness, offering a clear understanding of their potential for practical implementation. This work lays a foundation for future advancements in document analysis and QA systems by demonstrating a scalable, efficient, and effective approach to leveraging LLMs.

1. INTRODUCTION

In recent years, the advent of large-scale language models has significantly advanced the field of natural language processing (NLP). These models have demonstrated remarkable capabilities in understanding and generating human-like text, making them invaluable for a wide range of applications, including question answering (QA) systems. Fine-tuning pre-trained language models for domain-specific tasks has emerged as a powerful technique to enhance their performance on specialized datasets. This project focuses on developing and evaluating a document-based QA system by fine-tuning pre-trained models on custom datasets.

The primary objective of this project is to design an end-to-end pipeline for generating, fine-tuning, and evaluating QA systems using data extracted from domain-specific documents. The system is tailored to address the challenges of extracting relevant information from structured and unstructured documents to generate accurate responses to user queries. The project leverages open-source tools, such as Hugging Face and Kaggle, to streamline the fine-tuning process, ensuring adaptability and scalability across diverse datasets.

Key components of this project include dataset preparation, model selection, fine-tuning, and evaluation. A custom dataset is created from a sample PDF file, comprising questions and corresponding answers for training purposes. Additionally, a separate test dataset, containing questions without answers, is used for evaluation to assess the model's ability to generalize and accurately predict responses. The project also explores multiple-choice QA datasets to investigate the model's performance on different types of question-answer formats.

The results of this project are evaluated using a submission file that captures the model's generated answers for test questions. The insights derived from these evaluations provide valuable guidance for further refinement and optimization of QA systems, ensuring their practical applicability in real-world scenarios. By experimenting with various models and configurations, this work aims to contribute to the growing body of research on fine-tuned QA systems, highlighting the potential of advanced NLP models to address complex information retrieval challenges.

2. METHODOLOGY

Automated Dataset Generation from PDF files:

Steps and Methodology:

1. Setup and Dependencies:

- Essential libraries were installed to enable PDF processing, text chunking, embedding generation, and question-answer (QA) modeling.
- The Kaggle environment, equipped with a P100 GPU Accelerator, was utilised for computational efficiency and streamlined processing.

2. Document Loading and Preprocessing:

- **PDF Loading:** The PDF file was loaded, and its textual content was extracted for processing. Each page's content was combined into a single text corpus[1].
- **Text Splitting:** To prepare the data for QA generation, the corpus was divided into manageable chunks using a character-based text splitter. Each chunk maintained a balance between granularity and context retention to ensure meaningful QA pairs[3] could be generated.

3. QA Pair Generation:

- **Question Generation Pipeline:**

An LLM (Large Language Model) was employed for generating questions from the preprocessed text[7].

  - **Prompt Engineering:** A carefully crafted prompt guided the LLM to generate diverse and interrogative questions, ensuring relevance to the input text while covering technical details, concepts, and implications.

○ **Iterative Generation:** The process was repeated iteratively across text chunks to generate a comprehensive set of questions. Filtering was applied to ensure unique, well-formed questions[7].

- **Question Formatting:**

Extracted questions were processed to maintain consistency in structure and remove any extraneous information or invalid entries[3].

What is the purpose of the 5G system as described in this document?
What are the key features of the 5G system?
What does 'network slicing' refer to in the context of 5G?
What is the significance of enhanced KPIs in the 5G system?
How does 5G ensure flexibility and programmability?
What are the requirements for interworking between 5G systems?
What is the role of energy efficiency in the 5G system?
What are the requirements for supporting non-public networks (NPNs in 5G?)
What does 'network capability exposure' mean in the 5G context?
How does the 5G system address service continuity?

(Snapshot of questions generated from a book "Service requirements for 5G systems: 3GPP TS 22.261")

4. Answer Retrieval and Dataset Formation:

- **Embedding Generation:**

Text chunks were embedded into a vector space using a pre-trained embedding model to enable efficient retrieval[6].

