Dólares or Dollars? Unraveling the Bilingual Prowess of Financial LLMs Between Spanish and English

Xiao Zhang
The Fin AI
Singapore, Singapore
xiao.zhang@thefin.ai

Ruoyu Xiang
The Fin AI
Singapore, Singapore
ruoyu.xiang@thefin.ai

Chenhan Yuan
The University of Manchester
Manchester, UK
chenhan.yuan@postgrad.manchester.ac.uk

Duanyu Feng
Sichuan University
Chengdu, Sichuan, China
fengduanyu@stu.scu.edu.cn

Weiguang Han
Wuhan University
Wuhan, Hubei, China
han.wei.guang@whu.edu.cn

Alejandro Lopez-Lira
University of Florida
Gainesville, USA
alejandro.lopez-lira@warrington.ufl.edu

Xiao-Yang Liu
Columbia University
New York, NY, USA
XL2427@columbia.edu

Meikang Qiu
Augusta University
Augusta, USA
mqiu@augusta.edu

Sophia Ananiadou
Department of Computer Science,
The University of Manchester
Manchester, UK
Artificial Intelligence Research Centre
Tokyo, Japan
Archimedes/Athena Research Centre
Athens, Greece
sophia.ananiadou@manchester.ac.uk

Min Peng
Wuhan University
Wuhan, Hubei, China
pengm@whu.edu.cn

Jimin Huang
The Fin AI
Singapore, Singapore
jimin.huang@thefin.ai

Qianqian Xie*
The Fin AI
Singapore, Singapore
qianqian.xie@thefin.ai

ABSTRACT

Despite Spanish's pivotal role in the global finance industry, a pronounced gap exists in Spanish financial natural language processing (NLP) and application studies compared to English, especially in the era of large language models (LLMs). To bridge this gap, we unveil Toisón de Oro, the first bilingual framework that establishes instruction datasets, finetuned LLMs, and evaluation benchmark for financial LLMs in Spanish joint with English. We construct a rigorously curated bilingual instruction dataset including over 144K Spanish and English samples from 15 datasets covering 7 tasks. Harnessing this, we introduce FinMA-ES, an LLM designed for bilingual financial applications. We evaluate our model and existing LLMs using FLARE-ES, the first comprehensive bilingual evaluation benchmark with 21 datasets covering 9 tasks. The FLARE-ES benchmark results reveal a significant multilingual performance gap and bias in existing LLMs. FinMA-ES models surpass SOTA LLMs such as GPT-4 in Spanish financial tasks, due to strategic instruction tuning and leveraging data from diverse linguistic resources, highlighting the positive impact of cross-linguistic transfer. All our datasets, models, and benchmarks have been released[1].

CCS CONCEPTS

• **Applied computing** → **Economics**; • **Computing methodologies** → **Natural language processing**; **Language resources**.

KEYWORDS

Spanish; Bilingual; Large Language Models; Financial NLP

ACM Reference Format:
Xiao Zhang, Ruoyu Xiang, Chenhan Yuan, Duanyu Feng, Weiguang Han, Alejandro Lopez-Lira, Xiao-Yang Liu, Meikang Qiu, Sophia Ananiadou, Min Peng, Jimin Huang, and Qianqian Xie. 2024. Dólares or Dollars? Unraveling the Bilingual Prowess of Financial LLMs Between Spanish and English. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671554

*Corresponding author

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671554

[1] https://github.com/The-FinAI/PIXIU

1 INTRODUCTION

In light of digital transformations in the finance sector, recognizing the significance of global languages becomes paramount. A salient observation here is the substantial number of Spanish speakers. There are 485 million native Spanish speakers globally, asserting its position as the fourth most spoken language[14]. The U.S., a hub for fintech innovations, hosts 15 million individuals who speak Spanish as a secondary language. Additionally, the expanding digital user base necessitates that financial platforms consider such linguistic demographics, especially with Spanish speakers poised for significant growth by 2030 [11].

The integration of artificial intelligence (AI) in financial technologies (FinTech) has significantly accelerated advancements in the sector, particularly through the application of pre-trained language models (PLMs) [9] and recent large language models (LLMs) [24, 33] in natural language processing (NLP). These models, trained on vast amounts of text data, have the potential to understand, interpret, and generate human-like text, thereby transforming the way financial information is analyzed and processed [36]. They have been pivotal in transforming financial services, enabling sophisticated capabilities ranging from stock price forecasting to comprehensive financial analytics [17, 18, 38, 40].

Despite these advancements, a notable language disparity persists in the realm of FinTech [3, 10, 17, 18, 39]. The development and application of financial PLMs, including models like FinBERT [3], FLANG [29], and BloombergGPT [29], have predominantly concentrated on English. The recent financial LLM BloombergGPT [38], FinGPT [35] and PIXIU [40], continue this English-centric trend. Although there are attempts to adapt financial LLMs to other languages, such as Chinese, through models like DISC-FinLLM [4] and CFGPT [16], a significant gap remains in the development of Spanish-focused financial LLMs.

