The contribution of LLMs to relation extraction in the economic field

Mohamed Ettaleb, Véronique Moriceau, Mouna Kamel, Nathalie Aussenac-Gilles

The contribution of LLMs to relation extraction in the economic field

Mohamed Ettaleb[1], Véronique Moriceau[1], Mouna Kamel[1,2], Nathalie Aussenac-Gilles[1],
[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse
[2]Espace-Dev, Université de Perpignan
{Mohamed.Ettaleb, Veronique.Moriceau, Mouna.Kamel, Nathalie.Aussenac-gilles}@irit.fr

Abstract

Relation Extraction (RE) is a fundamental task in natural language processing, aimed at deducing semantic relationships between entities in a text. Traditional supervised relation extraction methods involve training models to annotate tokens representing entity mentions, followed by predicting the relationship between these entities. However, recent advancements have transformed this task into a sequence-to-sequence problem. This involves converting relationships between entities into target strings, which are then generated from the input text. Thus, language models now appear as a solution to this task and have already been used in numerous studies, with various levels of refinement, across different domains.
The objective of the present study is to evaluate the contribution of large language models (LLM) to the task of relation extraction in a specific domain (in this case, the economic domain), compared to smaller language models. To do this, we considered as a baseline a model based on the BERT architecture, trained in this domain, and four LLM, namely FinGPT specific to the financial domain, XLNet, ChatGLM, and Llama3, which are generalists. All these models were evaluated on the same extraction task, with zero-shot for the general-purpose LLM, as well as refinements through few-shot learning and fine-tuning. The experiments showed that the best performance in terms of F-score was achieved with fine-tuned LLM, with Llama3 achieving the highest performance.

1 Introduction

The goal of relation extraction (RE) task is to identify and classify relationships between entities in unstructured texts. In domain-specific fields like economic[1], this task is particularly challenging due to the complexity and diversity of linguistic expressions, as well as the presence of domain-specific terminology. Extracting meaningful domain relations from documents requires models that can handle the inherent ambiguities and varied structures present in texts.

Over the past decade, deep learning has led to significant advancements in RE tasks. Pretrained models like BERT (Devlin, 2018) and T5 (Raffel et al., 2020) have been extensively applied to general relation extraction, showing impressive results. In more specialized domains, models like GPT-FinRE (Rajpoot and Parikh, 2023) leverage OpenAI's models within an In-Context Learning (ICL) framework and use retrieval mechanisms to extract domain relations. Although these models exhibit great potential, directly using them for domain-specific tasks can lead to suboptimal performance. This is primarily due to their limited ability to fully perceive internal relationships, especially when entity mentions are ambiguous or when the sentence structures are highly complex, which is the case in many specific domains. The arrival of LLM represented a further step forward for NLP, and consequently for the task of extracting relations (Xu et al., 2023).

However, research has shown that using LLM does not result in significant performance gains compared with small models, particularly in the task of extracting relationships, which is similar to a classification problem (Lepagnol et al., 2024). A way to improve LLM performances for the RE task on specific domains is to refine them. Two techniques at least have proved their worth: few-shot learning and fine-tuning. The first one needs a simple set of prompts, while the second one, which is more costly, requires an annotated dataset and important computational resources.
The key research questions we aim to address in this paper are the following:

[1]In this paper, the term economic is used to encompass both the finance and business domains.

- whether and how can large models perform better than smaller models for relation extraction in the economic domain where entities hold rich and diverse information (e.g. a company name may represent the legal entity, products, people, or economic divisions), and relations highly depend on context?
- is fine-tuning of LLM effective for domain-specific relation extraction?
- do the performance improvements obtained by fine-tuning LLM justify the cost incurred?