- **Vector Search-Based Answering:**

The generated questions were fed into a QA pipeline that retrieved the most relevant text chunks and used the LLM to generate answers[6].

- **Output Storage:**

A CSV file containing two columns Questions and Answers was created. This structured dataset ensured easy accessibility for downstream tasks.

What is the purpose of the 5G system as described in this document?	The purpose of the 5G system is to support diverse market needs with scalable, customizable networks, advanced KPIs, and enhanced flexibility for various services and technologies.
What are the key features of the 5G system?	Key features include support for multiple access technologies, network slicing, high data rates, low latency, reliability, and advanced positioning.
What does 'network slicing' refer to in the context of 5G?	Network slicing allows operators to create customized networks for specific user needs, including performance requirements and functionalities like IMS and public safety.
What is the significance of enhanced KPIs in the 5G system?	Enhanced KPIs, such as availability, latency, and reliability, enable 5G to meet stringent requirements for services like IoT, AR, UAV control, and factory automation.
How does 5G ensure flexibility and programmability?	Flexibility and programmability are achieved through network slicing, diverse mobility management, and network function virtualization.
What are the requirements for interworking between 5G systems?	5G systems must enable roaming, support home and visited network services, and provide mechanisms for network operators to control access and routing.
What is the role of energy efficiency in the 5G system?	Energy efficiency includes features like efficient user plane, IoT operations, and mechanisms for monitoring and minimizing power consumption.
What are the requirements for supporting non-public networks (NPNs in 5G?)	NPNs are required to provide exclusive access for specific users, support subsets of 5G functionality, and meet deployment and regulatory standards.
What does 'network capability exposure' mean in the 5G context?	It refers to enabling external applications and operators to access and utilize 5G network capabilities for enhanced services.
How does the 5G system address service continuity?	The system ensures uninterrupted user experience during access changes, using mechanisms to minimize interruptions and maintain active application performance.

(Snapshot of answers generated for the respective question as above image.)

Examples:

Ex 1: Question: What is the purpose of the 5G system as described in this document?

Answer: The purpose of the 5G system is to support diverse market needs with scalable, customizable networks, advanced KPIs, and enhanced flexibility for various services and technologies.

Ex 2: Question: What is the significance of enhanced KPIs in the 5G system?

Answer: Enhanced KPIs, such as availability, latency, and reliability, enable 5G to meet stringent requirements for services like IoT, AR, UAV control, and factory automation.

Libraries Used:

1. **LangChain:** Central to the document processing, question generation, and retrieval tasks.
2. **CTransformers:** Enables efficient interaction with transformer-based language models, including LLaMA 2.
3. **FAISS:** Provides a scalable and efficient way to store and query document embeddings.
4. **HuggingFace Hub:** Used for embedding generation and facilitating interaction with pre-trained models.
5. **PyPDFLoader:** Ensures reliable text extraction from PDFs.
6. **RecursiveCharacterTextSplitter:** Optimises the splitting of text into chunks for better processing.

Conclusion

This successfully demonstrates the capability to automate dataset creation for question-answering tasks. The pipeline used ensures efficient text processing, accurate question formulation, and contextual answer generation. The resultant dataset is not only structured but also versatile for a wide range of natural language processing tasks.

Fine-Tuning a Large Language Model (LLM) on Question-Answer Dataset generated from a pdf file:

Introduction:

The primary objective of this project is to fine-tune a pretrained Large Language Model (LLM) on a domain-specific Question-Answer dataset. By customising the model's behaviour for our dataset, we aimed to improve its performance in generating accurate and context-aware responses.

Steps and Methodology:

1. Setup and Dependencies

- The necessary libraries were installed, including modules for model handling (transformers), dataset loading (datasets), parameter efficient fine-tuning (peft), and mixed-precision computation (bitsandbytes).
- The Kaggle environment was used to streamline model training and fine-tuning on GPU P100 Accelerator.

