Exploring Open Information Extraction for Portuguese Using Large Language Models

Bruno Cabral and **Marlo Souza** and **Daniela Barreiro Claro**
FORMAS - Research Center on Data and Natural Language
Institute of Computing, Federal University of Bahia
Salvador, Bahia - Brazil
{bruno.cabral,msouza,dclaro}@ufba.br

Abstract

In this work, we investigate the potential of Large Language Models (LLMs) for Open Information Extraction (OpenIE) in the Portuguese language. While most OpenIE methods are primarily optimized for English, only few works in the literature explore their uses for cross-lingual and multilingual scenarios. Despite the growing interest in Portuguese OpenIE methods, the use LLMs for Portuguese focused OpenIE is still an underdeveloped topic in the area.Our study addresses this research gap by examining the viability of using open and commercial LLMs with few-shot prompt engineering for Portuguese OpenIE. We provide an analysis of the performance of these LLMs in OpenIE tasks, revealing that they achieve performance metrics comparable to state-of-the-art systems. In addition, we have fine-tuned and launched an open LLM for OpenIE (PortOIE-Llama), which outperforms commercial LLMs in our experiments. Our findings highlight the potential of LLMs in Portuguese OpenIE tasks and suggest that further refinement and fine-tuning of larger models could enhance these results.

1 Introduction

The digital era is characterized by the exponential growth of data, a large part of which is unstructured text data from various sources such as books, blogs, articles, and more. Extracting valuable information from this vast data pool is a critical task; however, the challenge lies in uncovering relevant information embedded within the vast amount of data. Open Information Extraction (OpenIE) offers a solution to extract knowledge from extensive text collections, regardless of the domain (Banko et al., 2007; Batista et al., 2013).

Recent years have witnessed substantial advancements in generative AI models, particularly in large-scale language models like GPT-3 (Brown et al., 2020), spurred by the exponential growth of data availability and computational power needed for processing it (Gozalo-Brizuela and Garrido-Merchan, 2023). A significant advancement has been witnessed in Natural Language Processing (NLP) and digital image generation, capturing the attention of numerous individuals. A prime example of this growth is the rapid success of ChatGPT, which achieved the title of "Fastest Growing App of All Time" by amassing 100 million monthly active users within just two months[1].

Open Information Extraction (OpenIE) systems generate a structured representation of the information present in the original documents, typically in the form of relational tuples, for instance, (arg_1, rel, arg_2), where arg_1 and arg_2 are the arguments of the relation, usually described by noun phrases, and rel is a relation descriptor that describes the semantic relation between arg_1 and arg_2 (Gamallo, 2014).

For the Portuguese language, OpenIE has seen significant advancements in the last few years, although the application of Large Language Models (LLMs) remains relatively unexplored. Despite this, LLMs have shown capabilities in understanding and generating text that closely resembles human-like text, indicating a promising path for OpenIE tasks. This study aims to bridge this gap by examining the potential of both commercial and open-source LLMs when applied to Portuguese OpenIE, utilizing few-shot prompt engineering.

Our primary contribution revolves around an investigation into the potential of LLMs for OpenIE tasks in Portuguese. This contribution includes a comprehensive analysis of their performance and an evaluation of their ability to handle the complexities of the Portuguese language and their adapt-

[1] ChatGPT sets record for fastest-growing user base - analyst note, www.reuters.com/technology/chatgpt%2Dsets-record%2Dfastest-growing-user-base%2Danalyst-note-2023-02-01/ accessed November 5, 2023

ability to OpenIE tasks. Furthermore, we introduce and publicly release a fine-tuned LLM for OpenIE (PortOIE-Llama). We also examine how those LLMs compete against current OpenIE state-of-the-art systems for Portuguese.

This paper is structured as follows: Section 2 reviews the related work, Section 3 outlines the methodology and approach employed, Section 4 presents our experiments, results, and discussions, and Section 5 concludes our findings and discusses future research directions.

2 Related Work

The introduction of machine learning-based methodologies has indicated a new era for Open Information Extraction (OpenIE) systems (Stanovsky et al., 2018; Cui et al., 2018; Sun et al., 2018; Zhang et al., 2017). However, most of these systems have a particular emphasis on the English language (Claro et al., 2019), and their considerable dependence on annotated data presents substantial difficulties when extending them to other languages.

