ISSN 2224-087X. Електроніка та інформаційні технології. 2024. Випуск 26. С. 16–24
Electronics and information technologies. 2024. Issue 26. P. 16–24

UDK 004.89 DOIhttps://doi.org/10.30970/eli.26.2

METRIC-BASED COMPARISON OF FINE-TUNED LLAMA 2 AND MIXTRAL LARGE LANGUAGE MODELS FOR INSTRUCTION TASKS

B. Pavlyshenko, I. Bulka

System Design Department,
Ivan Franko National University of Lviv,
50 Drahomanova St., 79005 Lviv, Ukraine
bohdan.pavlyshenko@lnu.edu.ua,
ivan.bulka@lnu.edu.ua

The paper considers a comprehensive analysis and comparative study of two advanced Large Language Models (LLMs), namely LLaMA 2 and Mixtral, with a specific focus on their performance in executing instructional tasks. These models were fine-tuned using techniques such as LoRA and QLoRA, which were applied to extensive instruction datasets. The fine-tuning process was further enhanced by the implementation of Parameter-Efficient Fine-Tuning (PEFT) on NVIDIA A100 Tensor Core GPU instances, ensuring optimal performance. Both LLaMA 2 and Mixtral models were fine-tuned using the Hugging Face and PyTorch platforms, ensuring that similar parameters were maintained to facilitate a fair comparison. An inference was made using data not used in the initial training phase. This approach was adopted to test the models' ability to generalize and adapt to new, unseen data, thereby providing a more robust evaluation of their performance. An evaluation framework was established using the RAGAS library. The framework was designed to provide precise and reliable metrics, offering a comprehensive assessment of the models' performance. While the LLaMA 2 model demonstrates a faster rate of fine-tuning, it is susceptible to overfitting. On the other hand, Mixtrail, despite requiring more time for training, outperforms in evaluations, making it a more dependable tool for instructional tasks.

Keywords: LLMs, PEFT, Lora, Qlora, Mixtral, LLaMA, LLMs fine-tuning

Introduction

The progress in Natural Language Processing (NLP) has been significantly driven by advanced models like Transformers [1], BERT [2], and GPT [3], which use Attention Mechanisms [4] and similar technologies. These models have greatly improved tasks such as text classification, machine translation, sentiment analysis, and summarization. Newer models like GPT-3.5 [5], GPT-4 [6], Claude-2 [7], and BARD [8] have expanded NLP capabilities, introducing features like learning from context and handling tasks without prior specific training.

However, using these large language models (LLMs) for specialized tasks, especially in business, can be challenging. Typically, Supervised Fine-Tuning (SFT) [9] is used to adapt these general models to specific needs, but it's difficult to keep their broad language skills while also making them efficient for specific domains [10]. In commercial settings, it's also hard to measure how well these models perform because there isn't a standard way to evaluate them, and existing benchmarks [11] don't fit well with specialized industrial needs [12].

It's important to note that many of these models, like GPT-4, Claude-2, and BARD, are accessible through Application Programming Interfaces (APIs). However, using APIs raises issues about data privacy due to unclear data handling practices [13]. An alternative could be Retrieval Augmented Generation (RAG) [14], but this still involves sending data to external services. Also, frequent API use can become expensive. A better approach might be to fine-tune models on personal servers, which could be more cost-effective and secure.

Fine-tuned LLMs can be used in many ways and can greatly impact different areas of our lives. For example, they can be very useful in media and communication, helping to tell the difference between real news and fake or biased news [15]. LLMs can also be used in finance, where they can help analyze financial news in detail [16]. The fact that LLMs can be used in so many ways shows how important they are and how much potential they have for use in different areas.

This paper focuses on comparing two advanced LLMs, LLaMA 2 [17] and Mixtral [18], specifically on how well they perform instructional tasks. These models were fine-tuned using advanced techniques like LoRA [19] and QLoRA [20] on NVIDIA A100 Tensor Core GPUs [21], using methods that don't require many resources (Parameter-Efficient Fine-Tuning, PEFT) [22, 23].

