Scaling and Adapting Large Language Models for Portuguese Open Information Extraction: A Comparative Study of Fine-Tuning and LoRA

Alan Melo[1], Bruno Cabral[1][0000−0002−5221−2860], and Daniela Barreiro Claro[1][0000−0001−8586−1042]

FORMAS Research Center on Data and Natural Language
Institute of Computing - Federal University of Bahia – Salvador–Bahia, Brazil
{alan.melo,bruno.cabral,dclaro}@ufba.br
http://www.formas.ufba.br

Abstract. This paper comprehensively investigates the efficacy of different adaptation techniques for Large Language Models (LLMs) in the context of Open Information Extraction (OpenIE) for Portuguese. We compare Full Fine-Tuning (FFT) and Low-Rank Adaptation (LoRA) across a model with 0.5B parameters. Our study evaluates the impact of model size and adaptation method on OpenIE performance, considering precision, recall, and F1 scores, as well as computational efficiency during training and inference phases. We contribute to a high-performing LLM and novel insights into the trade-offs between model scale, adaptation technique, and cross-lingual transferability in the OpenIE task. Our findings reveal significant performance variations across different configurations, with LoRA demonstrating competitive results. We also analyze the linguistic nuances in the Portuguese OpenIE that pose challenges for models primarily trained on English data. This research advances our understanding of LLM adaptation for specialized NLP tasks and provides practical guidelines for deploying these models in resource-constrained and multilingual scenarios. Our work has implications for the broader cross-lingual open information extraction field and contributes to the ongoing discourse on efficient fine-tuning strategies for large pre-trained models.

Keywords: OpenIE · Language Model · Information Extraction.

1 Introduction

Open Information Extraction (OpenIE) is an NLP task that extracts structured data from documents [2]. Recently, such tasks have been inputted into different pipelines to facilitate the complexity of a set of NLP applications such as QA systems, mental maps, fake news approaches, etc. Following the evolution of the Large Language Model (LLM) domain, the OpenIE task has enlarged its approaches to employ LLM architectures.

As OpenIE has seen significant advancements in the Portuguese language in the last few years, the application of Large Language Models (LLMs) still needs

to be explored. As the evolution of LLMs increases the computational cost [10], advanced language model adaptation methods such as Full Fine-Tuning (FFT) and Low-Rank Adaptation (LoRA) are essential to improve performance and efficiency. Thus, the Portuguese language benefits from such adaptation methods to evolve its NLP applications.

This study aims to examine the potential of LLMs when applied to Portuguese OpenIE, evaluate the impact of model size and adaptation method, consider precision, recall, and F1 scores, and examine computational efficiency during training and inference phases. We contribute to a high-performing LLM and novel insights into the trade-offs between model scale, adaptation technique, and cross-lingual transferability in the OpenIE task for the Portuguese language.

Our results suggest a significance performance between the techniques and provide insights into the practical implications of their application in the Open Information Extraction task for the Portuguese language.

This work is organized as follows: Section 2 describes the OpenIE task, Section 3 outlines the related work, and positions our approach in the state of the art. Section 4 presents the Foundations and Adaptation techniques, Section 5 describes our methodology, and Section 6 presents our evaluations and the conclusion and future research directions following it.

2 Open Information Extraction (OpenIE)

We introduce the definition of Open Information Extraction (OpenIE) and the model to assess the performance of Large Language Models (LLMs) as triple extractors for the Portuguese language. This section explores the definition and potential applications of OpenIE, highlighting key contributions and ongoing challenges.

2.1 OpenIE Definition

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a sentence composed of tokens x_i. An OpenIE triple extractor is a function mapping X to a set $Y = \{y_1, y_2, \cdots, y_j\}$, where each element is a tuple $y_i = \{rel_i, arg1_i, arg2_i\}$ that encapsulates the information conveyed in sentence X.

We assume that tuples are always in the format $y = (arg_1, rel, arg_2)$, with $arg1$ and $arg2$ being noun phrases created from tokens in X, and rel representing a relation between arg_1 and arg_2. As it is common in the area, we do not consider extractions consisting of n-ary relations.