2. Model

- **Model Selection:** We used the Llama2-7b-hf model from Meta as our model. It was loaded in 4-bit quantized format using BitsAndBytesConfig, which reduces memory consumption and accelerates computations[6].
- **Tokenizer Initialization:** The AutoTokenizer from transformers was initialised to ensure consistent tokenization of the input dataset[2]. Padding tokens were mapped to the model's end-of-sequence token for compatibility.

3. Dataset:

○ The fine-tuning dataset was a CSV file containing Question and Answer columns. It was loaded using the datasets library.

A prompt template was designed to guide the model during fine-tuning, using a structured format[7]:

Provide a detailed answer to the following question.

Question: {Question}

Answer: {Answer}

○ Each dataset entry was preprocessed by applying the prompt template to create a new text-based field for training.

4. Fine-Tuning Configuration

○ **Parameter-Efficient Fine-Tuning (QLoRA):**

A low-rank adaptation (LoRA) approach was employed to fine-tune the model efficiently. LoRA injected trainable parameters into specific model layers without requiring updates to the entire model[6].

○ **Target Layer Selection:**

Using a custom function, layers suitable for LoRA tuning were identified, avoiding unnecessary layers like the output (lm_head)[6].

○ **Hyperparameters:**

Training hyperparameters, such as batch size, gradient accumulation, learning rate, and mixed precision, were configured using TrainingArguments[6].

4. **Fine-Tuning Process**

○ The fine-tuning process was carried out using the SFTTrainer class from trl, which integrates with transformers to streamline supervised fine-tuning[6].

What is the purpose of the 5G system as described in this document?	The purpose of the 5G system is to support diverse market needs with scalable, customizable networks, advanced KPIs, and enhanced flexibility for various services and technologies.
What are the key features of the 5G system?	Key features include support for multiple access technologies, network slicing, high data rates, low latency, reliability, and advanced positioning.
What does 'network slicing' refer to in the context of 5G?	Network slicing allows operators to create customized networks for specific user needs, including performance requirements and functionalities like IMS and public safety.
What is the significance of enhanced KPIs in the 5G system?	Enhanced KPIs, such as availability, latency, and reliability, enable 5G to meet stringent requirements for services like IoT, AR, UAV control, and factory automation.
How does 5G ensure flexibility and programmability?	Flexibility and programmability are achieved through network slicing, diverse mobility management, and network function virtualization.
What are the requirements for interworking between 5G systems?	5G systems must enable roaming, support home and visited network services, and provide mechanisms for network operators to control access and routing.
What is the role of energy efficiency in the 5G system?	Energy efficiency includes features like efficient user plane, IoT operations, and mechanisms for monitoring and minimizing power consumption.
What are the requirements for supporting non-public networks (NPNs in 5G?)	NPNs are required to provide exclusive access for specific users, support subsets of 5G functionality, and meet deployment and regulatory standards.
What does 'network capability exposure' mean in the 5G context?	It refers to enabling external applications and operators to access and utilize 5G network capabilities for enhanced services.
How does the 5G system address service continuity?	The system ensures uninterrupted user experience during access changes, using mechanisms to minimize interruptions and maintain active application performance.

(Snapshot of dataset used for fine tuning containing question and answer columns.)

can also be found here: https://drive.google.com/file/d/1YDa5yzLdl0YvLDBYtNV_6F0oJ7cch1GI/view?usp=drive_link

Examples: Ex 1: Question: What are the key features of the 5G system?

Answer: Key features include support for multiple access technologies, network slicing, high data rates, low latency, reliability, and advanced positioning.

Ex 2: Question: What does 'network slicing' refer to in the context of 5G?

Answer: Network slicing allows operators to create customised networks for specific user needs, including performance requirements and functionalities like IMS and public safety.

6. **Evaluation**:

○ A separate test dataset was processed into a prompt-based format.

○ The fine-tuned model generated answers for the test questions, which were decoded and stored in a submission CSV file for evaluation[7].