Various researchers, including Stanovsky et al. (2018), have proposed tagging-based models for OpenIE, viewing OpenIE as a sequence labeling task akin to Named Entity Recognition (NER). Since 2020, several works have employed Transformer architectures directly or in conjunction with BERT embedding (Devlin et al., 2018), such as that of Hohenecker et al. (2020), who analyzed various neural-based OpenIE architectures and introduced an ALBERT embedding block model.

Conversely, generative approaches to OpenIE model it as a sequence generation problem that produces a sequence of extractions (Cui et al., 2018). Authors, such as Cui et al. (2018) and Zhang et al. (2017), have also explored this approach, employing an encoder-decoder framework to learn high-confidence arguments and relation tuples bootstrapped from an OpenIE system. Contemporary studies have integrated BERT embeddings into their generative models. For instance, Kolluru et al. (2020a,b) launched OpenIE6 and IMoJIE, respectively, for the English language, employing a BERT encoder and an LSTM decoder to address the issue of redundant extractions in generative OpenIE models.

OpenIE systems for the Portuguese language have evolved, transitioning from rule-based dependency parsing (Oliveira et al., 2022) and linguistically driven patterns (Sena and Claro, 2019, 2020) to recent applications of supervised learning with deep neural networks, as seen in works like Multi2OIE (Ro et al., 2020) and PortNOIE (Cabral et al., 2022). These latter studies have shown significant enhancements in the F1 score compared to prior methods, corroborating the potential of neural network-based approaches for Portuguese OpenIE.

Applying Large Language Models (LLMs) for OpenIE is an emerging trend, albeit yet to be widely adopted. There are instances of use in related fields, such as Question Answering, Relation Extraction, and Information Extraction. Xu et al. (2023) explored the application of an LLM for few-shot relation extraction. Oppenlaender and Hämäläinen (2023), on the other hand, investigated the application of an LLM for question answering over a text corpus at scale with promising outcomes. Wei et al. (2023b) examined the use of LLMs system for zero-shot information extraction, proposing to frame it as a multi-turn question-answering problem. Lastly, Kolluru et al. (2022) investigated the use of Language Models, namely BERT and mT5 (Xue et al., 2021) for a two-stage generative OpenIE model, that initially identifies relations and then assembles the extractions for each relation.

In our approach, we explored open-source and commercial LLMs techniques, such as few-shot and prompt engineering, to assess their viability for Portuguese OpenIE. To our knowledge, this is the first work to assess the applicability of LLMs for OpenIE in the Portuguese language.

3 PT-OpenIE pipeline for LLMs

In this section, we describe our pipeline for PT-OpenIE, introducing our definition of Open Information Extraction (OpenIE) and guiding the pipeline for PT-OpenIE. Our pipeline describes each model to assess the performance of Large Language Models (LLMs) as triple extractors for the Portuguese language.

3.1 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. For simplicity, as is common in the area, we do not consider extractions consisting of n-ary relations.

3.2 Model Selection

We evaluate both open and commercial Large Language Models (LLMs). To select the best performing models at the time of writing (October 2023), we used the Chatbot Arena Leaderboard (Zheng et al., 2023). The top-performing models included OpenAI GPT-4 (OpenAI, 2023), Anthropic Claude-v1 (Anthropic, 2023), and OpenAI GPT-3.5-turbo (Brown et al., 2020), all of which are commercial models.

Using these models is only possible through a private REST API with a high cost for each call. However, we also wanted to compare the performance of open-source models. These models provide complete access and can be utilized locally.

On the Chatbot Arena Leaderboard, there are multiple fine-tuned open-source models from three foundational LLMs: LLaMA (Touvron et al., 2023a), LLaMA2 (Touvron et al., 2023b) and Falcon (Almazrouei et al., 2023). Foundational LLMs are base language models trained on a large corpus of text from the internet but without any task-specific data. These models learn to predict the next word in a sentence, which allows them to generate human-like text based on the input.

On the other hand, commercial models achieve good performance due to many factors, one of which may be alignment tuning. This process aims to obtain language models consistent with human expectations. For instance, the GPT-4 has been trained on a dataset of instructions. It has also undergone Reinforcement Learning from Human Feedback to better align with human preferences (Ouyang et al., 2022).