To answer these questions, we led several experiments, each of them involving a language model processed on the same corpus CORE (Borchert et al., 2023) which is a high-quality resource specifically designed for extracting economic relations. In this domain, preserving data confidentiality is a critical concern for organizations, particularly when dealing with sensitive economic information. Sharing data with third-party servers via APIs, which is often required for using proprietary LLMs, poses significant risks to privacy and security. As a result, organizations are increasingly prioritizing models that can be fully deployed, trained, and fine-tuned locally, ensuring that data never leaves their infrastructure. This approach not only addresses confidentiality concerns but also provides greater control over the training process, enabling the customization of models for specific tasks and datasets. These constraints strongly influenced our choice to focus on open-source models that could be installed and operated entirely on our servers, eliminating the need for external dependencies and ensuring compliance with strict data protection policies. The different tested models are a Language Model based on a BERT architecture, a economic specific LLM FinGPT (Wang et al., 2023) and three general LLM: ChatGLM2 (Team GLM et al., 2024), XLNet (Yang, 2019) and LLama3 (Dubey et al., 2024). These three models have been refined thanks to few-shot learning and fine-tuning techniques alternately. We report these experiments in the following. The rest of the paper is organized as follows. Section 2 presents related work for RE, limited to the sentence level, in specific domains and when using LLM. Section 3 outlines the problem and presents our methodology. Specific-domain resources used for our experiments are described in Section 4, and Section 5 gives and comments the obtained results. We propose in Section 6 a discussion, before concluding and giving perspectives to this work.

2 Related Work

2.1 Relation Extraction

Over the years, a variety of approaches have been developed for relation extraction (RE). The initial methods viewed RE as a multi-step process, beginning with named entity recognition and then moving on to relation classification (Zeng et al., 2014). More recently, transformer-based architectures have become the dominant approach (Wang et al., 2020), offering more powerful representations and enabling end-to-end extraction processes. Additionally, sequence-to-sequence (seq2seq) models have emerged as a promising technique for RE, demonstrating significant improvements in task performance (Cabot and Navigli, 2021; Josifoski et al., 2021).

Within these approaches, some are tailored towards extracting relationships from short sentences, typically identifying a single relationship between a pair of entities in each sentence. Others process longer texts, such as paragraphs or entire documents, where the model must extract all possible relationships among multiple pairs of entities.

2.2 Extracting Relations from a Sentence

Relation extraction at the sentence level is a significant focus in the field of Natural Language Processing (NLP) (Martínez-Rodríguez et al., 2020; Pawar et al., 2017). Many studies examine general types of relationships, such as hypernymy or cause-and-effect, using well-known manually annotated datasets like SemEval-2010 Task 8 (Hendrickx et al., 2019), ACE 2004 (Mitchell et al., 2005), and TACRED (Zhang et al., 2017). Deep learning methods have led to the development of various approaches to RE. For instance, (Khaldi et al., 2021) pioneered the development of knowledge-informed models for economic RE, employing simple neural architectures that necessitate no additional training for acquiring factual knowledge about entities, nor do they require alignment between each entity and its vector representation.

Recently, models fine-tuned specifically for the economic sector include FinGPT (Wang et al., 2023) and Fin-LLaMA (Todt et al., 2023), which were introduced in July 2023. FinGPT, built on OpenAI's GPT architecture, is optimized for economic by utilizing base models such as BLOOM and ChatGLM-

6B. It has been fine-tuned for relation extraction tasks, enabling it to identify predefined entity pairs and determine the relationship between each pair. Additionally, FinGPT can jointly extract all entity pairs from a given sentence, along with the relationships connecting them.
Methods based on extracting relations from a sentence generally identify a single relation between a pair of entities in each sentence, even if more than one relation exists. For example, these methods do not deal with enumerations and n-ary relations.

