Fine-Tuning LLMs for Self-Awareness and Identity

1st Guilherme G. Zanetti
Universidade Federal do Espírito Santo (UFES)
AUMO SA
Vitória, Brasil
guilherme.zanetti@edu.ufes.br

2nd Rafael F. Freitas
Universidade Federal do Espírito Santo (UFES)
AUMO SA
Vitória, Brasil
rafael.f.freitas@edu.ufes.br

***Abstract*—This work introduces AumoGPT, a Llama3 3B Instruct model finetuned using techniques such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA), with knowledge about the AUMO startup and who it is (AumoGPT) . Our experiments show that, with a small dataset, it is possible to finetune a model with specific knowledge while maintaining most of the original model capabilities.**

***Index Terms*—Large Language Models (LLMs), Generative Pre-trained Transformer (GPT), Low-Rank Adaptation (LoRA), Quantized Low-Rank Adaptation (QLoRA), Hellaswag, Massive Multitask Language Understanding (MMLU), Global Massive Multitask Language Understanding (Global MMLU), Global Massive Multitask Language Understanding for Portuguese (Global MMLU PT).**

I. Introduction

Large Language Models (LLMs) have become powerful tools for natural language processing since the introduction of the GPT architecture [1]. However, generic LLMs lack specific knowledge and identity tied to particular contexts, such as a company or brand. This paper presents the ”AumoGPT” project, which aims to fine-tune an open-source LLM to create a chatbot for the startup AUMO [2]. The goal is to enable the model to identify itself as AumoGPT, gain knowledge about AUMO [2], and maintain its general language abilities. The project targets a practical, scalable solution for personalized conversational agents.

II. Literature Review

Fine-tuning LLMs has been widely studied to adapt models to specific tasks. Techniques like Low-Rank Adaptation (LoRA) [3] and Quantized Low-Rank Adaptation (QLoRA) [4] allow efficient training on smaller hardware while preserving model performance. Previous works, such as those on Llama [5] and Qwen2.5 [6], show that fine-tuning can embed domain-specific knowledge without fully retraining a model from scratch. However, balancing specialized knowledge with general language skills remains a challenge. This project builds on these findings by applying QLoRA [4] to a small Llama3 3B Instruct model [5], aiming to create a self-aware chatbot with minimal loss of broader capabilities.

III. Methodology

The methodology involves fine-tuning an open-source LLM, specifically Llama3 3B Instruct [5], to develop AumoGPT. The process includes two main deliverables. First, a dataset of 298 examples about AUMO [2] and AumoGPT is generated to provide training data. Second, the LoRA [3] and QLoRA [4] fine-tuning techniques, implemented via the Oumi framework [7], are used to fine-tune the model. The training focuses on enabling the model to identify as AumoGPT and understand AUMO-specific information. Performance is evaluated quantitatively (e.g., accuracy on AUMO-related questions) and qualitatively (e.g., conversational coherence). Additional tests assess the retention of general language skills.

A. Dataset Generation

Datasets are generally one of the difficulties involved in training and fine-tuning a LLM. Since acquiring a specific dataset would require a lot of time and effort, the dataset used for this work was artificially generated using ChatGPT. The dataset was built all in Portuguese based on a Q&A model, which would simulate interactions between platform users and the model. The questions were built with available information on AUMO [2] and how the model would identify itself as, which had to be responded to accordingly. For the development of AumoGPT, a dataset was built with 298 examples, which had evenly distributed Q&A about AUMO and AumoGPT. This dataset was split into training, validation and test datasets, with 238, 30 and 30 examples respectively.

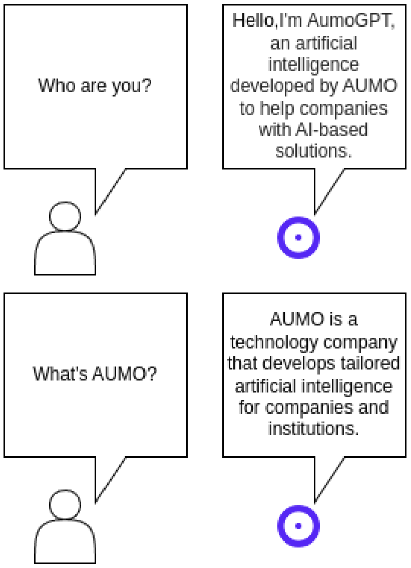
Who are you?
Hello,I'm AumoGPT, an artificial intelligence developed by AUMO to help companies with AI-based solutions.
What's AUMO?
AUMO is a technology company that develops tailored artificial intelligence for companies and institutions.

Fig. 1. Illustration of AumoGPT interacting with a user. The questions and responses present in this picture are part of the dataset.

B. LoRA/QLoRA Fine-tuning

The framework Oumi [7] was used for the finetuning implementation. Before finetuning, default parameters were chosen and used as a baseline.

Hyperparameter	**Baseline**	**QLoRA**	**LoRA**
Learning Rate	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$4 \cdot 10^{-5}$
Batch Size per Device	1	1	4
Number of Epochs	2	8	2
Optimizer	AdamW	AdamW	AdamW
Weigth Decay	0.01	0.01	0.01
Warmup Step	10	10	10
LoRA Rank	128	128	128
LoRA α	32	32	32
LoRA Dropout	0	0	0.2

TABLE I
HYPERPARAMETERS FOR FINETUNING

The finetuned models were evaluated using benchmarks described in Section IV. No method was used to change the hyperparameters and acquire the best perfomant model of each procedure (LoRA [3] and QLoRA [4]). The best hyperparameters, as shown in Table I, were empirically obtained by finetuning with different hyperparameters and then comparing the results by human interaction and by using the benchmarks presented in Section IV.

