PRICER: Leveraging Few-Shot Learning with Fine-Tuned Large Language Models for Unstructured Economic Data*

Matt Murtagh[†], PJ Wall and Declan O'Sullivan

School of Computer Science and Statistics, Trinity College Dublin, Ireland

Abstract

Accurate collection of economic data is crucial for metrics like the Consumer Price Index (CPI), informing policies on inflation and living costs. Traditional manual data collection methods from retail sources are labor-intensive and fraught with issues of scalability, accuracy, and data diversity. Our study introduces an OWL RDFS-based framework aligned with COICOP, and a transformer model, 'PRICER', to automate the extraction and structuring of online retail data into RDF. By iteratively fine-tuning PRICER—first with a broad DBPedia and Wikipedia knowledge base, then with specific online retail data—we achieve significant efficiency and accuracy improvements in data collection. Notably, PRICER shows marked performance gains in precision and recall after task-specific conditioning, validating our approach for converting unstructured text to structured knowledge. This advancement facilitates streamlined economic data aggregation and highlights PRICER's adaptability for broader standardised data processing applications. Future work will focus scaling the domain specific price dataset, refining the model's conditioning and exploring potential for other forms of technical data.

Keywords

Knowledge Graphs, Deep Learning, Large Language Models, Economic Data

1. Introduction

Statistical measures, such as the Consumer Price Index (CPI), rely heavily on highly heterogeneous price and consumption data to provide insights into inflation rates and the cost of living, influencing monetary policies and market strategies worldwide. Traditionally, this data is collected manually from a variety of retail sources, a process that is not only time-consuming but also prone to inaccuracies and inconsistencies. The Classification of Individual Consumption by Purpose (COICOP) [1], an international standard taxonomy developed by the United Nations, serves as a guideline for categorising such consumer goods and services. However, the application of COICOP in the face of highly heterogeneous and non-interoperable data formats presents significant challenges. In this context, the integration of advanced computational models and ontologies offers a promising avenue for transforming the landscape of economic data collection and analysis. By automating the extraction and structuring of data, we aim to

ESWC 24: Second International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data SemTech4STLD, May 26, 2024, Crete

[*] Repository for this paper is available here

[†] Corresponding author.

mmurtagh@tcd.ie (M. Murtagh); wallp2@tcd.ie (P. Wall); declan.osullivan@tcd.ie (D. O'Sullivan)

enhance the efficiency, accuracy, and interoperability of these vital economic indicators, setting the stage for more informed decision-making and policy development with data available in almost real-time.

In this paper, we propose a simple RDFS schema for the construction and population of price indices using an existing international standard. We introduce a framework for collecting and processing economic data from the internet directly into a knowledge graph, structured using the Resource Description Framework Schema (RDFS). This choice is due to the simplicity of the schema requirements, consisting of two datatype properties and the eighty classes that comprise the first module of the COICOP.

We employ Mistral 7B, an open-source 7 billion parameter model [2], and fine-tune it using Low Rank Adaptation (LoRA), a parameter-efficient fine-tuning (PEFT) method [3, 4]. Using a training dataset of prices collected automatically via a web crawler and manually annotated with the constraints of the RDFS schema, we test the accuracy of the model and report our results.

2. Related Work

Transforming unstructured text into structured knowledge graphs has long been a significant area of research, with notable contributions highlighting the potential of Open Information Extraction and entity recognition techniques [5, 6, 7]. Despite the established importance of these methodologies, the application of Large Language Models (LLMs) as agents for the collection and structuring of economic data represents a nascent and largely unexplored frontier. Prior investigations into LLMs concerning economic data have predominantly centered around enhancing the explainability of forecasts [8], deriving sentiment from financial narratives [9], and providing investment guidance [10], or simulating macroeconomic agents [11]. These studies, however, focus on processing or interpreting data that is already in a structured or semi-structured form in natural language, contrasting sharply with our ambition to structure raw, unstructured data directly into knowledge graphs.

The closest comparison to the current research is the Billion Prices Project [12], a now-defunct project that created an alternative indicator for inflation statistics by scraping online price data with particular utility towards countries or time periods where existing official statistics may be unreliable [13]. However, the Billion Prices Project still required manual classification, identification of appropriate retailers and highly specified web scraping scripts in order to function; most notably, all products and prices had to be manually categorised within the bounds of the COICOP after collection [12]. PRICER iterates on the goal of the Billion Prices Project by specifying an RDFS for the COICOP and directly classifying unstructured data within it.