We chose the fine-tuned Falcon-40B and LLaMA-65B models based on the OpenAssistant Conversations Dataset (Köpf et al., 2023) (OASST1), a human-annotated assistant-style conversation corpus with 161,443 messages. We selected this dataset due to its manual annotation, permissive license, and the inclusion of instructions in Portuguese. We also picked the LLaMA2-chat (Touvron et al., 2023b) with 7B and 70B, a successor to the LLaMA model, which was the fined-tuned model on instructions by the original team.

Table 1 summarizes the models used in this study.

3.3 Model Fine-tune

We employed Low-Rank Adaptation (LoRA) (Hu et al., 2021), an efficient technique for fine-tuning a Large Language Model (LLM). Training a foundational model is often an unattainable task for many due to its prohibitive costs. While pre-training is less costly, it remains within reach only for those with substantial resources. LoRA provides a solution to this challenge by representing model updates as low-rank factorizations, significantly reducing the size of update matrices, and enabling model fine-tuning at a fraction of the cost and time (Hu et al., 2021).

We employ three distinct, human-annotated datasets as training dataset for the finetune, which are described as follows:

- **Pragmatic** - This dataset, derived from Sena and Claro (2020), is human-annotated and consists of news data. They were composed of 400 sentences with 485 extractions.
- **Gamalho** - This dataset contains relations extracted from Portuguese texts using five different OpenIE systems validated by human experts. The linguistic resources were obtained from the studies by(Del Corro and Gemulla, 2013) and (Gamallo and Garcia, 2015). It is composed of 103 sentences with 346 extractions.
- **PUD 200** - These datasets include sentences sourced from news and the Portuguese portion of the Parallel Universal Dependencies (PUD) corpus (Nivre et al., 2020), as well as Wikipedia, with 337 extractions from 200 sentences.

We converted the sentences and extractions into the Alpaca (Taori et al., 2023) format, which consists of an instruction, an input, and an output. We translate the instruction into English: *"Given a sentence S, you make extractions in the ARG0, V, ARG1 format. Perform the extraction for the sentence below:"*. The input is the quoted sentence preceded by "S:", and the output uses the template: Extraction N: ARG0= " ..." V= "" ARG1= " ...".

We provide a real example of a sentence with its extraction from the PUD 200 dataset:

- **instruction**: Dada uma frase S você consegue fazer extrações no formato ARG0 , V, ARG1. Realize a extração para a frase abaixo:

- **input**: S: "Organismos que vivem em biomas marinhos devem estar adaptados a o sal presente em a água ."

- **output**: Extração 0: ARG0="Organismos" V="vivem em" ARG1="biomas marinhos"

We operated the Llama 2-Chat model (Touvron et al., 2023b) with 7B parameters as the basis for our fine-tuning. The training was conducted over two epochs, with a starting learning rate of 0.0002, which was reduced using a cosine scheduler. The LoRA parameters used were R of 32, Alpha of 16, and a dropout of 0.05. The fine-tuning was performed on the same cloud server, described in detail in the following section, at a cost of USD 0.65. This model is publicly available at HuggingFace[2].

Table 1: Summary of the utilized LLM models

Name	License	Model Size
OpenAI GPT-4	Commercial	N/A
Anthropic Claude-v1	Commercial	N/A
OpenAI GPT-3.5	Commercial	N/A
LLaMA-OASST1	Non commercial	65B
Falcon-OASST1	Apache 2	40B
LLaMA-2-Chat	Non commercial	7/70B

3.4 Dataset

The primary dataset used for evaluation is the *PUD 100*, a golden set based on the *PUD 200* dataset(Cabral et al., 2022). It is the second iteration of the dataset creation methodology employed for the creation of *PUD 200* utilized in our fine-tuning and is considered higher quality than it.

A team of academic annotators, experts in OpenIE, annotated the source dataset, that consists of sentences from news sources and Wikipedia of the Portuguese part of the Parallel Universal Dependencies (PUD) corpus (Nivre et al., 2020). It is composed of 100 sentences and 136 extractions. Although the dataset is relatively small, it is highly diverse and complex, presenting a wide range of linguistic phenomena. This complexity is beneficial for our study as it allows us to assess the robustness of the models in handling various linguistic phenomena.

