Fine-tune LLM for Code Generation

Anmol Ashri
RPTU Kaiserslautern, Department of Computer Science

Note: *This report contains a project documentation and reflection on the portfolio task submitted for the lecture Engineering with Generative AI in WiSe 2023-24. This report is an original work and will be scrutinised for plagiarism and potential LLM use.*

1 Portfolio documentation

1.1 Research Phase

This section gives a detailed analysis of the research phase about choosing a benchmarking dataset and a small pre-trained model enough to be Fine-tuned on google colab notebook.

1.1.1 Dataset

Fine-tune LLM for Code Generation task I took Python Code Instructions 18k alpaca dataset available on huggingface website[1]. It is a collection of Python problem descriptions and corresponding Python code solutions. It is derived from a larger dataset which contains coding instructions of about 120000 codes. The dataset is comprised of 18612 rows, each containing a unique problem description paired with a Python code solution. It has 4 columns:- Instruction, Inputs, Outputs, and prompts. I chose this dataset because because it contains lots of diverse types of Python programming tasks as for this task we need a natural instruction-based python code generation or function generation data. It will help the model to generalize better. This is a benchmark dataset that aligned with this task which requires the model to take natural language instructions to produce Python code, a lot of well-known models have been trained on this dataset and also it has been used widely in research [1]. The 'output' fields contain the Python code or functions that correctly accomplish the task described in the 'instruction' set which makes this dataset highly reliable.

1.1.2 Model

I chose LLaMA 2 [2](Language Model from Meta AI) is a series of transformer-based language models. It is an open-source second-generation pre-trained model which comes in several configuration variants such as :- 7B, 13B, 34B, and 70B (B for billion) parameters. This model is a benchmarked model and for this task, we have to choose a simple pre-trained LLM that should not be pre-trained on specific task so after analyzing the bigcode/big code-models-leaderboard[2] where it performs better than other models on various evaluation criteria. This model is trained on a variety of data which makes it suitable for this task. LLaMa 2 7B[3], the 7 billion parameter

[1] `https://huggingface.co/datasets/iamtarun/python_code_instructions_18k_alpaca`
[2] `https://huggingface.co/spaces/bigcode/bigcode-models-leaderboard`
[3] `https://huggingface.co/NousResearch/Llama-2-7b-hf`

model makes it feasible to finetune on the Google Colab platform [3] where there are limitations on computational resources.

1.2 Design Phase

The design Phase describes the Fine-Tuning Approach used in this task for the LLaMA 2 7B model; Innovation in Design prompts to generate synthetic dataset generation using the given AWS API LLM model and the evaluation metric used for this task.

1.2.1 Fine-Tuning Approach

This task involves using the pre-trained model in my case LLaMA 2 7B on Google Colab platform which makes it difficult to train and then test on Google Colab given the computational resources limitation so for effective use of LLaMA 2 on Google Colab makes it compulsory to use a Fine-Tuning approach.

I used the QLORA [4] (Quantized Low-Rank Optimization for Representation Approximation) fine-tuning approach which is designed to optimize LLMs with an emphasis on efficiency and resource reduction. It is a type of quantization approach that reduces the number of bits required to represent each number in a model's weight matrix, while utilizing low-rank approximations to reduce the overall size and complexity of the model and allows it be effectively used using limited computational resources. It works by replacing large matrix multiplication with low-rank equivalents, which require fewer parameters and computations. Why I chose it for this task due to the limitation of computational resources as using LLAMA 2 7B is expensive to use with Google Colab so using QLORA provides resource efficiency; it greatly increased the speed of execution as a testing model took hours but the model fine-tuned with QLORA took only about 25 minutes or so in my collab notebook execution. Using this approach not only increased the execution speed but also retained or increased model performance on the testing dataset.

1.2.2 Innovation in Design Prompts

This task involves the generation of a synthetic dataset generation task using a given AWS API LLM key (LLAMA 2 70B) and generating a dataset three times the size of the training dataset(in my case it would be the size of 1050 as my training dataset size used is 350). To generate the dataset it needs an innovative design prompt engineering so that it will result in generating a synthetic dataset that must align with the original dataset which is used in this task. Using AWS API, the synthetic dataset is generated which has 2 features:- Instruction and it's Output.

Synthetic_dataset_generation.ipynb file in my zip folder of this task shows the implementation of dataset generation using prompt engineering. I have used a few shot prompt engineering (7 shot examples) to generate examples similar to the original dataset. Meta instruction is provided to the AWS model which guides the generation process. It specifically sets criteria for the examples to be unique, complete, and relevant (as I tried generating without it, the model was generating duplicate, incomplete examples), also I have made sure it will add only the completed examples in the synthetic dataset file. I have given diverse example prompts just

like in the original dataset. Diversity is crucial for training a model to handle various programming scenarios. This approach of prompt engineering with meta instruction and a few shot examples, makes this approach allow for the generation of a large number of synthetic examples quickly and efficiently. Making repeated calls to API will generate different sets of examples, hence scalable which makes it a practical choice for creating extensive dataset, leading to more objective and varied training datasets.