2.3 Relation Extraction in the economic field

In the economic domain, RE systems are crucial for identifying specific relationships within texts, such as extracting and linking key performance indicators (KPIs) from economic documents (Hillebrand et al., 2022). Several datasets have been developed for RE using economic news, reports and earnings calls, including FinRED (Sharma et al., 2023), CorpusFR (Jabbari et al., 2020), Financial News Corpus (Wu et al., 2020), CORE (Borchert et al., 2023) and REFinD (Kaur et al., 2023).
Over the last few years, there has been a significant increase in research integrating financial datasets with GPT-based models like GPT-3 and GPT-4 to advance NLP applications (Mann et al., 2020). The leading methodologies generally fall into two categories: The first involves prompt engineering (White et al., 2023) with open-source LLMs, using their existing parameters. The second relies on supervised fine-tuning methods, such as Instruction Tuning (Ouyang et al., 2022), to create domain-specific LLMs that excel in financial tasks, among which:

- FinBERT (Araci, 2019) is a specialized model for financial sentiment analysis with under one billion parameters, fine-tuned on a financial corpus to excel in economic-related tasks.
- BloombergGPT (Wu et al., 2023) is a closed-source model derived from BLOOM, trained on a wide array of financial datasets to cover a broad spectrum of financial concepts.
- FinGPT (Yang et al., 2023) is an open-source LLM, fine-tuned from a general LLM (such as Llama2 or FinBert depending on FinGPT version) using low-rank adaptation methods to promote broader community accessibility.

2.4 Relation Extraction with fine-tuned LLM

Instruction tuning is a recent trend where supervised fine-tuning on a wide variety of tasks, often represented through demonstrations, has led to improved generalization in LLM (Wang et al., 2022). This approach aims to leverage the extensive knowledge gained by LLM during pre-training, making them more adaptable to new tasks. Various adaptation strategies have been developed to enhance fine-tuning in LLM, allowing for greater flexibility and efficiency. One such strategy is prefix-tuning (Li and Liang, 2021), where only a small segment, typically at the beginning (or "prefix") of the pre-trained transformers, is updated while keeping static the rest of the model parameters. This method reduces computational overhead and helps maintain the stability of the original model. Another notable strategy is Low-Rank Adaptation (LoRA) (Hu et al., 2021). Unlike traditional fine-tuning, which modifies the entire model, LoRA introduces injectable low-rank matrices that can be trained independently. This technique minimizes the risk of overfitting and significantly reduces the storage requirements for the fine-tuning process. A key benefit of LoRA is its compatibility with other strategies, including prefix-tuning, allowing for more comprehensive and adaptable fine-tuning approaches.

3 Task and Methodology

3.1 Task description

Given a sentence $S = \{w_1, w_2, ..., w_n\}$ consisting of n words, an entity E is defined as a contiguous span of words where $E = \{w_i, w_{i+1}, ..., w_j\}$ for indices $i, j \in \{1, ..., n\}$ and $i \leq j$. The goal is to extract a set of relation facts from the input sentence. Each fact is represented as a relation triplet. A relation triplet consists of three components: a first entity E_1, a relation $r \in \mathcal{R}$ from a predefined set of relation labels $\mathcal{R}$, and a second entity E_2. The triplet structure is formally expressed as (E_1, r, E_2). In the context of economic relation extraction, it is crucial to determine which model approach offers the best performance and efficiency. The methods examined in this paper include (1) Training BERT-based models, which leverage transformer networks to identify relationships between entities; (2) Applying zero-shot and few-shot learning techniques to LLM, where models are assessed with no specific or few examples; and (3) Fine-tuning LLM, offering a more tailored and precise

approach for domain-specific tasks. This study aims to evaluate these methods by comparing their performance in terms of accuracy. The goal is to determine the optimal strategy for relation extraction in economic texts, while highlighting the strengths and limitations of each technique.

3.2 Methodology for economic Relation Extraction Using LLM

In this section, we present a comprehensive evaluation strategy for economic relation extraction (BRE) leveraging generative and open-source large language models (LLMs) fine-tuned with task-specific data. We begin by developing efficient instructions adapted to the natural language and specified entities present in the CORE dataset, which we will discuss in more detail in the following section. Simultaneously, we establish optimal input and output configurations to enhance the model's understanding and task performance. Next, the PEFT framework is employed to facilitate efficient fine-tuning of the LLM, a process we will describe in the subsequent section. Following this, the fine-tuned models are utilized to generate inference results in the form of relation triplets from the provided text data through carefully crafted prompts. Finally, a direct extraction process is implemented to derive the relations from the generated triplets, effectively elucidating the connections between the specified entities within the text.