2.2 OpenIE applications

OpenIE has been used in a pipeline to serve diverse applications but is not limited to them.

- **Knowledge graph:** Extracting structured information from text, OpenIE facilitates the automatic creation of knowledge graphs, which are crucial for semantic search, question-answering systems, and decision support systems.
- **Text Summatization:** OpenIE can help identify key facts and relationships in a text, helping to generate concise summaries.
- **Querying Answering:** Triples extracted directly from questions in a QA system can provide factual answers or help increase implicit knowledge inference extraction.

2.3 OpenIE Challenges

Despite significant advances in Open Information Extraction (OpenIE), several challenges persist that hinder its broader applicability. Extracting relationships involving complex entities or those expressed through intricate linguistic constructions remains challenging. Furthermore, while OpenIE strives to be domain-agnostic, its performance can still vary greatly across different text genres and domains such as legal and health, necessitating ongoing efforts in domain adaptation. As real-time information extraction becomes more prevalent, enhancing the scalability and computational efficiency of OpenIE systems is critical. Additionally, the presence of semantic ambiguities and contextual dependencies in texts poses substantial challenges to extraction accuracy, emphasizing the need for continued research in this area. OpenIE continues to be a vibrant field of research with considerable potential to impact a wide range of applications, and advancements in machine learning—especially in deep and unsupervised learning—are expected to drive further improvements.

3 Related Work

Open Information Extraction (OpenIE) has evolved significantly since its inception, transitioning from rule-based approaches to sophisticated machine learning techniques. This section explores the progression of OpenIE systems, with a particular focus on recent advancements in Large Language Models (LLMs) and their application to multilingual scenarios, especially Portuguese.

Early OpenIE systems relied heavily on syntactic parsing and hand-crafted rules. OpenIE was introduced by Banko et al. [2], which focused on symbolic approaches based on predefined rules to describe each noun and verbal phrase presented in the sentence. Motivated by the distant relation between the noun phrase and its core, the dependency tree approach has emerged as a promising developed system as presented by Del Corro et al. [8]. These systems use predefined linguistic patterns to extract information. Although they are transparent and easy to debug, their effectiveness is limited by the coverage and complexity of the rules.

However, the field has since shifted towards machine learning approaches, which have demonstrated superior performance and adaptability [30, 7, 32, 38].

Zhou and colleagues [39] categorized modern OpenIE systems into two main types: sequence tagging models and generative models.

Sequence tagging models frame OpenIE as a token labeling task, assigning tags to indicate the role of each word in a sentence (e.g., argument, predicate) [39]. These models typically incorporate token embedding, contextual encoders (such as BERT [9]), and label decoders (often using Conditional Random Fields [20]). Notable works in this category include Stanovsky et al.[30], who introduced a Recurrent Neural Network (RNN) architecture for English OpenIE.

The lack of corpora has created a new wave to evolve resources, particularly those with the corpora for languages other than English[30]. As language models have seen significant usage employing neural networks, OpenIE evolved in describing the task and generating triples as prompting, particularly with English languages. Such systems can generalize from training data to extract triples from new texts. Recent approaches have increasingly employed neural networks, particularly transformer models, which have significantly improved the performance of OpenIE systems by capturing contextual and semantic nuances more effectively.

Generative approaches, on the other hand, model OpenIE as a sequence generation problem [7]. These models often employ encoder-decoder architectures to produce a sequence of extractions. Recent studies have integrated BERT embeddings into generative models, as seen in OpenIE6 and IMoJIE [17, 18], which address the issue of redundant extractions in generative OpenIE models.

While most OpenIE research has focused on English, there has been growing interest in developing multilingual and cross-lingual systems. Zhang and colleagues[38] proposed a semi-supervised cross-lingual approach, while Multi2OIE [27] utilized M-BERT for embedding and predicate extraction across multiple languages, including Portuguese and Spanish.