1.2.3 Evaluation Metric Selection Justification

To evaluate the model performance, I used BLEU [5] (Bilingual Evaluation Understudy) which is used in natural language processing to assess the quality of machine-generated text against reference text. It works by comparing n-grams of generated text to n-grams of reference text and calculates the score bases on the precision of matches. I selected this evaluation metric because it has been used in lot's of research related to Fine-tuning of LLM model [4,6]. It gives quantitative measures of how predicted generated text closely aligns with reference (in this case python code or function prediction based upon a given task with original output present in the dataset); N-gram coherence checks the correctness and functionality of output by taking into account about the arrangement of tokens(such as syntax in Python code).

1.3 Implmentation Phase

This section provides details about the implementation of this task. Following are the details of the implementation:-

- All the detailed step-wise implementation details have been explained in respective .ipynb files. Dataset loading, Fine-Tuning, testing and visualization are contained in GenAI.ipynb. Synthetic dataset details are contained Synthetic_dataset_generation.ipynb and interface implementation contained in the interface.ipynb file
- Some of the code is implemented taking reference from exercise.

1.4 Testing, Evaluation, and limitations

1.4.1 Testing, Evaluation

The main focus of this task is to select & implement a Fine-tuning approach and compare the code generation of different sets of models (Model A, B, C, D) based on evaluation metrics, and the generation of synthetic dataset. This section provides a detailed analysis of the testing of Models.

- Each type of Model is tested using BLEU score with tokenizer(max_length=128) and outputs max_new_tokens=100, these values are chosen due to computational resources limitations on collab platform.
- Fine-tuning greatly reduced the inference timing as it took hours to test Model A without Fine-tuning but due to fine-tuning other models took almost like 25 minutes or so.

- Although BLEU scores in general of all models are bad given the training-testing dataset and GPU limitations. Model A performs worst while Model B performs best of all as shown in figure 1. Model A performance bad is due to it was not trained on the training dataset and directly tested on the testing dataset with zero shot example prompt engineering.
- Model C performance is second last as it was trained on a synthetic dataset which shows there were irregularities in the synthetic dataset. Model D stood in second position in terms of BLEU score which makes sense as it was trained on the merged dataset.
- Choice of the testing dataset was made based upon the limitation of GPU resources on the Google Collab platform, as for this task we need to generate a synthetic dataset 3 times the size of the training dataset, so it would be obvious to reduce all dataset size. I tried to split the dataset size on the basis of 70-20-10: train-test-validation split. Testing set of 150 ensures coverage overall dataset to ensure model's performance is accurate and indication of performance in real-world scenarios.
- Giving meta instruction as part of prompts in generation of synthetic dataset and giving diverse examples improved quality of generated examples; although there were some irregularities still present but overall quality improved with this.

1.4.2 Limitations

There were some Limitations of this task which are explained in the following points:

- BLEU scores generated by the models are bad but as per my knowledge, it may be due to limited computational resources on the collab notebook. The chosen dataset size is much less for a model like LLAMA 2 7B, there are not enough samples for the model to generalize. Even with a large set of data, there is considerable resources needed for a fair BLEU score [4].
- Synthetic dataset generated using innovative prompt engineering, still there were some irregularities in a generation like repeated examples, incomplete fields; and there some instruction to complex functions or code which resulted in bad output by AWS API.
- I tried changing inference parameters like max_length, max_new_tokens which gave somewhat better results as shown in GenAI.ipynb but I implemented it on Fine-tuned models only.

1.5 Ethical Consideration

Task Like Code generation involves several ethical considerations which must be taken into consideration. Those considerations are discussed below:-

- **Bias**: Bias is a major problem in the generative models which they inherit and amplify it present in their training data. It is crucial to assess the model's outputs for biases that could lead to unfair results.
- **Accuracy and Reliabilty**: Code generation model models may not always produce accurate or reliable code(also seen in this task too). Result must be checked to prevent the errors which could lead to severe consequences, if implemented.