An example of this dataset is the following: **Sentence:** *Todos os médicos estavam armados, exceto eu.*. It can be translated as: *All the doctors were armed except me.*. Extractions in Portuguese and English are: Extraction 1: **ARG0=** O vestido **V=**é **ARG1=**contemporâneo (**ARG0=**The dress **V=**is **ARG1=**contemporary)

3.5 Prompt Engineering

Our methodology for deriving the optimal research prompt for the OpenIE task was a systematic, iterative process. We began by focusing on the first five sentences of the PUD 200 dataset, adjusting our prompt until it was able to correctly generate these sentences. It's important to note that these sentences were not used for evaluation or as few-shot examples in the prompt, but rather as a benchmark for the iterative refinement of our prompt.

Initially, we started with a simple prompt with the request to perform OpenIE extractions of a sentence. However, this naive approach did not yield satisfactory results, indicating that the model needed more explicit instructions to understand the task.

To improve the model's comprehension, we incorporated an extraction example into the prompt. This modification significantly enhanced the model's understanding of the task. We further refined the prompt by including a system definition, stating, "You are a very smart and accurate OpenIE...", this adjustment proved beneficial even in a 1-shot scenario, indicating that the model responded well to explicit role definitions.

Despite these improvements, we found that incorporating a comprehensive definition, such as the Wikipedia definition of the OpenIE task within the prompt led to unsatisfactory results. Finally, we formatted the examples in the key-value format with line breaks. This adjustment made the model's responses less conversational (e.g., "Yes, I can do an extraction...") and more structured, which was easier to parse.

After multiple rounds of adjustments, we finalized the following prompt:

> Você é um sistema muito inteligente e preciso de Extração de informação aberta. Dada uma frase S, você consegue fazer extrações no formato ARG0, V, ARG1, como por exemplo:
>
> S: “Maria é Professora de Banco de Dados”
>
> Extração 1:
>
> ARG0='Maria'

[2] https://huggingface.co/bratao/llama7b-finetuned-openie-lora

V='é'

ARG1='Professora de Banco de Dados'

[Few-Shot Examples]

Realize a extração para a frase abaixo:

S: **[SENTENCE]**

In this prompt, **[SENTENCE]** represents the sentence to be processed, and **[Few-Shot Examples]** are a few instances of OpenIE extractions in the context of the sentence. This approach is known as Few-shot Learning, where the Language Model is provided with a small number of example OpenIE extractions to facilitate its understanding of the task (Wang et al., 2020).

We also explored other prompts and techniques, such as Chain-of-Thought (Wei et al., 2023a) prompting. However, in our experiments, we found that the prompt mentioned above consistently produced outputs that met our expectations for this task. As a result, new attempts to understand the OpenIE task are encouraged; thus, leveraging prompt engineering on LLMs can be tackled as an open problem for PT-OpenIE.

3.6 Limitations

The first limitation of our approach is the relatively compact *PUD 100* dataset, which may circumscribe the broad applicability of our conclusions. Additionally, the quality of the prompt influences the efficacy of LLMs. As aforementioned, varying prompts could yield diverse outcomes. Concerning the task, the binary relation extraction framework we employed may not fully capture the complexity of some sentences. Moreover, the dataset employed rises only on binary relations. Lastly, LLMs can be subject to intrinsic biases in the training data, potentially affecting the quality and fairness of OpenIE tasks.

4 Experiments

We detail our empirical validation for extracting PT-OpenIE triples.

4.1 Experimental Design

Each pipeline stage was implemented in Python 3.10, leveraging the OpenAI and Anthropic libraries to utilize their Large Language Models (LLMs). For open-source local LLMs, we used the Llama.cpp project (Gerganov, 2023) to load the LLMs and predict the outputs.

The *temperature* of the model was set to 0.2, with *max_tokens* at 1000. We also set *top_p*, *frequency_penalty*, and *presence_penalty* to 0, ensuring no penalty is applied to tokens appearing multiple times in the output.

All local models were executed on a cloud server powered by an AMD EPYC 7003 with 30 vCPU, NVIDIA A100 GPU with 40GB of VRAM memory, and 200GB of RAM. This server was hosted on a cloud provider at a cost of $1.10/hr.