For Portuguese specifically, OpenIE systems have progressed from rule-based dependency analysis [23] and linguistically-oriented patterns [28, 29] to supervised learning with deep neural networks. Recent works like Multi2OIE [27] and PortNOIE [5] have shown significant improvements in F1 scores compared to earlier methods, highlighting the potential of neural network-based approaches for Portuguese OpenIE. This Portuguese method empowers OpenIE-PT for generative approaches layers [4].

The application of LLMs to OpenIE is an emerging trend, although not yet widely adopted. There are instances of LLM use in related fields such as Question Answering, Relation Extraction, and Information Extraction. Xu and others [36] explored the application of an LLM for few-shot relation extraction, while Oppenlaender et al.[24] investigated LLM use for question answering over large-scale text corpora with promising results.

Wei and colleagues [35] examined the use of LLM systems for zero-shot information extraction, proposing to frame it as a multi-step question answering problem. Kolluru et al.[19] investigated the use of Language Models, namely BERT and mT5 [37], for a two-stage generative OpenIE model that first identifies relations and then assembles extractions for each relation.

Recent research has explored various fine-tuning strategies for LLMs to adapt them to specific tasks efficiently. Full Fine-Tuning (FFT) involves updating all model parameters, which can be computationally expensive for large models [13]. In contrast, Parameter-Efficient Fine-Tuning (PEFT) methods aim to reduce the number of trainable parameters while maintaining performance [12].

Low-Rank Adaptation (LoRA), introduced by Hu and colleagues [14], is a PEFT method that has gained popularity due to its efficiency and effectiveness. LoRA adds trainable low-rank matrices to the attention layers of pre-trained models, significantly reducing the number of trainable parameters while achieving comparable performance to full fine-tuning in many tasks [14].

Cross-lingual transfer learning has shown promise in adapting models trained on high-resource languages to low-resource languages [6]. Recent studies have investigated the cross-lingual capabilities of LLMs and their potential for zero-shot and few-shot learning in multilingual settings [21, 15].

Cabral et al. [4] introduced a LoRA finetuned LLM for Portuguese based on LLaMA-2, achieving good results. In the context of OpenIE, cross-lingual transfer could potentially leverage the abundance of English training data to improve performance on languages with fewer resources, such as Portuguese. However, the effectiveness of such transfer learning approaches in OpenIE tasks, particularly when using LLMs, remains an open challenge.

Our work builds upon these foundations, exploring the application of LLMs to Portuguese OpenIE through fine-tuning and LoRA perspectives. We investigate the trade-offs between model size, adaptation technique, and cross-lingual transferability, contributing to the ongoing discourse on efficient fine-tuning strategies for large pre-trained models in multilingual scenarios.

4 Foundations of Large Language Models and Adaptation Techniques

This section explores the fundamentals and emerging technologies that form the backbone of modern natural language processing (NLP) systems. With a special focus on Large Language Models (LLMs), this part of the paper details the technical and theoretical innovations that have driven advancements in the field, as well as strategic adaptations such as Low-Rank Adaptation (LoRA) and fine-tuning, which enhance the functionality of these models in specific applications. The evolution of LLMs, from early n-gram based models to current transformer architectures, reflects significant progress in the ability to simulate human language comprehension and production, leading to notable improvements in efficiency and effectiveness across a variety of NLP tasks [3].

4.1 Large Language Models

Large Language Models are advanced machine learning algorithms designed to simulate the human capacity for understanding and producing natural language. These models are predominantly built using deep learning techniques and are

trained on extensive textual corpora to capture the grammatical, lexical, and semantic complexities of language. Historically, they have evolved from simple n-gram based models to sophisticated neural network architectures, such as recurrent neural networks (RNNs) and more recently, transformers, which utilize attention mechanisms to enhance contextual understanding and text generation [34].

Transformers, in particular, have revolutionized language modeling with their ability to process text sequences in parallel, resulting in significant gains in efficiency and effectiveness in NLP tasks. These models form the foundation for applications such as automated dialogue systems, open information extraction, machine translation, and others that require deep understanding and manipulation of human language [9].

Ongoing research and development in this area aim not only to improve the accuracy of these models but also to make them more accessible and ethical in their use. Recent advancements include the development of multilingual models capable of processing and generating text in multiple languages, which is particularly relevant for our study on Portuguese Open Information Extraction [6].