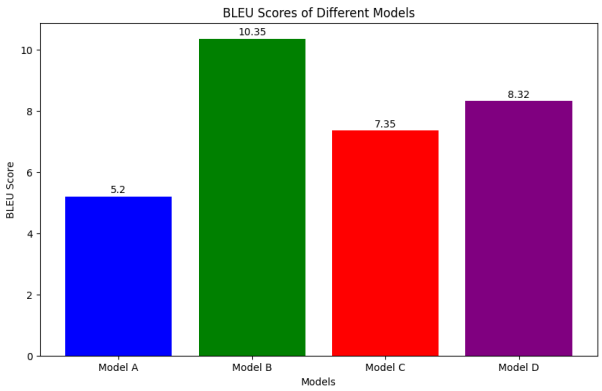

Figure 1: Bar Chart shows comparison models performance based upon BLEU score.

- **Irregularities**: Synthetic code generation tasks must be checked for irregularities such as generating repeated examples, incomplete examples etc.
- **Security**: Generated code could potentially introduce security vulnerabilities into a system. It is must to review and remove those vulnerabilities before use it for production usecase.

2 Reflection

This section contains reflection of the challenges, solution and potential areas for future improvement. Some of the part is already answered in the documentation section.

1. **What was the most interesting thing that you learned while working on the portfolio? What aspects did you find interesting or surprising?**
   Answer: The most interesting thing for me was using QLORA for Fine-tuning LLAMA 2 7B model which I used in this task on Google Collab since it is my first Generative AI project. During this task, I learned about QLORA in more detail and its application and advantages. Most interesting to me was how QLORA can reduce the inference time of a model from hours to just minutes in my case. Implementing it I came to know its impact on model use-case and furthermore, it preserves or enhances model accuracy.
2. **Which part of the portfolio are you (most) proud of? Why? What were the challenges you faced and how did you overcome them?**
   Answer: There are many things which I am proud of during the implementation of this task but to say most is generating a synthetic dataset that is 3 times the size of the training dataset using a given AWS API LLM model. Designed the function and combined prompt with meta instruction and examples which generated the examples similar to the dataset used in this task. Challenges like empty instructions or output fields, generating the same examples again and again which are given as 7-shot examples; successive repetitions of same examples. To overcome these issues, I designed the construct_prompt_with_7_examples() function which gives combined prompts with meta-instruction that directs AWS API to generate python code examples which are unique and

complete. Furthermore, To maintain diversity I changed some examples after generating a batch of synthetic datasets. generate_synthetic_examples() ensures that instruction and output both must be present before adding it to dataset.

3. **What adjustments to your design and implementation were necessary during the implementation phase? What would you change or do differently if you had to do the portfolio task a second time? What would be potential areas for future improvement ?**
   Answer: There were a lot of adjustments had to be made for this task because of computational resource limitations on the Google Collab platform such as choosing the right training-testing-validation dataset size, choosing max_new_tokens length for output and max_length input for the model, choosing a model which can be fine-tuned on Google Collab, generation of synthetic dataset length. If I have the chance to do same task second task, I would like to implement it with good GPU computation without any limitation (May be university DGX cluster or DFKI cluster) and with a new model LLAMA 3. Future improvements involve using a larger dataset with the new model to get a fair or good BLEU score since scores are not in this task.

4. **Include a brief section on ethical considerations when using these models on code generation tasks ?**
   Answer: Already answered in the documentation section above.

5. **From the lecture/course including guest lectures, what topic excited you the most? Why? What would you like to learn more about and why?**
   Answer: I liked "Knowledge Distillation" showcased in the lecture "Look Ma, I Shrunk BERT". This lecture covered innovative techniques like pseudo labeling which can be used to increase the model's learning efficiency and accuracy and other techniques for reducing the size of large language models like BERT while preserving their performance capabilities. I would like to know more use case as it can be used in increasing model efficiency irrespective of computational limitations. More tools and framework which can be used to implement it.

6. **Based on the content of the lecture taught during the semester and the task you have carried out in this portfolio, describe a project that you would like to do using generative AI and LLMs.**
   Answer: Utilizing generative models and techniques like knowledge distillation to create more efficient smaller models that can be deployed as continuous integration/continuous deployment(CI/CD) pipelines in software environments.

7. **What would be a good multiple-choice question for a test to see if a student has really understood the content? Design 3 Multiple Choice Questions (with at least 3 answer choices) keeping the following in mind:**

   a) Which aspect is most crucial for maintaining the team's productivity when generative models for code synthesis are deployed in a collaborative coding environment?
      i. Ability of model to generate code that strictly sticks to the project's existing codebase i.e. minimizing integration issues.

ii. Frequency with which model suggests alternative implementations for existing code.
iii. Model preference to latest programming language or libraries, regardless of their compatibility with existing codebase.

Answer: i

b) Fine-tuning an LLM model like LLaMA 7B using QLORA, which of the following statement reflects the primary advantage of QLORA approach?
i. It enables the fine-tuning of LLMs only on high-end GPUs with more than 50 GB of memory.
ii. It mainly focus on improving the model's accuracy by fine-tuning all layers of model with high precision.
iii. It focuses on fine-tuning of LLMs on GPU by reducing memory usage.