For each sentence in the *PUD 100* Dataset, we tokenized the sentence and fed the tokenized sentence to each LLM, along with the prompt described in the Prompt Engineering subsection. Models generated text outputs, which we parsed to extract the triples. The **LLaMA-2-7B-FT** is our fined-tuned model, and as it was explicitly fine-tuned for this task, it uses a custom prompt that is the same as it was trained on. For this reason, the few-shots prompt strategy was not used for this model.

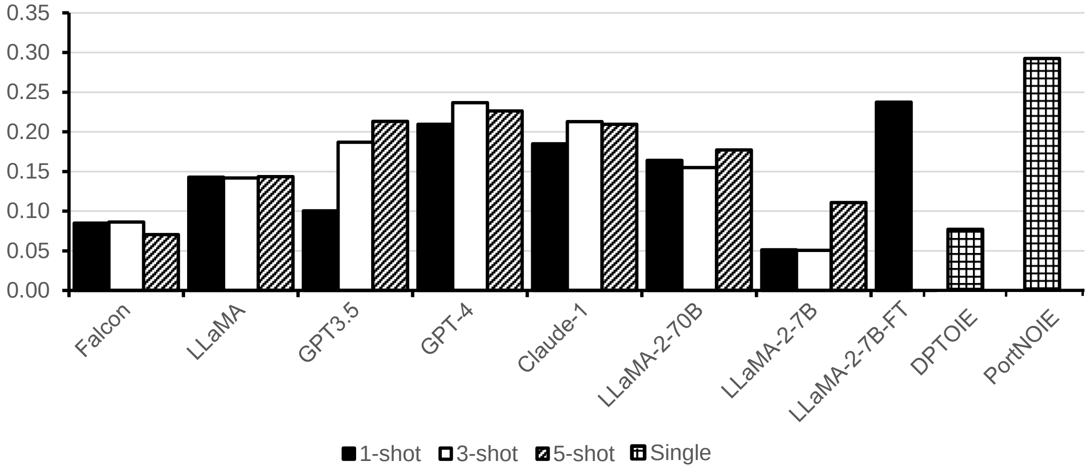
We reviewed existing Portuguese OpenIE systems for comparison, selecting DptOIE(Oliveira et al., 2022) and PortNOIE (Cabral et al., 2022). DptOIE employs Depth-First Search on the Dependency Tree for triple extraction, while PortNOIE is a deep neural network that claims to have achieved the best F1 metric result for Portuguese.

We used precision (P), recall (R), and the F1 measure to evaluate our extractor's quality. We adapted the evaluation code provided by Stanovsky et al. (2018), widely used in subsequent works (Ro et al., 2020; Kolluru et al., 2020a). By default, their benchmark uses a scoring method named **Lexical match**, which considers triple words as a match if they are at least 50% the same, regardless of the order. We also compared them using the **Perfect match** strategy, which considers the strings identical after removing punctuation.

These metrics compare the triples extracted by each model with the gold standard triples in the *PUD 100* Dataset. An exact match with the gold standard triple was considered a match. For lexical match, we used a relaxed matching strategy, considering a match if at least two components of the triple (arg1, rel, arg2) matched with the gold standard.

4.2 Results

The analysis of the results is organized into two parts. First, we present the F1 scores for perfect and lexical matches across different models using 1-shot, 3-shot, and 5-shot prompting strategies. Af-



	Perfect Match F1 ↑			Lexical Match F1 ↑		
Models	**1-shot**	**3-shot**	**5-shot**	**1-shot**	**3-shot**	**5-shot**
Falcon	0.0338	0.0344	0.0	0.0847	0.0862	0.0703
LLaMA	0.0380	0.0516	0.0603	0.1428	0.1419	0.1434
GPT3.5	0.0301	0.1007	0.0955	0.1005	0.1870	0.2132
GPT-4	0.1013	0.1065	0.0978	0.2094	0.2366	0.2262
Claude-1	0.0711	0.0972	0.0878	0.1850	0.2127	0.2094
LLaMA-2-70B	0.0447	0.0619	0.0655	0.1641	0.1547	0.1770
LLaMA-2-7B	0.0255	0.0144	0.0158	0.0510	0.0505	0.1106
LLaMA-2-7B-FT (PortOIE-Llama)	**0.1271**	N/A	N/A	**0.2372**	N/A	N/A

Table 2: F1 Measures of Different Models for PUD100 dataset using for 1,3 and 5-shots prompting for Perfect and Lexical Match

terward, a detailed performance analysis of the models using a 3-shot strategy on the *PUD 100* dataset. The results are presented in Table 2 and Table 3, respectively, with a visual comparison of the F1 scores of the different models using the best prompting strategy.