We also developed an evaluation framework [24] that suits both industrial and general uses, providing accurate and reliable performance metrics. This research could help guide future LLM applications and studies.

The paper is organized as follows: We start with a literature review on the evolution of LLMs and their use in specialized areas. Next, we discuss the data and methods used in our study. We then present our findings and discuss their implications for both research and practical applications. The paper concludes with a summary of key points and suggestions for further research.

Methodology

This research focused on adjusting two models, LLaMA 2 and Mixtrail. These models come in three different sizes: 7b, 13b, and 70b. However, due to limited resources, the smallest size, 7b, was chosen for this experiment. The fine-tuning was done using the PyTorch library.

The training dataset consists of two open-source datasets: instruct-v3 and alpaca, both of which are instructional collections designed for fine-tuning instructions. The instruct-v3 dataset is obtainable from GitHub (https://huggingface.co/datasets/mosaicml/instruct-v3) and consists of 56k entries. The alpaca repository is also hosted on GitHub (https://huggingface.co/datasets/tatsu-lab/alpaca) and has 52k records. Were used datapoints from these datasets where the instructions were shorter than 2048 tokens.

The two models were trained on the same dataset, which was segmented into three distinct sections: training, validation, and testing, which contained 83k, 10k, and 3k records respectively. The training segment was utilized to fine-tune the models, the validation segment to evaluate the training outcomes during the fine-tuning procedure, and the testing segment to assess the final model outcomes.

All instructions were formatted according to the following template:

```
    <s>[INST] Below is an instruction that describes a task. Write a response that appropriately completes the request.

     {input} [/INST]

     {response}</s>
```

Example:

<s>[INST] Below is an instruction that describes a task. Write a response that appropriately completes the request. Identify the odd one out. Twitter, Instagram, Telegram[/INST]Telegram</s>

Both models were configured using LoraConfig [25], with the following parameters:

Table 1. Lora Config for LLMs fine-tuning

Parameter	Paremeter description	Value
lora_alpha	LoRA scaling factor	16
lora_dropout	Dropout parameter to reduce overfitting	0.1
r	Matrix rank, relates to the amount of trainable parameters	64

The models were fine-tuned using PEFT [26, 27], which helped to reduce the hardware resources required.

Given that both models required tens of gigabytes of RAM, NVIDIA A100 Tensor Core GPU instances were used for fine-tuning. The parameters for fine-tuning were as follows:

Table 2: Training parameters for LLMs fine-tuning

Parameter	Parameter description	Value
num_train_epochs	Number of training epochs	2
per_device_train_batch_size	Batch size	4
warmup_steps	The number of warm-out steps	0.03
learning_rate	Learning rate	2.5e-5
bf16	16-bit floating point format	True
max_seq_length	Max number of tokens	1024

The models were evaluated using the test dataset, which was not used during the fine-tuning process. Evaluating LLMs is a complex task, and for this study, two techniques were applied. Firstly, instructions, expected responses, and actual responses were sent to the GPT-4 model, which was asked to assign a score between 1 and 10. A higher score indicated a better adherence to the instructions. Secondly, the RAGAS library (which can be obtained by the link: https://docs.ragas.io/en/stable/) was used to evaluate the model results using two metrics: Answer Correctness [28] and Answer Semantic Similarity [29].

Experiments and Results

When comparing how different models learn over time, the LLaMA 2 model stands out because it learns very quickly, as shown by its fast-dropping training loss. However, when we look at the validation loss, which helps us understand how well the model performs on new, unseen data, it becomes clear that LLaMA 2 tends to overfit quickly. This means it performs well on the training data but not as well on new data, as noted in the literature [30]. As training continues, it is expected that LLaMA 2 will continue to show better results on the training data but worse results on the validation data. Attempts to fix this issue by using a technique called dropout have not been very successful in changing this trend.