Answer: iii

c) In the RAG models, which component produces relevant context that refined the generation process?
i. The generator, which is responsible for producing final output without external context.
ii. Encoder, which generates output by encoding the input into a fixed-length vector.
iii. Retriever, which uses a knowledge source to obtain context documents that are used to generate the output based on given input.

Answer: iii

References

[1] Bo Shen, Jiaxin Zhang, Taihong Chen, Daoguang Zan, Bing Geng, An Fu, Muhan Zeng, Ailun Yu, Jichuan Ji, Jingyang Zhao, Yuenan Guo, and Qianxiang Wang. Pangu-coder2: Boosting large language models for code with ranking feedback, 2023.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[3] Purnawansyah, Zahrizhal Ali, Herdianti Darwis, Lutfi Budi Ilmawan, Sitti Rahmah Jabir, and Abdul Rachman Manga. Memory efficient with parameter efficient fine-tuning for code generation using quantization. In *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–6, 2024.

[4] Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In Philipp Koehn, Barry

Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore, December 2023. Association for Computational Linguistics.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[6] Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. Teaching large language models an unseen language on the fly, 2024.

Informed Consent of Participation

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

You are invited to participate in the field study **LLM Education** initiated and conducted by Applied Machine Learning group. The research is supervised by **Sebastian J Vollmer**. Please note:

- Your participation is entirely voluntary and can be withdrawn at any time.
- The field study will last approximately 6 weeks.
- We will record the documentation submitted by the student at the end of the portfolio exam.
- All records and data will be subject to standard data use policies.

If you have any questions or complaints about the whole informed consent process please contact Sebastian J Vollmer (E-Mail: sebastian.vollmer@dfki.de).

Purpose and Goal of this Research

Ability to create multiple choice questions of LLMs. Can we prompt an LLM to create good multiple choice questions? Your participation will help us achieve this goal. The results of this research may be presented at scientific or professional meetings or published in scientific proceedings and journals.

Participation and Compensation

Your participation in this study is completely voluntary and is unpaid.

Procedure

After confirming the informed consent the procedure is as follows:

1. Student submits the exam with deliverables
2. Student also creates questions of their choice based on instructions of good multiple choice questions.
3. We compare LLM's output to MCQs created by students.

The complete procedure of this field study will last approximately 6 weeks.

Risks and Benefits

There are no risks associated with this field study. We hope that the information obtained from your participation may help to bring forward the research in this field. The confirmation of participation in this study can be obtained directly from the researchers.

Data Protection and Confidentiality

We are planning to publish our results from this and other sessions in scientific articles or other media. These publications will neither include your name nor cannot be associated with your identity. Any demographic information will be published anonymized and in aggregated form. Contact details (such as e-mails) can be used to track potential infection chains or to send you further details about the research. Your contact details will not be passed on to other third parties. Any data or information obtained in this field study will be treated confidentially, will be saved encrypted, and cannot be viewed by anyone outside this research project unless we have you sign a separate permission form allowing us to use them. All data you provide in this field study will be subject of the General Data Protection Regulation (GDPR) of the European Union (EU) and treated in compliance with the GDPR. Faculty and administrators from the campus will not have access to raw data or transcripts. This precaution will prevent your individual comments from having any negative repercussions. During the study, we log experimental data, and take notes during the field study. Raw data and material will be retained securely and compliance with the GDPR, for no longer than necessary or if you contact the researchers to destroy or delete them immediately. As with any publication or online-related activity, the risk of a breach of confidentiality or anonymity is always possible. According to the GDPR, the researchers will inform the participant if a breach of confidential data was detected.

Identification of Investigators

If you have any questions or concerns about the research, please feel free to contact:
Sebastian J Vollmer
Principal Investigator Trippstadter Str. 122
67663 Kaiserslautern, Germany sebastian.vollmer@dfki.de

☒ I understand the explanation provided to me. I understand and will follow the hygiene rules of the institution. I understand that this declaration of consent is revocable at any time. I have been given a copy of this form. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this field study.

☒ I agree that the researchers will and take notes during the field study. I understand that all data will be treated confidentially and in compliance with the GDPR. I understand that the material will be anonymized and cannot be associated with my name. I understand that full anonymity cannot be guaranteed and a breach of confidentiality is always possible. From the consent of publication, I cannot derive any rights (such as any explicit acknowledgment, financial benefit, or co-authorship). I understand that the material can be published worldwide and may be the subject of a press release linked to social media or other promotional activities. Before publication, I can revoke my consent at any time. Once the material has been committed to publication it will not be possible to revoke the consent.

Kaiserslautern, 29.04.2024

Signature **Place, Date**