Considering only the LLMs in the **Perfect Match** scenario, the LLaMA-2-7B-FT model, our finetuned version of the LLaMA-2-7B model, called PortOIE-Llama, outperforms other models in all scenarios with a score of 0.1271 as shown in Table 2. The original model, the LLaMA-2-7B, performs considerably worse, with the highest F1 of 0.0255, a 5-fold performance increase. This finetuned model is better than the second-best model, the commercial GPT-4 model, with scores of 0.1013, 0.1065, and 0.0978, respectively.

Our results indicate a somewhat unexpected trend: the 5-shot prompting strategy was not better as prompts with fewer examples. The performance of the Falcon model plummets dramatically in the 5-shot scenario, reaching scores near 0. This outcome defies the conventional expectation that more prompts would lead to better performance. For instance, the GPT-4 model, which excelled in the 1-shot and 3-shot scenarios with scores of 0.1013 and 0.1065, respectively, saw a slight dip in performance in the 5-shot scenario. This provides valuable insight into optimizing prompting strategies, demonstrating that more prompts do not necessarily equate to better performance.

In the detailed performance analysis using the best-shot performance for each model on the Lexical Match PUD100 dataset (Table 3), the PortNOIE model exhibits the highest precision score of 0.3269 and the highest F1 score of 0.2905. It also has the lowest cost of 1k and the shortest average prediction time, making it the most efficient model. For the LLMs, the LLaMA-2-7B-FT model, although having the second-highest F1 score of 0.2372, comes with a significantly lower cost and a shorter prediction time compared to the GPT-4

Table 3: F1 Measures of Different Models for PUD100 dataset using the best performing prompting strategy for Lexical Match

Model	**Precision ↑**	**Recall ↑**	**F1 ↑**	**1k Cost ↓**	**Avg Pred. Time ↓**
Falcon(3-shot)	0.1041	0.0735	0.0862	$1.16	3.8 seconds
LLaMA(5-shot)	0.1472	0.1397	0.1433	$1.10	3.6 seconds
GPT3.5(5-shot)	0.2132	0.2132	0.2132	$1.20	2.7 seconds
GPT-4(3-shot)	0.1980	0.2941	0.2366	$36.80	4.2 seconds
Claude-1(3-shot)	0.1813	0.2573	0.2127	$3.20	3.5 seconds
LLaMA-2-70B(5-shot)	0.1597	0.1985	0.1770	$1.16	3.8 seconds
LLaMA-2-7B(5-shot)	0.1196	0.1029	0.1106	$0.45	1.5 seconds
LLaMA-2-7B-FT (PortOIE-Llama)	0.28	0.2058	0.2372	$0.45	1.5 seconds
DPTOIE	0.0408	0.0787	0.0757	$1.62	5.3 seconds
PortNOIE	**0.3269**	0.2615	**0.2905**	**$0.15**	**0.5 seconds**

model.

The LLaMA-2-7B-FT model, our fine-tuned version of the LLaMA-2-7B model (PortOIE-Llama), was the best performing LLM, followed by GPT-4 a commercial LLM model in both perfect and lexical match scenarios. Nevertheless, considering cost-effectiveness and efficiency, the PortNOIE model stands out in the 3-shot strategy on the PUD100 dataset. This suggests that while large language models like GPT-4 can provide good performance, other factors, such as cost and speed, must also be considered in practical applications.

4.3 Qualitative Analysis

We evaluate the qualitative results of multiple large language models on the PT-OpenIE task. The input sentence was: "Teoricamente, um casal poderia abrir quatro contas Tesco e ganhar 3% em £12,000 – £360", and the expected output was "ARG0='um casal' V='poderia abrir' ARG1='quatro contas Tesco'".