4.2 Low-Rank Adaptation (LoRA)

LoRA is a model adaptation technique that leverages the principle of matrix decomposition to efficiently modify the weights of a pre-trained model, allowing for adaptation with significantly reduced computational cost. Unlike traditional fine-tuning, which adjusts all model parameters, LoRA focuses on adapting a fraction of these parameters through the addition of low-rank projections [14].

At the core of LoRA is the idea that transformation matrices in language models, such as those found in the attention and feed-forward layers of transformers, can be approximated by products of lower-dimensional matrices. Mathematically, this approach involves introducing two smaller matrices, A and B, where the product AB serves as a low-rank approximation for updating the original weight matrix, W. This product does not replace W but is added to it, allowing the original model to be extended with new learning capabilities without directly altering its pre-existing structure [14].

The primary advantage of using LoRA in language models is twofold: first, it reduces the number of parameters that need to be adjusted during adaptation, decreasing the demand for computational resources and training data. Second, by keeping most of the original model structure unchanged, LoRA preserves the prior knowledge embedded in the model, minimizing the risk of catastrophic forgetting, a common problem in more invasive adaptations [16].

LoRA proves particularly useful in scenarios where computational resources are limited or when it is necessary to adapt models on devices with restricted processing capacity. For these reasons, LoRA is a strategic choice for adapting language models to specific tasks, maintaining a balance between effectiveness and efficiency [14].

4.3 Fine-tuning

Fine-tuning, in the context of large language models (LLMs), refers to the process of adjusting the pre-trained weights of a model on a specific dataset or task, with the aim of adapting the model to perform better in particular scenarios. This method involves continuing the training of the model from its initial pre-trained configuration, typically using a lower learning rate, to refine its parameters without losing the generalizations learned during extensive pre-training [13].

The fine-tuning process involves several critical steps: selection of a training dataset that is representative of the final task, choice of a learning rate that is sufficiently small to avoid losing useful information already acquired, and careful adjustment of the number of training epochs to prevent overfitting. The success of fine-tuning depends not only on the quality and size of the dataset but also on a good regularization strategy and monitoring of the model's generalization [25].

The main advantage of fine-tuning is its ability to produce highly specialized models for specific tasks, which can result in a significant increase in performance compared to generic pre-trained models. However, this method also presents challenges, such as the need for large volumes of task-specific data and the risk of overfitting, especially in tasks with limited datasets [11].

For these reasons, fine-tuning is a powerful and widely used technique in adapting language models for natural language processing (NLP) applications, offering a flexible approach to personalizing artificial intelligence models across a variety of domains and tasks [31].

In the context of our study on Portuguese Open Information Extraction, we explore both LoRA and fine-tuning techniques to adapt large language models, comparing their effectiveness and efficiency in this specific cross-lingual task. This comparative analysis contributes to the ongoing discourse on efficient adaptation strategies for large pre-trained models in multilingual scenarios [1].

5 Methodology

5.1 Corpus

In our study, we employed the TransAlign corpus [26], specifically designed to enhance the availability of high-quality training data for Open Information Extraction (OpenIE) in under-resourced languages. The corpus, developed through a cross-lingual alignment of data from resource-rich languages like English to Portuguese, comprises 96,067 high-quality triples. These triples are aligned to reflect Brazilian Portuguese grammatical structures, using advanced translation models and handcrafted rules.

5.2 Pre-processing

In the initial processing of the TransAlign corpus data, each triple and sentence was transformed into a format close to natural language, using a specific com-

mand system. This transformation aimed to prepare the data for more efficient manipulation in language models.

ChatML, a template format utilizing Jinja[1] syntax, was employed to transform a list of chat messages into a formatted string that can be directly used by language models for training or inference. In the context of TransAlign, this template was adapted to organize the information from the triples (ARG0, V, ARG1) into a format simulating conversations, facilitating the training of models in OpenIE tasks [33].