2.1. Classification of Individual Consumption by Purpose

The Classification of Individual Consumption by Purpose (COICOP) is a hierarchical framework developed by the United Nations Statistics Division to standardize the categorization of goods and services consumed by individuals and households. Its primary aim is to facilitate international comparisons of consumer behavior and inflation rates by providing a consistent taxonomy for classifying consumer expenditure. COICOP divides consumer expenditures into broad categories such as food and beverages, housing, health, education, and transportation, which are further subdivided into more detailed classifications. This systematic approach enables researchers, policymakers, and statisticians to aggregate and analyze data in a manner that reflects the functional aspects of consumer spending, thereby enhancing the accuracy and comparability of economic indicators such as the Consumer Price Index (CPI). The adoption of COICOP across national statistical agencies underscores its significance in economic analysis, offering a robust framework for understanding consumption patterns and guiding economic policy decisions [1].

For the purpose of this paper, we have transposed the first and largest section of the COICOP, covering food, beverages, and tobacco, into a simplified RDFS schema. This schema captures the hierarchical nature of the COICOP via a straightforward subclassing mechanism. We include just two properties in our schema: ***:hasPrice*** and ***:collectionDate***. These are designed to demonstrate how a basic RDFS framework can structurally reflect a portion of the COICOP classification in a manageable way for this research context[1].

.

2.2. Transformers and Large Language Models

NLP models have long used representation of text as vector embeddings to classify and understand paragraphs and documents of text, allowing for the semantic meaning of words and subwords to be represented in a high-dimensional space [14, 15, 16]. The introduction of the self-attention mechanism, where an additional embedding value is given to represent the importance of every other word in a sequence to a given word, in addition to positional encodings to allow for highly parallelised processing dramatically changed the field [17]. The introduction of transformer models such as BERT, trained on large diverse datasets, brought on the advent of pre-trained models that can be fine tuned for specific tasks [18]. The high benchmark performance and proliferation of these new models led to a diversity of variants, making slight changes to the model structure for increased memory efficiency [19], larger datasets [20] and performance [21]. The scaling up of data and parameters used in these models has brought on the proliferation of LLMs. The release of GPT-3, a 175 billion parameter model, has popularised the LLM not only in academic spheres but also in the public imagination, with strong task-agnostic performance and few shot learning at scale, with far less necessity to fine-tune on specific tasks [22]. Concurrently, there has been a plethora of open-source LLM models developed and released to the public at similar scale to commercial LLMs. Bloom, an open-source collaboration of major research players in the area, developed a 176B parameter model that is open for researchers to download, test, fine-tune and integrate into existing research [23]. Similarly, Falcon LLM produced a competitive openly available 7B parameter model and demonstrated that training a model on properly filtered and deduplicated web based data alone can be as effective as training on a mix of such data and highly curated training datasets [24], a lesson that this paper utilises in its approach. Meta similarly released LLaMA, with sizes ranging from 7B to 60B, outperforming larger models such as GPT-3 on a number of

[1] This schema is purely demonstrative; a more complete, semantically rich ontology is outside the scope of this paper. We make it available for download here.

NLP benchmarks [25]. Mistral 7B, the model used for this paper, is an open source LLM that outperforms LLaMA of sizes up to 34B on most tasks, incorporating a sliding windows attention mechanism that allows for cheaper training [2].

2.3. Limitations of Large Language Models

While LLMs are powerful tools and provide an exciting new avenue for research, they also have significant issues. One major flaw is the tendency for LLMs to ”hallucinate”, a phenomenon where output in response to a prompt is plausible sounding but nonsensical or factually incorrect [26, 27], rather than admitting a lack of knowledge. These hallucinations are particularly problematic in fields that require highly accurate answers and where fact-checking might be a lengthy, time-limited process. This has led to LLMs being characterised as ”Stochastic Parrots”, mirroring human responses or generating code in a probable manner rather than reflecting some underlying meaning [28]. LLMs have achieved their scale largely due to aforementioned large volumes of data produced and gathered on the internet. However, this has led to issues with models reflecting human biases in racist, sexist, extreme or other forms of derogatory or harmful dialogue found online and then incorporated in to datasets without sufficient filtering or quality control [28]. Finally, the scale of LLMs requires long term training with clusters of hundreds of power-intensive GPUs, meaning that training and running these models can come with significant environmental costs [28]. This incentivises the use of smaller, more efficient LLMs. While Mistral 7B, the model fine-tuned as part of this paper is significantly smaller than many LLMs such as GPT, we acknowledge that the training of such models still incurs an environmental cost. Similarly, while the dataset used to fine tune the model is extracted from the internet, it is far smaller than the large scale datasets used for the primary training of these models, and can therefore be filtered more effectively.