Falcon: The extraction for this system was not as expected. It seems to have misunderstood the sentence structure and extracted the currency amount as the argument, which is not the main focus of the sentence. The output was:

ARG0="Teoricamente, um casal poderia abrir quatro contas Tesco e ganhar 3% em £12,000 – £360." V="em" ARG1="£12,000 – £360."

LLaMA: This system performed well and produced the expected output. It correctly identified the main entities and relations in the sentence. The output was:

ARG0 = "um casal" V = "poderia abrir" ARG1 = "quatro contas Tesco"

GPT3.5: This system was not able to perform the task, as it is not designed for open information extraction. It suggested using other tools or libraries for this purpose.

GPT-4: This system performed well, but it also extracted additional information that was not part of the expected output. The output was:

ARG0="um casal" V="poderia abrir" ARG1="quatro contas Tesco" ARG2="e ganhar" ARG3="3% em £12,000 – £360"

Claude-1: This system seems to have misunderstood the sentence structure, similar to Falcon. The output was:

ARG0 ="Teoricamente," V ="um casal" ARG1="poderia abrir quatro contas Tesco e ganhar 3% em £12,000 – £360. "

LLaMA-2-70B: This system performed well and produced two extractions from the sentence, both of which are relevant. The output was:

Extração 1: ARG0="um casal" V="poderia" ARG1="abrir quatro contas Tesco" Extração 2: ARG0="ganhar" V="3%" ARG1="£12,000 – £360"

LLaMA-2-7B: This system seems to have misunderstood the task and asked for more input instead of providing an extraction.

LLaMA-2-7B-FT: This system performed well and produced the expected output. The output was:

Extração 0: ARG0="um casal" V="poderia *abrir" ARG1="quatro contas Tesco"*

In summary, LLaMA, GPT-4, LLaMA-2-70B, and LLaMA-2-7B-FT were able to extract the expected triples, while Falcon and Claude-1 had difficulties with the sentence structure. GPT3.5 and LLaMA-2-7B were not able to perform the task. The LLaMA-2-7B performed significantly worse than the 70B version, demonstrating that the LLM size was a considerable factor for this problem.

5 Conclusion and Future Work

This research explored the efficacy of Large Language Models (LLMs) in the context of Open Information Extraction (OpenIE) for Portuguese. We conducted experiments by employing diverse prompting strategies and comparing the performance of several models, namely GPT-4, GPT3.5, LLaMA, Falcon, Claude-1, and LLaMA-2, against established Portuguese OpenIE systems such as DptOIE and PortNOIE. Additionally, a fine-tuned LLM based on LLaMA-2 7B, from now on called PortOIE-Llama, was developed and evaluated.

The results revealed that our fine-tuned LLM (PortOIE-Llama) consistently outperformed other LLMs in F1 scores under both perfect and lexical match scenarios, surpassing the larger commercial LLM, GPT-4.

However, despite the high F1 scores achieved by the LLMs, they remain resource-intensive. Furthermore, PortNOIE demonstrated superior performance not only in terms of performance metrics but also in cost-effectiveness and speed of predictions, achieving the highest precision score, the highest F1 score, the lowest cost for 1k predictions, and the shortest average prediction time. This suggests that while LLMs like GPT-4 and LLaMA can offer remarkable performance, a specialized model remains the optimal choice for OpenIE in Portuguese.

The LLaMA-2 model, fine-tuned on OpenIE, our PortOIE-Llama, exhibits significant potential for future exploration despite being developed under many constraints. It was fine-tuned using a dataset with a limited number of Portuguese examples, and the original LLaMA-2 model is not optimized for Portuguese as the majority of its dataset is in English. Furthermore, we employed an efficient fine-tuning technique, Low-Rank Adaptation (LoRA) (Hu et al., 2021), which, while enabling the creation of such a model with limited resources, only trains a small percentage of the original LLM. It is reasonable to anticipate that fine-tuning with more data, using a larger LLM that better understands Portuguese and is specifically designed for the OpenIE task could yield superior results.

In conclusion, this work contributes to the understanding of Large Language Models' application in OpenIE for Portuguese. The findings of this research have practical implications for creating efficient and cost-effective OpenIE systems for Portuguese. Future research could explore optimizing using various prompting strategies and evaluate these models' performance on larger and more diverse datasets. This model is publicly available at HuggingFace Models[3]. The data and code are available at *https://github.com/FORMAS/openie_generative*.