Axolotl[2], a language model training platform, was used to train the models using the processed data. The choice of Axolotl was due to its efficiency in managing and optimizing the training of large-scale language models, especially in configurations involving complex data adaptations, as is the case with multilingual OpenIE [14].

After applying the ChatML template, the data were converted to the JSONL format. This format is particularly useful for AI model training, as it allows each line of the file to contain a complete JSON object, representing a single data entry. This simplifies data loading and batch processing during training, contributing to the overall efficiency of the process.

5.3 Hyperparameters

We explored the training configurations of decoder-only language models, specifically focusing on Full Fine-Tuning for the Qwen2 0.5 model (0.5B parameters) and LoRA on the same model. The Qwen2 0.5B model was selected due to its compatibility with the training dataset, which includes data in Portuguese, and its smaller size allows for stable training with limited computational resources. The training process was standardized to ensure comparability across different model scales and techniques.

In our experiments, we standardized the configurations for each model, setting the batch size to 64. We adjusted the gradient accumulation settings accordingly to maintain equivalent batch size efficiency between the two models. Specifically, we used four accumulations for the LoRA model and eight accumulations for the Full Fine-Tuning model. This difference in the number of micro-batches, with more frequent accumulations in the Full Fine-Tuning model, was necessary for the lower gradient variability observed during training with LoRA. The learning rate was initialized at 5×10^{-5} and adjusted according to a cosine learning rate scheduler. Our models underwent training over three epochs with a brief warm-up phase of 10 steps to stabilize the learning rate at the beginning of training.

The loss function utilized was the Unsloth cross entropy loss, which enhances the traditional cross-entropy calculation, offering refined loss management crucial for the stability and performance of model training. The optimizer of choice was the Paged AdamW 8bit[3], incorporating 8-bit quantization for optimizer

[1] available at https://jinja.palletsprojects.com
[2] available at https://github.com/OpenAccess-AI-Collective/axolotl
[3] https://github.com/TimDettmers/bitsandbytes

states to reduce memory requirements significantly while preserving training efficacy.

Additional configurations included gradient checkpointing, a critical memory optimization technique that saves only selected intermediate states during the forward pass, thus reducing overall memory consumption. The models were trained using weights with BF16 precision, which strikes an optimal balance between computational performance and numerical precision.

For the LoRA configuration, a rank of 64 was introduced, resulting in 35,192,832 trainable parameters. Furthermore, the following target modules were selected: q_proj, k_proj, v_proj, o_proj, gate_proj, down_proj, up_proj.

5.4 Experimental Setup

Our experiments were conducted on a customer hardware platform consisting of an NVIDIA RTX 3060 GPU with 12GB VRAM, supported by 32GB of system RAM. The software environment was based on Ubuntu, using the Axolotl training tool. The models were handled in the GGUF F16 format, optimized for high-performance computing tasks, which facilitated efficient model loading and inference operations. For model interaction and manipulation, we employed the llama-cpp-python [4] library, which integrates seamlessly with Python, enabling sophisticated data processing and model tuning capabilities. Additionally, Weights & Biases (wandb) was used to collect system metrics, such as GPU power usage, and track training parameters, including loss, during the training process, allowing for the generation of insightful graphs and data collection. This comprehensive setup ensured that our experimental procedures were not only efficient but also reproducible.

5.5 Experiments

Models were trained using the complete TransAlign corpus, with 1% reserved for validation. Performance was evaluated using precision, recall, and F1-score metrics on the PUD 100 dataset, which is highly annotated and reliable [22]. The evaluation code developed by Stanovsky et al. (2018) was employed, based on a lexical matching metric to assess the accuracy of extracted triples [30].

Results will be presented in Table 1, including precision, recall, and F1 metrics for each evaluated model, allowing for a clear comparison of model performance under different configurations and adaptations.

6 Evaluation

This section outlines the methodology used to compare the two models. We assess quantitative differences employing metrics such as F1-score, precision, and recall. For qualitative analysis, we investigate the triples generated by both

[4] https://github.com/abetlen/llama-cpp-python

models to gauge their effectiveness in capturing pertinent information. All model outputs were produced under consistent generation settings: a temperature of 0.2, a top-p of 0.95, a min-p of 0.05, and a top-k of 40. These parameters were chosen to ensure the stability and relevance of the generated content.