3. Approach

3.1. Model Overview and Structure

This section introduces the structure and workflow of our model. The core objective is to transform HTML data from online retailers into a knowledge graph format, adhering to our COICOP RDFS schema. This includes the identification and extraction of product prices and classifications. The foundational model, Mistral 7B, undergoes a two-phase fine-tuning process: initially with DBPedia data to grasp information structuring, followed by domain-specific adaptation using online retail data. Figure 1 depicts the overall architecture of PRICER. The model employs a transformer-based approach, leveraging the capabilities of Mistral7B for semantic understanding and knowledge extraction. The fine-tuning process is bifurcated into an initial phase focusing on general knowledge graph formation, and a subsequent phase targeting the COICOP RDFS for retail data.

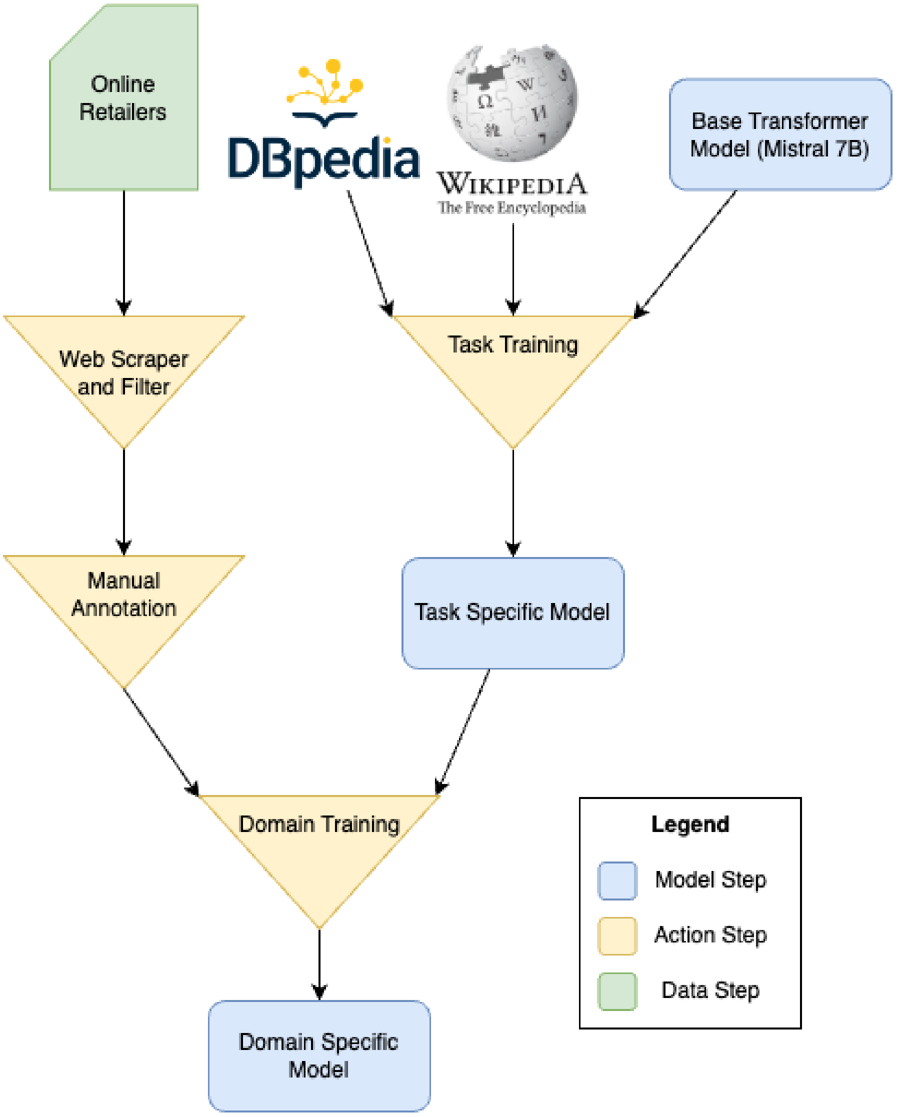

Figure 1: An illustrative diagram showcasing the layered fine-tuning approach and the transition from general information parsing to domain-specific knowledge extraction.