3.2.1 Instruction-Based Fine-Tuning Design

LLM are typically released with a recommended prompt template to ensure effective interaction with the model during inference. A prompt template refers to a structured string with placeholders that are populated with input data, guiding the model to produce the desired output (Lyu et al., 2024). To construct an instruction-based fine-tuning dataset, it is essential to design the instruction, input, and output formats. A prompt can contain any of the the following components (Irfan and Murray, 2023):

Instruction - a specific task of instruction you want the model to perform.

Context - can involve external information or additional context that can steer the model to better responses.

Input Data - is the input or question that we are interested to find a response for.

Output Indicator - indicates the type of format of the output.

Not all the components are required for every prompt, and their inclusion depends on the specific task at hand.

In our fine-tuning design, we incorporate three key components into our prompt: the instruction, the input sentence, and the output format. The instruction is defined as: *"What is the relationship between {E1} and {E2} in the context of the input sentence. Choose an answer from: {list_of_relations}}"*. This helps direct the model's attention towards identifying the correct relationship between the specified entities. To further clarify the expected response, we append the output format: (*(E1, Relation, E2)*, ensuring that the model generates relation triplets in a consistent and structured manner. This predefined prompt format is then applied throughout the fine-tuning data-set to guide the model's training and improve its performance in relation extraction tasks.

3.2.2 Efficient Fine-Tuning of LLM for Relation Extraction

To mitigate the significant computational costs of fine-tuning LLM and address the limitations of RE tasks, an efficient solution is required. We employ the PEFT framework (Mangrulkar et al., 2022), which significantly reduces the number of trainable parameters while maintaining high performance. PEFT is compatible with a variety of open-source LLM, such as Llama3-8B (Dubey et al., 2024), ChatGLM2-6B (Team GLM et al., 2024), and XL-Net (Yang, 2019), etc. Specifically, the LoRA method is applied to the Query (Q) and Value (V) matrices within the Gated Query Attention (GQA) section, which are then combined with the Key (K) part to compute the attention mechanism as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The generated attention is passed through several network layers to extract the relations between the given entities in the input text. The overall process for predicting the relation triplets can be formulated as :

$$p_\theta(Y|X, P) = \prod_{i=1}^{m} p_\theta(y_i|X, P, y_{<i})$$

Where $X = [x_1, x_2, ..., x_n]$ represents the input text sequence, $Y = [y_1, y_2, ..., y_m]$ represents the target sequence, and P is the prompt. By leveraging the PEFT framework, we address

the challenge of limited perceptual capabilities in generic open-source LLM, while simultaneously improving the understanding and generalization of domain-specific texts. This approach enhances the precision of relation extraction and is widely applicable to domain-specific BRE.

4 Resources from the economic field

In this section, we present the resources we used for relation extraction in the economic field, including the dataset and models tested throughout our experiments.

4.1 Dataset

We used the CORE dataset (Borchert et al., 2023), a high-quality resource specifically designed for extracting company relations, which are a subset of economic relations. Unlike distantly supervised datasets, CORE is manually annotated, covering a broad range of relation types and entity categories, including named entities, common nouns, and pronouns. The dataset focuses on economic entities such as companies, brands, and products, making relation extraction more challenging due to the varied contexts in which these entities appear. Annotators labeled 12 predefined relation types (see Figure 1), ensuring high data quality through multiple validation rounds. The annotated instances were randomly divided into a training set (4000 instances) and a test set (708 intances), each split containing all available relations types. We chose this dataset because its focus on economic entities aligns with the objectives of our research, enabling us to evaluate the performance of our models in real-world economic contexts. Furthermore, the high-quality, manually annotated nature of CORE ensures that our results are grounded in accurate and reliable data, which is crucial for the success of fine-tuning and evaluating LLMs in the context of economic relation extraction.