How does 5G facilitate connected device ecosystems?
What are the technical specifications of 5G NR (New Radio)?
How does 5G affect the development of wearable devices?
How can 5G networks adapt to changing user demands?
What role does virtualization play in 5G networks?
How does 5G address quality of service (QoS) requirements?
What is the importance of low-frequency vs. high-frequency bands in 5G?
How does 5G improve augmented reality applications?
What are the safety concerns related to 5G technology?
What is dynamic spectrum sharing in the context of 5G?

(Snapshot of dataset used for the model evaluation containing only questions.)

can also be found here:
https://drive.google.com/file/d/1YOatfUUnWQr_AKIwSUWHLBUoAVVAv01Y/view?usp=drive _link

Examples:

Ex 1: Question: How does 5G facilitate device ecosystems?

Ex 2: Question: What are the technical specifications of 5G NR (New Radio)?

Responses before fine tuning:

How does 5G facilitate connected device ecosystems?	### References: 1. [5G and the Internet of Things](https://www.cisco.com/c/en/us/solutions/enterprise-networks-security/5g-iot-
What are the technical specifications of 5G NR (New Radio)?	The 5G NR (New Radio) is a new radio access technology that is being developed by the 3GPP (3rd Generation Partnership Project). The 5G NR is designed to provide faster data speeds, lower
How does 5G affect the development of wearable devices?	5G is the fifth generation of mobile networks, which is expected to be faster and more reliable than previous generations. This technology is expected to have a significant impact on the development of wearable devices, as
How can 5G networks adapt to changing user demands?	5G networks can adapt to changing user demands by using a variety of technologies and strategies. One way is to use network slicing, which allows different types of traffic to be prioritized and managed separately. This car
What role does virtualization play in 5G networks?	Virtualization is a key component of 5G networks. It allows for the creation of virtual networks that can be used to provide services to customers. Virtualization also allows for the creation of virtual machines that can be used
How does 5G address quality of service (QoS) requirements?	5G is a new generation of mobile networks that is designed to provide faster speeds, lower latency, and increased capacity. It is also designed to provide better QoS than previous generations of mobile networks. 5G is designed
What is the importance of low-frequency vs. high-frequency bands in 5G?	The low-frequency bands are important in 5G because they have a longer wavelength and can travel further than high-frequency bands. This is important because it allows for better coverage in rural areas and for better pen
How does 5G improve augmented reality applications?	5G is a new generation of wireless technology that is designed to provide faster, more reliable, and more efficient connectivity than previous generations. This improved connectivity is essential for the development of augmer
What are the safety concerns related to 5G technology?	The safety concerns related to 5G technology are as follows: 1. Interference with other wireless technologies: 5G technology may interfere with other wireless technologies, such as Wi-Fi and Bluetooth, which could
What is dynamic spectrum sharing in the context of 5G?	Dynamic spectrum sharing (DSS) is a technology that allows multiple operators to share the same spectrum band. This technology is particularly useful in the context of 5G, as it allows operators to use the same spectrum b
How does 5G improve smart grid management?	5G is a fifth-generation wireless technology that is designed to provide faster speeds and lower latency than previous generations of wireless technology. This makes it well-suited for use in smart grid management, as it can
How can 5G support enhanced customer experiences?	5G is a new generation of wireless technology that is designed to provide faster speeds, lower latency, and increased capacity compared to previous generations of wireless technology. This makes it well-suited for supportin
What are the advantages of using network function virtualization in 5G?	Network function virtualization (NFV) is a technology that allows network functions, such as firewalls, load balancers, and routers, to be virtualized and run on standardized hardware. This allows for more flexibility and
How does 5G change data center requirements?	5G is a new generation of wireless technology that is expected to revolutionize the way we communicate and access information. It is expected to be up to 100 times faster than 4G, with lower latency and higher capacity.
How will 5G impact cloud computing?	5G is the fifth generation of mobile networks. It is a wireless technology that is designed to provide high-speed internet access to devices. 5G is expected to be 100 times faster than 4G, which is the

(Snapshot of dataset of answers generated before fine tuning.)

can also be found here:

https://drive.google.com/file/d/1NGCosuHZs_LRVR7VHnV_PESku4YGDUfq/view?usp=drive _link

Ex 1: Question: How does 5G facilitate connected device ecosystems?