3.2. Task Specification, Data Preparation and Training

As discussed, the specific task of PRICER is to generate RDF triples from HTML text scraped from online retailers, within the parameters defined in the COICOP RDFS schema. Specifically, the input data is text data, while the output data are RDF triples. To achieve this, the Mistral 7B model is fine-tuned in two phases; task-specific fine tuning and domain-specific fine tuning. We outline both of these processes here.

3.2.1. Phase 1: Task-Specific Fine Tuning

The initial training phase is designed to prime the model with the capability to transform unstructured textual information into structured data aligned with knowledge graph schemas. This foundational training leverages a dataset of 18,000 pairs, each consisting of text extracted from Wikipedia and the corresponding structured data in the form of DBPedia triples. The objective is to imbue the model with an intrinsic understanding of how to discern and map complex information from natural language text to a structured format that mirrors the organised nature of knowledge graphs. This begins with the acquisition of textual data, for which we utilise the Wikipedia API's "random page" feature. This tool enables us to systematically retrieve the full textual content from a diverse array of Wikipedia pages. The randomness ensures a wide coverage of topics, which allows the model to learn from a broad spectrum of subjects and contexts. Parallel to textual data extraction, we engage in the retrieval of structured data corresponding to each Wikipedia page. This is achieved through the DBPedia SPARQL endpoint, a query service that allows us to download the triples associated with the Wikipedia pages that have been extracted. DBPedia, being a structured version of Wikipedia's content based on the Wikipedia page's information box, offers a rich set of triples that represent the information on the pages in a structured RDF format. For the purpose of this research, we make the assumption that the information contained in the infoboxes on Wikipedia pages are directly related to the text content of those pages. To ensure that the collected RDF data is valid, we employ RDFlib, a Python library designed to work with RDF data. We drop any observations that cannot be parsed by RDFlib. Upon successful extraction and validation, we proceed to pair each piece of Wikipedia text with its corresponding set of DBPedia triples. This pairing forms the basis of our training dataset, where the model learns to associate the unstructured text with structured knowledge representations. Moreover, we manually conduct a final review to eliminate any duplicates.

3.2.2. Phase 2: Domain-specific Fine-tuning

The foundation of our domain-specific dataset begins with an automated data collection process. Utilising Python scripts that integrate Selenium and BeautifulSoup, we devised a methodical approach to navigate and extract content from web pages of a prominent Irish online retailer. Selenium, a tool for browser automation, enables our scripts to interact with the web pages dynamically to access the necessary product information. BeautifulSoup complements this by parsing the HTML content retrieved by Selenium, allowing for efficient extraction of data embedded within the web pages. Upon retrieving the HTML content, our focus shifts to the extraction of relevant product information. Given the diverse structure of web pages, we employ a filtering process to isolate the data of interest. This process begins with the identification and extraction of content enclosed within ”div” tags, a common HTML element used to group block-level content on web pages. This step is narrows the data to segments more likely to contain product information.

Following the initial isolation, we further refine the dataset through a targeted keyword search. This search is designed to identify and retain content featuring key terms indicative of retail-specific details, such as ”price” and the euro symbol (€). This keyword filtering ensures that the extracted excerpts are relevant to our objective, focusing on product prices and possible classifications. With the filtered set of HTML product excerpts in hand, the next phase involves the manual annotation of this data within the COICOP RDFS framework. Each excerpt is thoroughly reviewed and labeled with appropriate COICOP RDFS labels. This encompasses both product types and price details. Table 1 demonstrates an example of extracted text and the corresponding RDF triples chosen to represent that text as part of the annotation process. Following this process, we produced a dataset of 1000 observations of gathered data.

Extracted Text	**RDF Triples**

€2.10 €9.25/kg Quantity controls Quantity of Kearns Pork Sausages 227G Add	`<http://example.org/coicop#KearnsSausages> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.org/coicop#MeatFreshChilledFrozen> .`
	`<http://example.org/coicop#KearnsSausages> <http://example.org/coicop#hasPrice> ”2.10”`<http://www.w3.org/2001/XMLSchema#decimal> `.`
	`<http://example.org/coicop#KearnsSausages> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.org/coicop#LiveAnimalsMeatAndOtherParts> .`
	`<http://example.org/coicop#KearnsSausages> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.org/coicop#Food> .`

Table 1
Example of text extracted from retailer and corresponding labelled RDF triples

For both the Wikipedia-DBpedia and prices datasets, consistent with the Mistral 7B format, each observation is divided into the system prompt, user prompt and response in the format *System Prompt [INST] User Prompt [/INST] Response.* ”Text to RDF” was provided as the system prompt common to all observations. The user prompt was the relevant text data, in this case either the Wikipedia text data or the scraped price data. Finally, the response was either the DBPedia triples or COICOP triples. These individual text observations were encoded using the Mistral tokenizer.