4.2 Models for economic Relation Extraction

In our experiments, we evaluated several models for their performance in economic relation extraction at the sentence level, as we mentioned earlier. These models were chosen because they are open-source and can be easily deployed locally:

- **XLNet (Extra-Long Transformer Network)** : a language model based on the Transformer architecture, developped by Google. Its major innovation lies in the use of Permutation Language Modeling (PLM), allowing the model to consider different word orders. It also includes a segment-level recurrent mechanism and two-stream self-attention to better capture distant dependencies and bidirectional relations. XLNet utilizes several datasets, including BooksCorpus and the English version of Wikipedia. Additionally, it incorporates Giga5, ClueWeb 2012-B, and Common Crawl.

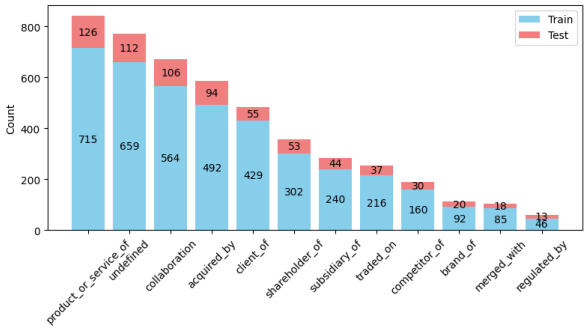

Figure 1: Relation types and distribution in the data-set

- **ChatGLM**: A bilingual language model optimized for question-answering and dialogue in Chinese and English, ChatGLM is based on the General Language Model (GLM) framework with 6.2 billion parameters. The model's pre-training data includes 1.2 terabytes of English text and 1.25 terabyte of Chinese text. In our experiments, we specifically used the ChatGLM2-6B version.

- **Llama-3**: Developed by Meta, Llama-3 is a family of LLM with 8 or 70 billion parameters. It is optimized for instruction-based tasks and excels in dialogue use cases, outperforming many open-source chat models. Llama 3 is pretrained on over 15T tokens that were all collected from publicly available sources. In our experiments, we specifically used the Llama-3 model with 8 billion parameters (Llama3-8B).

- **FinGPT**: Is an open-source framework designed for financial large language models (FinLLM), enabling the analysis and extraction of insights from financial data. It is trained on datasets like news and tweet sentiment analysis to support domain-specific tasks in economic.

- **BizBERT**: A fine-tuned version of BERT, BizBERT is trained on economic-specific datasets BizREL ((Khaldi et al., 2021)) and uses BERT's pre-trained language model (PLM) to encode sentences, focusing on economic entities and relations.

5 Experiments and Results

This section presents the experimental setup, results, and their analysis. We aim to address the following key research questions:

- **RQ1: Whether and how large models can perform better than smaller models?** We evaluate several models with differents sizes and compare their performance in economic relation extraction.
- **RQ2: Is fine-tuning of LLM effective for domain-specific relation extraction?** We explore whether refining LLM with techniques like N-shot learning or fine-tuning enhances their performance in extracting relations specific to a domain, such as economic.
- **RQ3: Do the performance improvements obtained by fine-tuning LLM justify the cost incurred?** The goal is to determine if the improvements in relation extraction accuracy justify the higher computational resources required for fine-tuning LLMs.

5.1 Experimental Setup

In order to answer these research questions, we conducted extensive experiments on domain-specific datasets.

Baseline: We used BizBERT (Khaldi et al., 2021) as a baseline

Evaluation Metrics : We used Precision, Recall, and F1-Score to evaluate the performance of the models.

Hyperparameters and Environment : For the CORE dataset, we fine-tuned the LLM for 8 epochs with a learning rate of 1e-4. The batch size was set to 4, and the gradient accumulation steps were 8. All experiments were conducted on a single NVIDIA RTX8000 (24 Go RAM).