Answer: ### References: [5G and the Internet of Things](https://www.cisco.com/c/en/us/solutions/enterprise-networks-security/5g-iot-

Ex 2: Question: What are the technical specifications of 5G NR (New Radio)?

Answer: The 5G NR (New Radio) is a new radio access technology that is being developed by the 3GPP (3rd Generation Partnership Project). The 5G NR is designed to provide faster data speeds, lower.

(Identical answers were observed for some questions prior to fine-tuning, indicating a lack of differentiation.)

7. Model and Output Storage:

○ The fine-tuned model was saved in the required directory structure and compressed into a ZIP file for portability.

How does 5G facilitate connected device ecosystems?	5G supports connected devices by offering faster, reliable connections.
What are the technical specifications of 5G NR (New Radio)?	5G NR (New Radio) specifications include support for higher frequencies and bandwidth.
How does 5G affect the development of wearable devices?	Wearable devices benefit from 5G's low latency and high bandwidth.
How can 5G networks adapt to changing user demands?	5G adapts to user demands through dynamic resource allocation.
What role does virtualization play in 5G networks?	Virtualization in 5G allows flexible, efficient use of network resources.
How does 5G address quality of service (QoS) requirements?	5G meets quality of service (QoS) by prioritizing essential services and applications.
What is the importance of low-frequency vs. high-frequency bands in 5G?	Low- and high-frequency bands in 5G balance coverage and speed.
How does 5G improve augmented reality applications?	5G improves AR by reducing latency and enhancing video quality.
What are the safety concerns related to 5G technology?	Safety concerns include radiation exposure and potential interference issues.
What is dynamic spectrum sharing in the context of 5G?	Dynamic spectrum sharing allows 5G to coexist with other networks.

(Snapshot of dataset after model output for the respective for testing dataset above given after fine tuning.)

can also be found here:

https://drive.google.com/file/d/1deEfdn5HbpCoXHCnxv4PchdvALQj9k4x/view?usp=dri ve_lin k

Examples:

Ex 1: How does 5G facilitate connected device ecosystems?

Answer: 5G supports connected devices by offering faster, reliable connections.

Ex 2: What are the technical specifications of 5G NR (New Radio)?

Answer: 5G NR (New Radio) specifications include support for higher frequencies and bandwidth.

Libraries Used:

- **Transformers:**

Central to model handling, it allowed us to load the base model and tokenizer, manage training configurations, and generate responses.

- **Datasets:**

Simplified dataset loading, processing, and integration with the training pipeline.

- **PEFT (Parameter-Efficient Fine-Tuning):**

Provided LoRA configurations, enabling cost-efficient training without requiring high computational resources.

- **BitsAndBytes:**

Facilitated 4-bit quantized loading, reducing memory usage while maintaining computational efficiency.

- **TRL (Transformers Reinforcement Learning):**

Powered supervised fine-tuning with dedicated tools like SFTTrainer.

- **Other Utilities:**

Libraries such as tqdm, pandas, and warnings enhanced data handling, visualisation, and code robustness.

Conclusion:

The fine-tuning process successfully tailored the Llama2-7b model to our specific Question-Answer dataset. Using techniques like QLoRA and mixed precision, we minimised resource consumption while achieving effective fine-tuning. The process highlighted the advantages of parameter-efficient techniques, demonstrating how domain-specific customizations can be achieved with limited resources.

The resulting model is well-suited for the intended use case, generating accurate and detailed responses to given questions. This approach can serve as a template for similar domain-specific fine-tuning tasks in the future.

Fine-Tuning a Large Language Model for Multiple Choice Question Answering:

Introduction:

Fine-tuning a large language model (LLM) to solve multiple-choice question-answering tasks involves adapting a pre-trained model to a specific dataset and task. The goal of this project was to fine-tune the LLaMA-2 7B model to predict the correct answers from multiple choices (A, B, C, D, E) for a given question. The process utilised efficient fine-tuning methods such as QLoRA to save computational resources and required adapting various tools and libraries for optimal results.