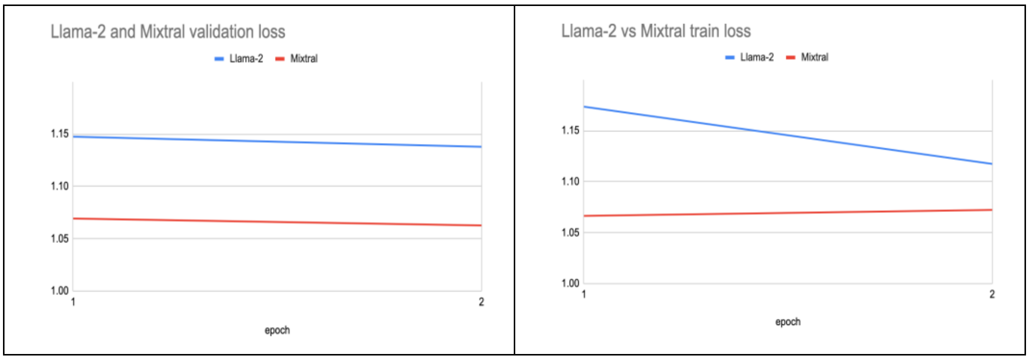

Fig. 1. Training loss comparison

The table shows that Mixtrail takes more than double the time to train compared to LLaMA 2. This is important for projects with limited budgets, as saving time means saving money. The training was done using an NVIDIA A100 Tensor Core GPU and involved about 83,000 training records. If the number of records increases or if larger models, such as 13 billion or 70 billion parameter models, are used instead of a 7 billion parameter model, the costs could increase significantly.

Table 3. Training time comparison

Model	**LLaMA 2**	**Mixtral**
Training time, hours:mins	3:27	7:36

Metrics are crucial for measuring how well a model performs specific tasks. The better the metrics, the more capable the model is of handling these tasks.

Were evaluated the basic versions of the LLaMA 2 and Mixtrail models and compared them to their enhanced versions after fine-tuning. This comparison aims to understand how beneficial fine-tuning is for the tasks we are focusing on.

We will assess the models using three main metrics: GPT-4 score, Answer Correctness, and Answer Semantic Similarity.

Table 4. Metrics comparison

Model	Gpt-4 score (max 10)	Answer Correctness	Answer Semantic Simiilarity
LLaMA 2, base	7.21	0.66	0.91
LLaMA 2, fine-tuned	6.96	0.63	0.91
Mixtral, base	7.12	0.62	0.91
Mixtral, fine-tuned	7.51	0.67	0.91

When assessing how well models perform, it's crucial to understand the limitations of the metrics used. The Answer Semantic Similarity metric, which measures how closely the model's response matches the actual answer in terms of meaning, might not be ideal for tasks involving detailed instructions, especially in specialized areas like physics or mathematics. This is because it may not capture all the nuanced information in such instructions.

Another metric, Answer Correctness, looks at both how semantically similar the model's answer is to the correct answer and how factually accurate it is. This makes it a more comprehensive measure.

In tests where these metrics are applied, GPT-4 is used to determine which model scores highest on these instruction tasks. Among the models evaluated, the Mixtral model fine-tuned for this purpose shows the best performance. Interestingly, the LLaMA 2 base model ranks second based on the Answer Correctness metric.

Discussions and Future direction

The comparison between the LLaMA 2 and Mixtrail models shows how they perform when carrying out instruction tasks. The results show that LLaMA 2 adjusts quickly, but it tends to overfit. On the other hand, Mixtrail takes more time to train but performs better in evaluations, making it a better choice for instruction tasks. The results highlight the importance of fine-tuning to improve the performance of LLMs for specific tasks.

The study used the RAGAS library to create an evaluation framework. This framework has been effective in measuring the models' performance by providing accurate and reliable metrics. It opens up opportunities for future analysis and applications of LLMs, making the fine-tuning process easier and providing a method for quick and efficient performance measurement and metrics evaluation.

The research emphasizes the need to find a balance between maintaining general language capabilities and achieving efficiency in specific tasks, a common challenge in business applications. These insights can guide future research and applications in machine learning, particularly those focusing on Large Language Models and their use in instruction tasks.