5.2 Performance Evaluation

We aimed to compare the effectiveness of LLM against smaller, more traditional models, such as BERT-based models, in order to assess how well they adapt to domain-specific tasks like BRE. The results of our evaluation, presented in Table 1, provide the performance of various models, including XLNet, ChatGLM, BizBERT, FinGPT, and Llama3, on the CORE dataset.

We began by testing the models using zero-shot and few-shot learning techniques. In zero-shot learning, the model directly predicts relationships without prior task-specific examples, relying solely on its pre-trained knowledge. For few-shot learning, we included three examples in the prompt as demonstrations to guide the model. These examples consisted of a sentence with annotated entities and their corresponding relationships, helping the model understand the expected output format and contextual cues. Few-shot learning leverages the model's ability to generalize from limited task-specific data, making it particularly useful for scenarios with minimal annotated resources. BizBERT was retrained on the CORE training data. Similarly, FinGPT involved fine-tuning the BLOOM model (Le Scao et al., 2023) on the CORE dataset. This technique adjusts the model's parameters while preserving its general pre-trained knowledge. Fine-tuning is particularly effective for adapting large language models to specialized domains, as it enables them to align closely with the target task's requirements. The results include comparisons between models tested in zero-shot and few-shot settings, as well as those subjected to fine-tuning, highlighting the differences in their adaptability and performance under varying levels of task-specific training.

From Table 1, it is evident that large language models like Llama3 and ChatGLM consistently outperform traditional BERT-based models like BizBERT, particularly when fine-tuned for domain-specific tasks such as BRE. Fine-tuning significantly enhances performance, as seen in the increase of F1 scores from 0.69–0.70 in zero- and few-shot learning to 0.80 after fine-tuning Llama3. The results confirm that fine-tuning LLM on a task like BRE is highly effective, leading to substantial improvements in F1 scores. The comparison underscores the potential of LLM to outperform smaller models, especially when adapted to specialized tasks, making them the most efficient and accurate solutions in these experiments.

5.3 Effectiveness of Fine-Tuning LLM

To validate the effectiveness of fine-tuning large language models, we conducted experiments using the CORE data-set. We fine-tuned Llama3 using LoRA on varying proportions of the CORE

Method	Zero-shot	Few-shot	Fine-tuning	Retrained
BizBERT	unavailable	unavailable	unavailable	0.71
XLNet	0.54	0.58	0.76	unavailable
ChatGLM	0.56	0.59	0.78	unavailable
FinGPT	0.38	0.41	0.76	unavailable
Llama3	**0.69**	**0.70**	**0.80**	unavailable

Table 1: The F1 score comparison of models on CORE dataset.

training data (4000 instances): 10%, 30%, 50%, and 70%, and compared the results with the model fine-tuned on the entire data-set. Llama3 was selected for these experiments because it yielded the best results in our evaluations, demonstrating superior performance in economic relations extraction tasks compared to other models. As shown in Table 2, the performance of the model significantly improves with fine-tuning, even when using a small portion of the data. This demonstrates that fine-tuning is a much more effective strategy for domain-specific tasks. For example, the results clearly show that the model's performance on the BRE task continues to improve as more training data is incorporated. By fine-tuning with 30% of the training data, the F1 score reached 0.75, already surpassing the performance of the model fine-tuned with fewer data. Notably, the gains become more gradual beyond 50% of the training data, where the F1 score reaches 0.77, and when using 70% of the data, the F1 score improves slightly to 0.78. This plateau in performance suggests that fine-tuning on a substantial subset of the training data (around 50-70%) is sufficient to achieve robust generalization, highlighting the importance of data quality over sheer quantity. Fine-tuning with the entire dataset yields the best result, with an F1 score of 0.80, confirming that fine-tuning is an essential step for achieving state-of-the-art performance in economic relation extraction tasks.

6 Discussion and Conclusion

The results of our experiments demonstrate that fine-tuning LLM is a highly effective strategy for improving performance on domain-specific tasks, such as economic Relation Extraction. Across our trials, models like Llama3 consistently outperformed smaller BERT-based models and exhibited significant performance gains when fine-tuned with domain-specific data. This study supports the hypothesis that while large models may not always show substantial improvement in general tasks, their adaptation to specialized domains is crucial for realizing their full potential.