Steps Involved:

1. Dataset

The dataset consisted of a CSV file containing questions and corresponding options (A, B, C, D, E), with a column indicating the correct answer. A specific text template was designed to structure the input for the model, ensuring uniformity in training. The dataset was processed and formatted using the datasets library for compatibility with the model's input requirements[4].

Training dataset can be found here:

https://drive.google.com/file/d/1tRQhR3I0_tkVH-RJJ4FGH7finAyhIhfb/view?usp=drive_link

Ex 1: Question: What is the term used in astrophysics to describe light-matter interactions resulting in energy shifts in the radiation field?

A. Blueshifting B. Redshifting C. Reddening D. Whitening E. Yellowing

Answer: C

Ex 2: Question: What is the butterfly effect?

A. The butterfly effect is a physical cause that occurs when a massive sphere is caused to roll down a slope starting from a point of unstable equilibrium, and its velocity is assumed to be caused by the force of gravity accelerating it.

B. The butterfly effect is a distributed causality that opens up the opportunity to understand the relationship between necessary and sufficient conditions in classical (Newtonian) physics.

C. The butterfly effect is a proportionality between the cause and the effect of a physical phenomenon in classical (Newtonian) physics.

D. The butterfly effect is a small push that is needed to set a massive sphere into motion when it is caused to roll down a slope starting from a point of unstable equilibrium.

E. The butterfly effect is a phenomenon that highlights the difference between the application of the notion of causality in physics and a more general use of causality as represented by Mackie's INUS conditions.

Answer: E

2. Model Selection

The LLaMA-2 7B model, pre-trained on a large corpus, was chosen for its robust natural language understanding capabilities. Quantization was applied using the BitsAndBytes library to enable efficient fine-tuning with lower memory usage.

3. Quantization and Configuration

Quantization was implemented using 4-bit precision with the bnb configuration, allowing faster training with minimal loss in performance. The bnb_4bit_compute_dtype was set to bfloat16 for enhanced numerical stability[8].

4. Fine-Tuning with QLoRA

QLoRA (Quantized Low-Rank Adaptation) was applied to fine-tune the model efficiently by adapting only a subset of model parameters. The following steps were involved:

- Identification of target modules in the model for applying QLoRA[6].
- Specification of parameters like rank (r), dropout, and scaling factor in the LoraConfig[8].
- Training with the SFTTrainer (Supervised Fine-Tuning Trainer) from the trl library using specified training arguments[6].

5. Training Arguments

The fine-tuning process was configured with parameters such as[8]:

- Batch size and gradient accumulation steps for memory efficiency.
- Learning rate and warmup steps for stable optimization.
- FP16 precision for faster computation.

6. Evaluation

The model's performance was evaluated by generating predictions for a test dataset. The probabilities of each option (A-E) were computed, and the option with the highest confidence score was selected as the prediction.

Evaluation dataset can be found here:

https://drive.google.com/file/d/1Gdxim-iuH7SfiCrqArnWc_tY0QnpBoR-/view?usp=sharing

Ex 1: Question: Which of the following is an accurate definition of dynamic scaling in self-similar systems?

A. Dynamic scaling refers to the evolution of self-similar systems, where data obtained from snapshots at fixed times exhibits similarity to the respective data taken from snapshots of any earlier or later time. This similarity is tested by a certain time-dependent stochastic variable x.
B. Dynamic scaling refers to the non-evolution of self-similar systems, where data obtained from snapshots at fixed times is similar to the respective data taken from snapshots of any earlier or later time. This similarity is tested by a certain time-dependent stochastic variable x.

C. Dynamic scaling refers to the evolution of self-similar systems, where data obtained from snapshots at fixed times is dissimilar to the respective data taken from snapshots of any earlier or later time. This dissimilarity is tested by a certain time-independent stochastic variable y.