Acknowledgments

This material is partially based upon work supported by the FAPESB under grant INCITE PIE0002/2022. This material is partially supported by the FAPESB TIC 0002/2015. This material is partially based upon work supported by CAPES Financial code 001.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Anthropic. 2023. Meet claude. `https://www.anthropic.com/product`. Accessed: 2023-04-03.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction for the web. In IJCAI, volume 7, pages 2670–2676.

David Soares Batista, David Forte, Rui Silva, Bruno Martins, and Mário Silva. 2013. Extracçao de relaçoes semânticas de textos em português explorando a dbpédia e a wikipédia. Linguamatica, 5(1):41–57.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric

[3] https://huggingface.co/bratao/llama7b-finetuned-openie-lora

Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Bruno Cabral, Marlo Souza, and Daniela Barreiro Claro. 2022. Portnoie: A neural framework for open information extraction for the portuguese language. In International Conference on Computational Processing of the Portuguese Language, pages 243–255. Springer.

D.B. Claro, M. Souza, C. Castellã Xavier, and L. Oliveira. 2019. Multilingual open information extraction: Challenges and opportunities. Information, 10(7):228.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. arXiv preprint arXiv:1805.04270.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In Proceedings of the 22nd international conference on World Wide Web, pages 355–366. ACM.

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

Pablo Gamallo. 2014. An Overview of Open Information Extraction (Invited talk). In 3rd Symposium on Languages, Applications and Technologies, volume 38 of OpenAccess Series in Informatics (OASIcs), pages 13–16, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Pablo Gamallo and Marcos Garcia. 2015. Multilingual open information extraction. In Portuguese Conference on Artificial Intelligence, pages 711–722. Springer.

Georgi Gerganov. 2023. llama.cpp. `https://github.com/ggerganov/llama.cpp`. GitHub repository.

Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchan. 2023. Chatgpt is not all you need. a state of the art review of large generative ai models.

Patrick Hohenecker, Frank Mtumbuka, Vid Kocijan, and Thomas Lukasiewicz. 2020. Systematic comparison of neural architectures and training approaches for open information extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8554–8565.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Soumen Chakrabarti, et al. 2020a. Openie6: Iterative grid labeling and coordination analysis for open information extraction. arXiv preprint arXiv:2010.03147.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020b. IMoJIE: Iterative memory-based joint open information extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5871–5886, Online. Association for Computational Linguistics.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . 2022. 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), pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 4034–4043, Marseille, France. European Language Resources Association.

Leandro Oliveira, Daniela Barreiro Claro, and Marlo Souza. 2022. Dptoie: A portuguese open information extraction based on dependency analysis. Artif. Intell. Rev., 56(7):7015–7046.

OpenAI. 2023. Gpt-4 technical report.

Jonas Oppenlaender and Joonas Hämäläinen. 2023. Mapping the challenges of hci: An application and evaluation of chatgpt and gpt-4 for cost-efficient question answering.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Youngbin Ro, Yukyung Lee, and Pilsung Kang. 2020. Multiˆ 2oie: Multilingual open information extraction based on multi-head attention with bert. arXiv preprint arXiv:2009.08128.

Cleiton Fernando Lima Sena and Daniela Barreiro Claro. 2019. Inferportoie: A portuguese open information extraction system with inferences. Natural Language Engineering, 25(2):287–306.

Cleiton Fernando Lima Sena and Daniela Barreiro Claro. 2020. Pragmaticoie: A pragmatic open information extraction for portuguese language. Knowl. Inf. Syst., 62(9):3811–3836.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. 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), pages 885–895.

Mingming Sun, Xu Li, Xin Wang, Miao Fan, Yue Feng, and Ping Li. 2018. 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, pages 556–564. ACM.

Rohan Taori, Ishaan Gulrajani, Tianhao Zhang, Yves Dubois, Xiang Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023b. Zero-shot information extraction via chatting with chatgpt.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction?

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Sheng Zhang, Kevin Duh, and Benjamin Van Durme. 2017. 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, pages 64–70.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.