6.1 Quantitative Analysis

During the inference phase, the video memory consumption was 1.2 GB for both models, and the token generation rates were similar, at 2500 tokens per second.

Table 1. Evaluation Scores

Modelo	F1	precision	recall
Qwen2 0.5b LoRA	0.2797	0.33	0.2427
Qwen2 0.5b	0.2712	0.32	0.2353

During the training phase, the LoRA model consumed less energy than the Full Fine-Tuning (FFT) model. This is because LoRA reduces the number of trainable parameters by focusing only on smaller, low-rank matrices, leaving the majority of the pre-trained weights untouched. This approach significantly lowers the computational load, resulting in less memory usage and reduced energy consumption, compared to FFT, which updates all model parameters, as illustrated in Figure 1.

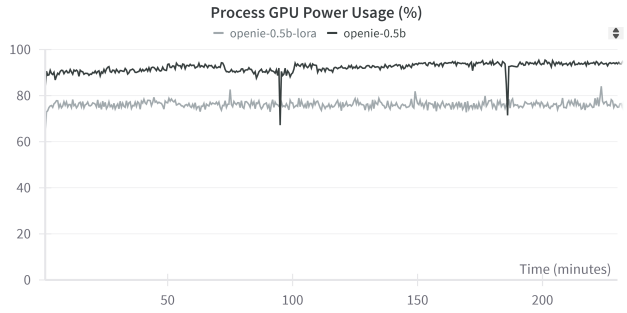

Fig. 1. Graph illustrating the comparative energy consumption of the LoRA and FFT models during the training phase.

It was observed that the model trained with LoRA achieved better results across all metrics compared to the model that underwent full fine-tuning for OpenIE task as it can be observed in Figure 2.

The timing metrics from inference on the sentence "Alan, who studies at UFBA, is a member of Formas," reveal minimal differences between Full Fine-Tuning (FFT) and LoRA. FFT took 228.14 ms, while LoRA slightly improved this to 226.55 ms. Eval times were 4.89 ms per token for FFT and 4.67 ms for LoRA. Prompt eval and sample times were nearly identical for both methods. These minor differences, ranging from 0.17 ms to 5.36 ms, suggest that while LoRA is marginally faster, the impact on practical performance is negligible.

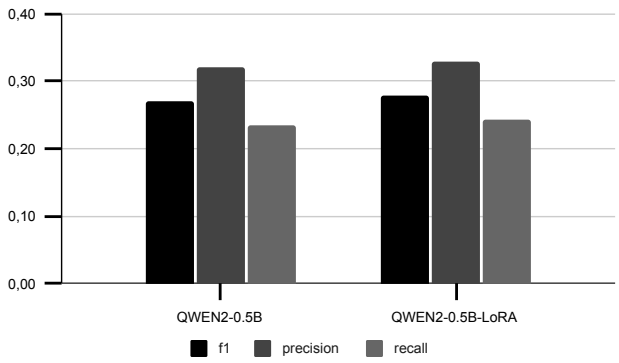

Fig. 2. Graph depicting the differences in metrics between the validated models.

6.2 Qualitative Analysis

Upon analyzing some of the triples generated by both models, a significant degradation in generation quality was noticeable in the model that underwent full fine-tuning, whereas the model trained with LoRA demonstrated greater resistance to hallucinations. The following extractions exemplify this:

Sentence: "Ela começou no RSC no meio dos anos sessenta, atuando como internada de asilo no Marat/Sade."

FFT: *arg1:* Ani - *rel:* começou em - *arg2:* o RSC **Analysis of FFT**: The FFT model hallucinates by incorrectly identifying "Ani" as the subject of the action "começou em". "Ani" does not appear in the sentence, showing a clear error in entity recognition and relation mapping.

LoRA: *arg1*: Ela - *rel:* começou em - *arg2:* o RSC

Analysis of LoRA: The LoRA model accurately identifies "Ela" as the subject who "começou em o RSC", perfectly aligning with the content of the original sentence without adding or distorting information.