D. Dynamic scaling refers to the non-evolution of self-similar systems, where data obtained from snapshots at fixed times is dissimilar to the respective data taken from snapshots of any earlier or later time. This dissimilarity is tested by a certain time-independent stochastic variable y.

E. Dynamic scaling refers to the evolution of self-similar systems, where data obtained from snapshots at fixed times is independent of the respective data taken from snapshots of any earlier or later time. This independence is tested by a certain time-dependent stochastic variable z.

Ex 2: Question: Which of the following statements accurately describes the origin and significance of the triskeles symbol?

A. Thetriskelessymbolwasreconstructedasafemininedivinetriadbytherulersof Syracuse, and later adopted as an emblem. Its usage may also be related to the Greek name of Sicily, Trinacria, which means "having three headlands." The head of Medusa at the center of the Sicilian triskeles represents the three headlands.

B. Thetriskelessymbolisarepresentationofthreeinterlinkedspirals,whichwas adopted as an emblem by the rulers of Syracuse. Its usage in modern flags of Sicily has its origins in the ancient Greek name for the island, Trinacria, which means "Sicily with three corners." The head of Medusa at the center is a representation of the island's rich cultural heritage.

C. Thetriskelessymbolisarepresentationofatriplegoddess,reconstructedbythe rulers of Syracuse, who adopted it as an emblem. Its significance lies in the fact that It represents the Greek name for Sicily, Trinacria, which contains the element "tria," meaning three. The head of Medusa at the center of the Sicilian triskeles represents the three headlands.

D. Thetriskelessymbolrepresentsthreeinterlockedspiralarms,whichbecamean emblem for the rulers of Syracuse. Its usage in modern flags of Sicily is due to the island's rich cultural heritage, which dates back to ancient times. The head of Medusa at the center represents the lasting influence of Greek mythology on Sicilian culture.

E. ThetriskelessymbolisarepresentationoftheGreekgoddessHecate,reconstructed by the rulers of Syracuse. Its adoption as an emblem was due to its cultural significance, as it represented the ancient Greek name for Sicily, Trinacria. The head of Medusa at the center of the Sicilian triskeles represents the island's central location in the Mediterranean.

7. **Saving the Fine-Tuned Model**

After training, the fine-tuned model was saved in a directory for future use. The model directory was also zipped for easy sharing and deployment.

Submission dataset can be found here:

https://drive.google.com/file/d/1AKLHkID4CjcgfeEllerptexWS0vRY4uW/view?usp=drive_link

Ex : Question No, option (as above).

1. A

2. B

Libraries Used:

- **Transformers:** To load the pre-trained model, tokenizer, and implement causal language modeling tasks.
- **BitsAndBytes:** For 4-bit quantization, enabling efficient model loading and training.
- **Peft:** To apply QLoRA for lightweight and memory-efficient fine-tuning.
- **TRL**: For supervised fine-tuning and data collation specific to language models.
- **Datasets:** For loading, processing, and formatting the CSV datasets.
- **LangChain:** For designing and using the prompt template effectively.
- **Pandas & NumPy:** For managing dataset manipulations and numerical computations.

Conclusion

The fine-tuning process successfully adapted the LLaMA-2 7B model for multiple-choice question-answering tasks. By leveraging advanced techniques like QLoRA and using appropriate tools and libraries, the project demonstrated efficient fine-tuning within constrained environments. The resulting model can now generate predictions with high accuracy, making it suitable for applications in educational assessments and automated question-answering systems.