3.3. Training Specification

The training and evaluation was performed on a cluster of four NVIDIA RTX A5000 GPUs. We fine-tune the model using Low Rank Adaptation (LoRA), a parameter-efficient fine-tuning (PEFT) method [3, 4]. This approach freezes the weights from previous training sessions, enabling the training of large models with reduced computational resources. Training time for the DBPedia sample was 12 hours, while training time for the smaller price dataset was just half an hour. The model was evaluated on the test dataset, which comprised of 10 percent of the total scraped data gathered and labelled. Each row of text gathered from the product web page in the test dataset was tokenized and inputted to the model. The text output, formatted in n-triples, was parsed with RDFlib for evaluation. This is a complex multi-label classification task. One row of input data results in numerous triple outputs. As such, to evaluate the model, we first provide the raw input text collected from the website from each observation of our test data. The generated triples are then recorded and parsed using RDFlib and any duplicate triples are dropped. These triples are then compared against the those that have been pre-labelled by hand. A correct observation is considered that where the *http://example.org/coicop#hasPrice* relation or *http://www.w3.org/1999/02/22-rdf-syntax-ns#type* relation has been correctly estimated. All triples, both those generated by the model and manually labelled input datasets are in n-triples format.

4. Results

The empirical evaluation of PRICER, as shown in the structured comparative analysis (Table 2), demonstrates the influence of model conditioning through targeted fine-tuning. We analyse a number of performance metrics across various model configurations, including precision, recall, and F1 score. The pre-training with DBPedia and Wikipedia data emerges as a critical factor in enhancing the model's proficiency in transforming unstructured retail text into structured knowledge graphs that align with the COICOP schema.

Table 2

Comparative Performance Metrics across all training configurations. Displayed are the results for the base Mistral 7B model without any fine-tuning, the base model trained with just the collected HTML dataset (M+P), and models fine-tuned with DBPedia and Wikipedia data at 900 (M+DBP900+P), 9,000 (M+DBP9K+P), and 18,000 (M+DBP18K+P) observations, respectively, before further fine-tuning with the HTML dataset.

Metric	Base 7B[2]	M+P	M+DBP900+P	M+DBP9K+P	M+DBP18K+P
Precision	00.00	0.26	0.26	0.66	0.49
Recall	00.00	0.22	0.27	0.50	0.51
F1 Score	00.00	0.23	0.26	0.53	0.46

The results from the model fine-tuned with 9,000 DBPedia and Wikipedia observations (M+DBP9K+P) indicates a significant peak in precision (0.66) and a robust recall (0.50), suggests an optimal balance of general and domain-specific knowledge for our task. Notably, the precision demonstrates peaks at 9,000 observations, but a subsequent decline is observed when further expanded to 18,000 observations (0.49). This trend prompts a critical examination of the relationship between data volume and model performance, particularly the balance between broadening the model's knowledge base and the potential for overfitting or diminishing returns on additional data.

[2] The base model has never seen data categorised in the schema and is therefore not expected to be able to complete this task. It is included here for demonstration

The iterative conditioning process, especially with an expansive corpus of 18,000 DBPedia and Wikipedia observations (M+DBP18K+P), marks a significant enhancement in the model's capabilities, outperforming both the base model and configurations fine-tuned with smaller datasets in terms of precision and recall. However, the observed decline in precision from the M+DBP9K+P to the M+DBP18K+P configuration indicates a need for further investigation into data selection and model tuning methodologies.

While the results indicate an improvement in the model's ability to generate correctly associated triples with task-specific tuning, achieving correct interpretations in approximately half of the tested observations highlights substantial avenues for improvement. This observation accentuates the necessity for ongoing refinement of the model, aiming to enhance its accuracy and reliability.

5. Limitations

While this work shows promise, it must be acknowledged that this has been achieved on a relatively simple schema with just eighty unique classes and two unique data properties. Additionally, the HTML price data is trained on that of only one online retailer. Thus, the model in its current form will not work as a general purpose annotator for other forms of data, and will need to be expanded with and tested with data from other online retailers to become more source agnostic. Additionally, the model remains untested with a more semantically rich schema or ontology. We fully acknowledge these as current limitations for this model and intend to expand on these areas in future work.