Sentence: "O ano passado foi um ano incrível e estou pronto para que possamos voltar ainda melhor em 2017."

FFT: *arg1*: O ano passado - *rel:* estou pronto para que possamos voltar - *arg2:* ainda melhor em 2017

Analysis of FFT: In this extraction, the FFT model creates a hallucinated relation by connecting "O ano passado" with "estou pronto para que possamos voltar", which is not a logical or supported relation within the context of the sentence.

LoRA: *arg1:* O ano - *rel:* passado foi - *arg2:* um ano incrível

Analysis of LoRA: The LoRA model performs a more accurate and textually grounded extraction, identifying that "O ano passado foi um ano incrível", which is a direct and faithful interpretation of what is explicitly mentioned in the sentence.

Sentence: "Após algumas cirurgias geniais e muito trabalho duro de reabilitação, eu tive uma recuperação completa."

FFT Extraction: *arg1*: eu tive uma recuperação completa - *rel*: inicialmente - *arg2*: uma recuperação completa

Analysis of FFT: The FFT model incorrectly incorporates redundancy and an unsupported relation "inicialmente," which distorts the meaning of the sentence. It also repeats "uma recuperação completa" for both subject and object, showing a segmentation error.

LoRA Extraction: *arg1:* eu - *rel:* tive - *arg2:* uma recuperação completa

Analysis of LoRA: The LoRA model's extraction is precise and aligned with the sentence content, correctly identifying "eu" as the subject, "tive" as the verb indicating action, and "uma recuperação completa" as the object, efficiently capturing the sentence's information structure.

These examples illustrate how the LoRA model tends to produce extractions that are more faithful to the original text, avoiding adding non-existent information or distorting relationships between sentence elements, unlike the FFT model, which showed a propensity for hallucinations possibly due to overfitting or insufficient generalization during training.

Acknowledgments This material is partially based work supported by the FAPESB under grant TIC002/2015 and TO CCE 0022/2023.

7 Conclusions and Future Directions

In this work, we evaluate the performance of large language models (LLMs) using Full Fine-tuning and LoRA adaptations, analyzing both quantitative and qualitative measures. We use metrics such as F1, precision, and recall to assess quantitative outcomes. Qualitative analysis is conducted through example sets. Our findings indicate that LoRA adaptations are as effective as Full Fine-tuning, with the added benefit of lower energy consumption and competitive results. We plan to conduct further experiments with a larger language model to deepen our understanding of LLM behavior under these adaptations.

References

1. Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations. In: Jurafsky, D., Chai, J., Schluter, N.,

Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4623–4637. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.421, https://aclanthology.org/2020.acl-main.421
2. Banko, M.: Open Information Extraction for the Web. Ph.D. thesis, University of Washington (2009)
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020)
4. Cabral, B., Claro, D., Souza, M.: Exploring open information extraction for Portuguese using large language models. In: Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H.G., Amaro, R. (eds.) Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. pp. 127–136. Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain (Mar 2024), https://aclanthology.org/2024.propor-1.13
5. Cabral, B., Souza, M., Claro, D.B.: Portnoie: A neural framework for open information extraction for the portuguese language. In: International Conference on Computational Processing of the Portuguese Language. pp. 243–255. Springer (2022)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2020), https://arxiv.org/abs/1911.02116
7. Cui, L., Wei, F., Zhou, M.: Neural open information extraction. arXiv preprint arXiv:1805.04270 (2018)
8. Del Corro, L., Gemulla, R.: Clausie: clause-based open information extraction. In: Proceedings of the 22nd international conference on World Wide Web. pp. 355–366. ACM (2013)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Gozalo-Brizuela, R., Garrido-Merchan, E.C.: Chatgpt is not all you need. a state of the art review of large generative ai models (2023)
11. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. ArXiv **abs/2004.10964** (2020), https://api.semanticscholar.org/CorpusID:216080466
12. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp (2019), https://arxiv.org/abs/1902.00751
13. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification (2018), https://arxiv.org/abs/1801.06146
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
15. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization (2020), https://arxiv.org/abs/2003.11080

16. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (Mar 2017). https://doi.org/10.1073/pnas.1611835114, http://dx.doi.org/10.1073/pnas.1611835114
17. Kolluru, K., Adlakha, V., Aggarwal, S., Chakrabarti, S., et al.: Openie6: Iterative grid labeling and coordination analysis for open information extraction. arXiv preprint arXiv:2010.03147 (2020)
18. Kolluru, K., Aggarwal, S., Rathore, V., Mausam, Chakrabarti, S.: IMoJIE: Iterative memory-based joint open information extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5871–5886. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.521, https://aclanthology.org/2020.acl-main.521
19. Kolluru, K., Mohammed, M., Mittal, S., Chakrabarti, S., ., M.: Alignment-augmented consistent translation for multilingual open information extraction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2502–2517. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.179, https://aclanthology.org/2022.acl-long.179
20. Lafferty, J.D., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (2001), https://api.semanticscholar.org/CorpusID:219683473
21. Lin, X.V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P.S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., Stoyanov, V., Li, X.: Few-shot learning with multilingual language models (2022), https://arxiv.org/abs/2112.10668
22. Nivre, J., de Marneffe, M.C., Ginter, F., Hajic, J., Manning, C.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D.: Universal dependencies v2: An evergrowing multilingual treebank collection. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4034–4043. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.497
23. Oliveira, L., Claro, D.B., Souza, M.: Dptoie: A portuguese open information extraction based on dependency analysis. Artif. Intell. Rev. **56**(7), 7015–7046 (dec 2022). https://doi.org/10.1007/s10462-022-10349-4
24. Oppenlaender, J., Hämäläinen, J.: Mapping the challenges of hci: An application and evaluation of chatgpt and gpt-4 for cost-efficient question answering (2023)
25. Peters, M.E., Ruder, S., Smith, N.A.: To tune or not to tune? adapting pretrained representations to diverse tasks (2019), https://arxiv.org/abs/1903.05987
26. Rios, A., Cabral, B., Claro, D., Cavalcante, R., Souza, M.: TransAlign: An automated corpus generation through cross-linguistic data alignment for open information extraction. In: Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H.G., Amaro, R. (eds.) Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. pp. 196–206. Association for Computational Linguistics, Santiago de Compostela, Galicia/Spain (Mar 2024), https://aclanthology.org/2024.propor-1.20
27. Ro, Y., Lee, Y., Kang, P.: Multi^ 2oie: Multilingual open information extraction based on multi-head attention with bert. arXiv preprint arXiv:2009.08128 (2020)

28. Sena, C.F.L., Claro, D.B.: Inferportoie: A portuguese open information extraction system with inferences. Natural Language Engineering **25**(2), 287–306 (2019). https://doi.org/10.1017/S135132491800044X
29. Sena, C.F.L., Claro, D.B.: Pragmaticoie: A pragmatic open information extraction for portuguese language. Knowl. Inf. Syst. **62**(9), 3811–3836 (sep 2020). https://doi.org/10.1007/s10115-020-01442-7
30. Stanovsky, G., Michael, J., Zettlemoyer, L., Dagan, I.: Supervised open information extraction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 885–895 (2018)
31. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? (2020), https://arxiv.org/abs/1905.05583
32. Sun, M., Li, X., Wang, X., Fan, M., Feng, Y., Li, P.: Logician: a unified end-to-end neural approach for open-domain information extraction. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 556–564. ACM (2018)
33. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762
35. Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., Han, W.: Zero-shot information extraction via chatting with chatgpt (2023)
36. Xu, X., Zhu, Y., Wang, X., Zhang, N.: How to unleash the power of large language models for few-shot relation extraction? (2023)
37. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer (2021)
38. Zhang, S., Duh, K., Van Durme, B.: Mt/ie: Cross-lingual open information extraction with neural sequence-to-sequence models. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 64–70 (2017)
39. Zhou, S., Yu, B., Sun, A., Long, C., Li, J., Yu, H., Sun, J., Li, Y.: A survey on neural open information extraction: Current status and future directions (2022)