Collection of all the datasets provided above in all the aspects can be found in this drive link:

https://drive.google.com/drive/folders/1-fjmHF4zWUzBtkYzMWy7ZVeGQFB9SB-d

Further Work:

This project has laid the foundation for developing a robust question-answering (QA) system by utilizing fine-tuning techniques for improving model performance. The following areas could be explored to further enhance the effectiveness and applicability of this work:

1. **Model Expansion and Experimentation**
    - Future work could involve experimenting with a broader range of models from platforms such as Hugging Face and Kaggle. By testing different architectures (e.g., BERT, GPT-3, T5) and variants of pre-trained models, it would be possible to determine which ones perform best for specific types of question-answer datasets.
    - Fine-tuning not just on text-based datasets, but also on more complex structured data (e.g., tables, code, etc.), would further expand the model's capabilities in handling diverse query formats.
2. **Dataset Enhancement**
    - The current dataset generation process can be expanded to include a more diverse set of PDFs to increase the robustness of the model. By incorporating PDFs from different domains (e.g., medical, technical, literary), the model can be trained to handle a wide range of topics.
    - Augmenting the dataset by including additional metadata or context (such as summaries or contextual links between questions and answers) could help the model better understand complex relationships and improve the relevance of its responses.
3. **Incorporating Multi-Modal Inputs**
    - Incorporating multi-modal data, such as images or tables within PDFs, could help improve the model's ability to handle diverse forms of content. This would also enable the QA system to be more effective in real-world applications where inputs are often not limited to text alone.

4. **Real-time Question Answering**
    - The next step in the project could involve deploying the fine-tuned model into a real-time QA system. This would include integrating the model into an interactive interface where users can ask questions, and the model provides immediate responses. Implementing a feedback loop to improve model performance based on user input would also help in continuous learning and adaptation.
5. **Evaluation with Different Metrics**
    - The evaluation process could be expanded by integrating more advanced metrics, such as F1 score, BLEU score, or ROUGE score, to assess the quality of the generated answers. A deeper analysis of false positives and false negatives could reveal more specific areas for improvement.
    - Additionally, performance evaluation could be conducted on a broader set of benchmarks, ensuring the model performs effectively across diverse question types, including multi-choice, open-ended, and fact-based questions.
6. **Cross-Lingual and Multilingual QA Systems**
    - Extending the project to support multilingual question answering could significantly increase the applicability of the system. Fine-tuning models on multi-lingual datasets would allow the system to handle questions and answers in multiple languages, thus opening up new possibilities for global applications.
7. **Application to Domain-Specific QA Systems**
    - Fine-tuning the model for specific domains, such as legal, medical, or scientific domains, could improve the relevance and accuracy of answers. By customizing the QA model for these specialized fields, the system could provide more valuable and precise insights tailored to users in these industries.

REFERENCES

[1] Nikahat Mulla & Prachi Gharpure Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications.

https://link.springer.com/article/10.1007/s13748-023-00295-9 .

[2] Kumar, V., Chaki, R., Talluri, S.T., Ramakrishnan, G., Li, Y.-F., Haffari, G.: Question generation from paragraphs: a tale of two hierarchical models. (2019)

[3] Wang, W., Hao, T., Liu, W.: Automatic question generation for learning evaluation in medicine. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 4823 LNCS, pp. 242–251 (2008). https://doi.org/10.1007/978-3-540-78139-4_22

[4] Araki, J., Rajagopal, D., Sankaranarayanan, S., Holm, S., Yamakawa, Y., Mitamura, T.: Generating questions and multiple- choice answers using semantic analysis of texts. In: COL- ING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers. pp. 1125–1136 (2016)

[5] Kunichika, H., Katayama, T., Hirashima, T., Takeuchi, A.: Auto- mated question generation methods for intelligent english learning systems and its evaluation. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 62, 1209–1215 (2005)

[6] Fine-tuning and Utilization Methods of Domain-specific LLMs Cheonsu Jeong1 Dr. Jeong is Principal Consultant & the Technical Leader for AI Automation at SAMSUNG SDS

[7] Automatic Question-Answer Generation for Long-Tail Knowledge: Rohan Kumar, Youngmin Kim, Sunitha Ravi, Haitian Sun, Christos Faloutsos, Ruslan Salakhudinov, Minji Yoon.

[8] Quantity Affects Quality: Fine-Tuning on LLM’s Multiple-Choice Question Abilities Hsuan-Lei Shao, Wei-Hsin Wang and Sieh-Chuen Huang.