One key observation from the experiments is that fine-tuning even on a fraction of the available data (30-50%) yielded substantial improvements. However, further increases in data usage led to diminishing returns, indicating that optimal performance can be achieved without needing the full dataset. This underscores the importance of efficient resource allocation in training, as fine-tuning large models can be computationally expensive. Moreover, fine-tuning open-source LLM locally offers a compelling alternative to propriety solutions, especially in privacy-sensitive domains like economic, where data confidentiality is a critical factor.

In conclusion, this study demonstrates that fine-tuning LLM for domain-specific relation extraction not only improves performance but also offers a cost-effective and scalable solution.

For future work, we aim to focus on the extraction of multiple triplets from paragraphs, where several relationships need to be identified within the same text. Additionally, we plan to investigate the extraction of n-ary relations, extending the traditional binary relations extraction approach to handle more complex relationships involving multiple entities.

7 Acknowledgement

This work was carried out within the ECLADATTA project funded by the French National Research Agency under grant ANR-22-CE23-0020.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.

Philipp Borchert, Jochen De Weerdt, Kristof Coussement, Arno De Caigny, and Marie-Francine Moens. 2023. CORE: A few-shot company relation classification dataset for robust domain adaptation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Fine-tuning Setting	Precision	Recall	F1 score
Llama3 + 10% Training Data	0.75	0.72	0.73
Llama3 + 30% Training Data	0.78	0.74	0.75
Llama3 + 50% Training Data	0.80	0.75	0.77
Llama3 + 70% Training Data	0.81	0.77	0.78
Llama3 + All Training Data	**0.82**	**0.79**	**0.80**

Table 2: Impact of Fine-Tuning Techniques on LLM Performance in BRE

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *CoRR*, abs/1911.10422.

Lars Hillebrand, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. Kpi-bert: A joint named entity recognition and relation extraction model for financial reports. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Muhammad Irfan and Liam Murray. 2023. Micro-credential: A guide to prompt writing and engineering in higher education: A tool for artificial intelligence in llm. Technical report, University of Limerick.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A french corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2293–2299.

Martin Josifoski, Nicola De Cao, Maxime Peyrard, and Robert West. 2021. Genie: Generative information extraction. *CoRR*, abs/2112.08340.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hadjer Khaldi, Farah Benamara, Amine Abdaoui, Nathalie Aussenac-Gilles, and EunBee Kang. 2021. Multilevel entity-informed business relation extraction. In *International Conference on Applications of Natural Language to Information Systems*, pages 105–118. Springer.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*.

Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, and Sophie Rosset. 2024. Small language models are good too: An empirical study of zero-shot classification. *arXiv preprint arXiv:2404.11122*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2024. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *CoRR*, abs/2402.18540.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1.

José-Lázaro Martínez-Rodríguez, Aidan Hogan, and Ivan López-Arévalo. 2020. Information extraction meets the semantic web: A survey. *Semantic Web*, 11(2):255–335.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. 2017. Relation extraction : A survey. *CoRR*, abs/1712.05191.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pawan Kumar Rajpoot and Ankur Parikh. 2023. Gptfinre: In-context learning for financial relation extraction using large language models. *CoRR*, abs/2306.17519.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2023. Finred: A dataset for relation extraction in financial domain. *CoRR*, abs/2306.03736.

Team Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Pedram Babaei William Todt, Ramtin Babaei, and P Babaei. 2023. Fin-llama: Efficient finetuning of quantized llms for finance.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *CoRR*, abs/2310.04793.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *CoRR*, abs/2302.11382.

Haoyu Wu, Qing Lei, Xinyue Zhang, and Zhengqian Luo. 2020. Creating a large-scale financial news corpus for relation extraction. In *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 259–263. IEEE.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? *Preprint*, arXiv:2305.01555.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *CoRR*, abs/2306.06031.

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on empirical methods in natural language processing*.