IV. EXPERIMENTS

To evaluate the quality of the fine-tuned model, two experiments were performed:

1. General Language Skills: This experiment aims to measure the deterioration in the base LLM's language capabilities, like commonsense knowledge, question answering and Portuguese ability. For such, three open benchmarks were used: Massive Multitask Language Understanding (MMLU) [8], Hellaswag [9] and Global Massive Multitask Language Understanding (Global MMLU) [10] for Portuguese (Global MMLU PT). The two fine-tuned models (QLoRA Model and LoRA Model) and two baseline models (Llama3 3b Instruct with and without 4bit quantization) were tested against these benchmarks.

2. AUMO Knowledge: The second experiment aims to evaluate the model's knowledge of the AUMO startup and if it acts as expected from AumoGPT. For this, the 30 questions from the test database were answered by the trained models and manually evaluated as good answer - one AumoGPT should answer - or a bad answer, which is not according to AUMO's directives.

V. RESULTS

The results of the finetuning experiments show that, although Hellaswag [9] score did not change much between models, MMLU [8] and Global MMLU [10] for Portuguese scores, in some models, achieved values up to 14% lower than the respective QLoRA/LoRA baseline model, indicating generalization loss. A QLoRA generalization loss example is shown in Figure 2.

The QLoRA/LoRA hyperparameters in Table I were used to acquire the results in Table II. Table II shows that it is possible to acquire specific-purpose models while maintaining close benchmark scores or even increasing them, suggesting that the finetune model had preserved most of the QLoRA/LoRA original model capabilities while obtaining a good accuracy related to AUMO's directives as shown in Table III.QLoRA model higher accuracy in Table III can be explained as a consequence of higher number of epochs when finetuning, described on Table I, with the tradeoff of lower benchmark scores when compared with it's LoRA model counterpart.

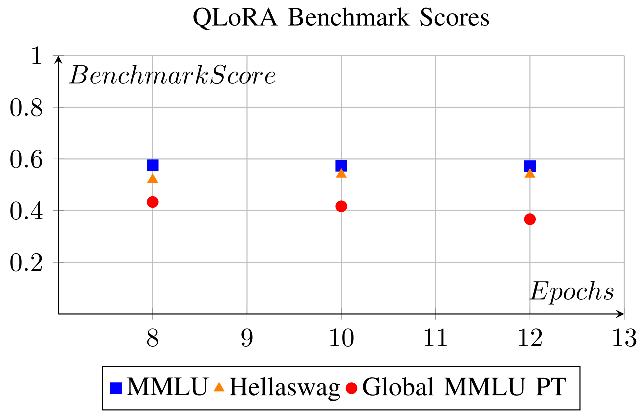

Fig. 2. Finetune effects in benchmarks scores when changing QLoRA "Number of Epochs" hyperparameter in Table I

Model	**MMLU**	**Global MMLU PT**	**Hellswag**
Baseline Model	0.624	0.475	0.52
LoRA Model	**0.63**	**0.475**	**0.52**
Baseline Quantized Model	0.585	0.458	0.5
QLoRA Model	0.575	0.433	0.52

TABLE II
BENCHMARKS RESULT

Model	**Good Answers**	**Accuracy**
LoRA Model	23/30	76.7%
QLoRA Model	**27/30**	**90.0%**

TABLE III
AUMO KNOWLEDGE RESULTS

VI. CONCLUSION

This project demonstrates a practical approach to fine-tuning LLMs for self-awareness and identity, using AumoGPT as a case study. By leveraging LoRA [3] / QLoRA [4] and a structured methodology, it addresses the need for personalized chatbots in startups like AUMO [2] . The work highlights the potential of open-source models for cost-effective, tailored solutions. Future efforts will focus on scaling the dataset and refining evaluation methods to enhance performance further.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] AUMO, “AUMO: Transformamos dados em soluções de IA avançadas,” 2025, [Online; accessed 20-Mar-2025]. [Online]. Available: https://www.aumo.ai/

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv*, p. 2106.09685, 2021.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” *arXiv*, p. 2305.14314, 2023.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” *arXiv*, p. 2302.13971, 2023.

[6] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu *et al.*, “Qwen2.5 technical report,” *arXiv*, p. 2412.15115, 2024.

[7] Oumi, “Oumi: Let’s build better AI open is the path forward,” 2025, [Online; accessed 20-Mar-2025]. [Online]. Available: https://www.oumi.ai/

[8] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, “Measuring massive multitask language understanding,” *arXiv*, p. 2009.03300, 2020.

[9] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, “Hellaswag: Can a machine really finish your sentence?” *arXiv*, p. 1905.07830, 2019.

[10] S. Singh, A. Romanou, C. Fourrier, D. I. Adelani, J. G. Ngui, D. Vila-Suero, P. Limkonchotiwat, K. Marchisio, W. Q. Leong, Y. Susanto, R. Ng, S. Longpre, W.-Y. Ko, S. Ruder, M. Smith, A. Bosselut, A. Oh, A. F. T. Martins, L. Choshen, D. Ippolito, E. Ferrante, M. Fadaee, B. Ermis, and S. Hooker, “Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation,” *arXiv*, p. 2412.03304, 2024.