As we move forward, the focus of our study will be broadened to assess these models within various specialized domains, not merely instructional tasks. There will also be a concentrated effort to fine-tune techniques and enhance evaluation metrics employed in this study. This initiative could pave the way for the creation of more effective and efficient LLMs, elevating the field of Natural Language Processing to new heights. Noteworthy, we aim to focus extensively on improving the 'reasoning' component in future research.

Conclusion

The comparison of the LLaMA 2 and Mixtrail models has given us important information about how they work with instructional tasks. LLaMA 2 is quicker to fine-tune, but it can overfit. On the other hand, Mixtrail takes longer to train, but it performs better in tests, making it more dependable for instructional tasks.

The study uses the RAGAS library to create a way to evaluate how well LLMs perform. The information from this study could help guide future research, especially in machine learning where LLMs are used for instructional tasks.

In conclusion, these results remind us that fine-tuning is important for improving how well LLMs perform. They also highlight the importance of finding a balance between general language skills and efficiency in specific tasks, especially in business settings.

References

[1] *Lin Tianyang, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu.* "A survey of transformers." AI open 3 (2022): 111-132.

[2] *Devlin Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova*. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[3] *Liu Xiao, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang*. "GPT understands, too." AI Open (2023).

[4] *Vaswani Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin.* "Attention is all you need." Advances in neural information processing systems 30 (2017).

[5] *Brown Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al*. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[6] OpenAI, R. "Gpt-4 technical report. arxiv 2303.08774." *View in Article* 2, no. 5 (2023).

[7] *Bai Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain et al*. "Training a helpful and harmless assistant with reinforcement learning from human feedback." arXiv preprint arXiv:2204.05862 (2022).

[8] *Zhang Zheng, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao.* "Balancing specialized and general skills in llms: The impact of modern tuning and data strategy." arXiv preprint arXiv:2310.04945 (2023).

[9] *Alt Christoph, Marc Hübner, and Leonhard Hennig.* "Fine-tuning pre-trained transformer language models to distantly supervised relation extraction." *a*rXiv preprint arXiv:1906.08646 (2019).

[10] *Zhang Zheng, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao.* "Balancing specialized and general skills in llms: The impact of modern tuning and data strategy." arXiv preprint arXiv:2310.04945 (2023).

[11] *Wang Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman*. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." arXiv preprint arXiv:1804.07461 (2018).

[12] *Zhang Zheng, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao*. "Balancing specialized and general skills in llms: The impact of modern tuning and data strategy." arXiv preprint arXiv:2310.04945 (2023).

[13] *Yao Yifan, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang.* "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly." High-Confidence Computing (2024): 100211.

[14] *Lewis Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al.* "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems. 33. (2020): 9459-9474.

[15] *Pavlyshenko Bohdan M.* "Analysis of disinformation and fake news detection using fine-tuned large language model." arXiv preprint arXiv:2309.04704 (2023).

[16] *Pavlyshenko Bohdan M.* "Financial News Analytics Using Fine-Tuned Llama 2 GPT Model." arXiv preprint arXiv:2308.13032 (2023).

[17] *Touvron Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov et al.* "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).

[18] *Jiang Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot et al.* "Mixtral of experts." arXiv preprint arXiv:2401.04088 (2024).

[19] *Qin Haotong, Xudong Ma, Xingyu Zheng, Xiaoyang Li, Yang Zhang, Shouda Liu, Jie Luo, Xianglong Liu, and Michele Magno.* "Accurate LoRA-Finetuning Quantization of LLMs via Information Retention." arXiv preprint arXiv:2402.05445 (2024).

[20] *Dettmers Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer*. "Qlora: Efficient finetuning of quantized llms." Advances in Neural Information Processing Systems. 36. (2024).

[21] *Choquette Jack, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky.* "Nvidia a100 tensor core gpu: Performance and innovation." IEEE Micro 41, no. 2 (2021): 29-35.

[22] *Ding Ning, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu et al*. "Parameter-efficient fine-tuning of large-scale pre-trained language models." Nature Machine Intelligence 5, no. 3 (2023): 220-235.

[23] *Fu Zihao, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier*. "On the effectiveness of parameter-efficient fine-tuning." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, pp. 12799-12807. 2023.