6. Conclusion and Future Work

This study presents a pioneering approach to the automation of economic data collection through the development of an RDFS-based schema, aimed at streamlining the extraction and structuring of online price and product data. By extending the COICOP taxonomy to an RDFS schema and training a transformer model on online data categorized within it, this research suggests an approach to enhance the efficiency and accuracy of data collection processes and establishes a foundation for the application of such methodologies across various domains requiring standardized data handling. The employment of Mistral 7B, fine-tuned with LoRA, showcases the potential of leveraging large language models (LLMs) in the context of economic data collection, providing a promising avenue toward the automation of complex data processing tasks.

Moving forward, our research will pivot around three key areas: significantly expanding our datasets with a larger array of labeled price data and a richer collection from Wikipedia and DBpedia to boost the model's precision and understanding of economic complexities, while exploring a diverse range of LLMs beyond Mistral 7B. Additionally, we intend to explore performance on a more semantically rich schema or ontology containing more complex relationships between objects as well as additional data features such as quantity and base pricing for CPI calculation. This multifaceted approach aims to refine our model's capability to accurately capture the nuances of economic data, thereby enhancing vital indicators like the Consumer

Price Index. By testing various models for their efficacy in integrating and structuring economic data within our RDFS-based schema, we anticipate identifying models that not only elevate performance but also exhibit greater adaptability and efficiency in the dynamic field of economic data analysis and the collection of official statistics.

7. Acknowledgments

The authors would like to thank the Science Foundation Ireland ADAPT Centre and Centre for Research Training in Artificial Intelligence for funding and support.

References

[1] United Nations Statistics, Classification of Individual Consumption According to Purpose (COICOP) 2018, 2018. URL: https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf.

[2] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, Mistral 7B, arXiv preprint arXiv:2310.06825 (2023).

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

[4] A. Chavan, Z. Liu, D. Gupta, E. Xing, Z. Shen, One-for-All: Generalized LoRA for Parameter-Efficient Fine-tuning, arXiv preprint arXiv:2306.07967 (2023).

[5] J. L. Martinez-Rodriguez, I. López-Arévalo, A. B. Rios-Alvarado, Openie-based approach for knowledge graph construction from text, Expert Systems with Applications 113 (2018) 339–355. Publisher: Elsevier.

[6] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, Robust disambiguation of named entities in text, 2011, pp. 782–792.

[7] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach, Transactions of the Association for Computational Linguistics 2 (2014) 231–244. Publisher: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

[8] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, Y. Lu, Temporal Data Meets LLM–Explainable Financial Time Series Forecasting, arXiv preprint arXiv:2306.11025 (2023).

[9] A. H. Huang, H. Wang, Y. Yang, FinBERT: A large language model for extracting information from financial text, Contemporary Accounting Research 40 (2023) 806–841. Publisher: Wiley Online Library.

[10] H. Yang, X.-Y. Liu, C. D. Wang, Fingpt: Open-source financial large language models, arXiv preprint arXiv:2306.06031 (2023).

[11] N. Li, C. Gao, Y. Li, Q. Liao, Large language model-empowered agents for simulating macroeconomic activities, arXiv preprint arXiv:2310.10436 (2023).

[12] A. Cavallo, R. Rigobon, The billion prices project: Using online prices for measurement and research, Journal of Economic Perspectives 30 (2016) 151–178. Publisher: American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203-2418.

[13] A. Cavallo, M. Bertolotto, Filling the Gap in Argentina's Inflation Data, Available at SSRN 2782104 (2016).

[14] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, 2014, pp. 1532–1543.

[15] K. W. Church, Word2Vec, Natural Language Engineering 23 (2017) 155–162. Publisher: Cambridge University Press.

[16] J. H. Lau, T. Baldwin, An empirical evaluation of doc2vec with practical insights into document embedding generation, arXiv preprint arXiv:1607.05368 (2016).

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[20] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, Big bird: Transformers for longer sequences, Advances in neural information processing systems 33 (2020) 17283–17297.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[23] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[24] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, J. Launay, The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only, arXiv preprint arXiv:2306.01116 (2023).

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[26] Z. Ji, Y. Tiezheng, Y. Xu, N. Lee, E. Ishii, P. Fung, Towards mitigating LLM hallucination via self reflection, 2023.

[27] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, L. Yuan, Llm lies: Hallucinations are not bugs, but features as adversarial examples, arXiv preprint arXiv:2310.01469 (2023).

[28] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?⊠, 2021, pp. 610–623.