[24] *Desmond Michael, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M. Johnson*. "EvaluLLM: LLM assisted evaluation of generative outputs." In Companion Proceedings of the 29th International Conference on Intelligent User Interfaces, pp. 30-32. 2024.

[25] *Dettmers Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer*. "Qlora: Efficient finetuning of quantized llms." Advances in Neural Information Processing Systems 36 (2024).

[26] *Ding Ning, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu et al*. "Parameter-efficient fine-tuning of large-scale pre-trained language models." Nature Machine Intelligence 5, no. 3 (2023): 220-235.

[27] *Fu Zihao, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier*. "On the effectiveness of parameter-efficient fine-tuning." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, pp. 12799-12807. 2023.

[28] *Zuccon Guido, Bevan Koopman*. "Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness." arXiv preprint arXiv:2302.13793 (2023).

[29] *Xia Peipei, Li Zhang, and Fanzhang Li.* "Learning similarity with cosine similarity ensemble." Information sciences 307 (2015): 39-52.

[30] *Srivastava Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov*. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15, no. 1 (2014): 1929-1958.

МЕТРИЧНЕ ПОРІВНЯННЯ ТОНКО НАЛАШТОВАНИХ МОВНИХ МОДЕЛЕЙ LLaMA 2 ТА MIXTRAL LARGE ДЛЯ ЗАВДАНЬ З ІНСТРУКЦІЯМИ

Б. Павлишенко, І. Булка

Кафедра системного проектування,
Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 Львів, Україна
bohdan.pavlyshenko@lnu.edu.ua,
ivan.bulka@lnu.edu.ua

У роботі проведено комплексний аналіз і порівняльне дослідження двох великих мовних моделей, а саме LLaMA 2 і Mixtral, з акцентом на їхній продуктивності при виконанні навчальних завдань. Це моделі з відкритим кодом і доступні для широкого загалу. Дані моделі уже навчені на великих наборах даних і нашею задачею було їхнє тонке налаштування.

Моделі були налаштовані за допомогою методів, таких як LoRA та QLoRA, які застосовувалися до великих наборів даних інструкцій. Основною ідеєю даних методів - є ефективне використання ресурсів за рахунок оптимізації тренувального процесу і параметрів. Тренувальний датасет складався з 10 тисяч інструкцій. Задачею тонкого налаштування було навчити моделі ефективно слідувати інструкціям.

Процес тонкого налаштування було покращено завдяки реалізації параметрів-ефективного тонкого налаштування (PEFT) з використанням графічного процесора NVIDIA A100 Tensor Core GPU, що забезпечує оптимальну продуктивність. Обидві моделі LLaMA 2 і Mixtral були налаштовані за допомогою платформ Hugging Face і PyTorch, використовуючи підтримку однакових параметрів для забезпечення порівняння. Тонке налаштування обох моделей відбувалось на протязі 2 епох.

На протязі навчання, моделі евалюейтились, щоб зрозуміти наскільки ефективно проходить навчання. Основною метрикою для евалюейшину моделі була лосс функція.

Аналіз моделей був зроблений на основі даних, які не були включені у фазу навчання і евалюейшину. Цей підхід було прийнято для перевірки здатності моделей узагальнювати та адаптуватися до нових, невідомих даних, забезпечуючи таким чином більш надійну оцінку їх ефективності. З допомогою бібліотеки RAGAS і спеціально створеної метрики GPT-4 score було створено систему оцінювання ефективності LLM моделей. Дані моделі оцінювались на основі декількох метрик для більш надійної оцінки.

Хоча модель LLaMA 2 демонструє швидшу швидкість тонкого налаштування, вона чутлива до перенавчання. З іншого боку, Mixtrail, незважаючи на те, що вимагає більше часу для навчання, показує кращі метрики, що робить його більш надійним інструментом для виконання завдань зв’язаних зі виконанням інструкцій.

Ключові слова: великі мовні моделі PEFT, Lora, Qlora, Mixtral, LLaMA, тонке налаштування великих мовних моделей.

The article was received by the editorial office on 14.05.2024.
Accepted for publication on 21.05.2024.