To address it, we propose Toisón de Oro, a novel bilingual framework meticulously designed to bridge the research gap. "Toisón de Oro" includes the first open-source multi-task and Spanish-English bilingual instructional data (FIT-ES) with 151k samples from 15 datasets covering 7 tasks, the first open-source Spanish-English bilingual evaluation benchmark (FLARE-ES) with 21 datasets covering 9 tasks, and the pioneering open-source Spanish-English bilingual financial LLM, FinMA-ES. This model is derived by finetuning LLaMA2 7B model [33] using FIT-ES. The contributions of "Toisón de Oro" are manifold: 1) We provide open access to all components, promoting widespread adoption and further innovation in the financial NLP sector. 2) Our bilingual solution addresses the critical need for multilingual financial analysis, enabling global firms to navigate and analyze diverse data sets effectively. 3) The scalable multi-task framework of "Toisón de Oro" supports a broad spectrum of financial applications, facilitating a unified and efficient approach to financial analytics.

To build the multi-task and bilingual instruction data, we sourced from Spanish and English financial tasks such as sentiment analysis, news headline classification, annual report text summarization, and examination question answering. This led to the creation of Financial Instruction Tuning Data in Spanish and English (FIT-ES), combining task-specific Spanish instructions with the respective data samples. In pursuit of enhancing bilingual financial analytics, we propose the bilingual financial LLM, FinMA-ES by finetuning LLaMA2 7B with FIT-ES. For model evaluation, we further build the bilingual FLARE-ES benchmark by including validationa and test set from FIT-ES, and extra 6 unseen datasets and 2 unseen tasks. This deliberate inclusion is aimed at evaluating the model's generalization capabilities across a diverse array of financial contexts and tasks.

We evaluate the performance of FinMA and exisiting SOTA LLMs with FLARE-ES. Our experimental analysis revealed three critical insights: 1) Existing LLMs, including GPT-4, demonstrate a pronounced performance gap in Spanish financial tasks, highlighting a significant disparity in their effectiveness across languages. 2) The FinMA-ES models excel in bilingual financial analysis, surpassing established models like GPT-4 in key Spanish financial tasks. This superiority underscores the pivotal role of instruction tuning, employing both target language and high-resource language datasets to enhance model capabilities. 3) Interestingly, fine-tuning LLMs with data from low-resource languages not only addresses gaps in those languages but also unexpectedly boosts the models' performance in high-resource language datasets, suggesting a beneficial cross-linguistic transfer effect.

Our contributions can be summarized as follows: 1) We created the first bilingual framework specifically designed for Spanish-English financial NLP and prediction tasks. 2) We developed the first bilingual and multi-task instruction tuning data. 3) We developed and fine-tuned the FinMA-ES model, the first LLM optimized for processing and understanding bilingual financial data on both Spanish and English. 4) We established the FLARE-ES benchmark, the first open-source comprehensive set of evaluations that allows for the cross-lingual assessment of models on both Spanish financial tasks and English financial tasks. 5) Our FLARE-ES benchmark evaluation indicates that FinMA-ES models notably outperform leading LLMs like GPT-4 in Spanish financial tasks due to strategic instruction tuning and data from both low- and high-resource languages, revealing a critical multilingual performance disparity and the unexpected benefits of cross-linguistic transfer.

2 RELATED WORK

Financial Language Models There is a pronounced absence of models, both PLMs and LLMs, specifically designed for Spanish or bilingual Spanish-English applications. Models such as finBERT [3] and FLANG [29] excel in processing English financial texts but offer limited utility in cross-lingual scenarios. Similarly, the advanced BloombergGPT [38], despite its massive scale, perpetuates this English-centric approach. Other recent developments, like FinGPT [35], InvestLM [42], and PIXIU [40], continue to focus predominantly on English. While there have been efforts to create financial LLMs for other languages, such as Chinese, with models like DISC-FinLLM [4] and CFGPT [16], Spanish remains notably underserved. This significant gap in Spanish and bilingual financial language models highlights the unique importance and potential impact of our work in developing bilingual financial LLMs, aiming to bridge this linguistic divide and cater to a wider, more diverse audience in the financial domain.

Financial Evaluation Benchmark In the sphere of financial NLP, there has been a significant focus on developing benchmarks

for English and Chinese. [29] introduced the FLUE benchmark, offering a diverse set of financial NLP tasks in English. Complementing this, [40] developed FLARE, another English benchmark for evaluating financial LLMs with a broad range of tasks. [35] also proposed an English benchmark for financial LLMs. In the Chinese context, the BBT-CFLEB benchmark by [19] and FinEval by [43] have been instrumental in advancing Chinese financial NLP. Further contributing to this are DISC-FinLLM by [4], which offer unique perspectives and tasks for evaluating Chinese financial language models. Despite these strides in English and Chinese benchmarks, the absence of comprehensive Spanish financial NLP benchmarks is evident, underscoring a significant gap in the field.

Open Sourced Large Language Models In the current landscape of AI democratization, while general models like LLaMA [33] and its instruction-following variants, Alpaca [32], and Vicuna-13B [7], have shown significant progress, the development in financial LLMs exhibits a language bias. English financial LLMs have been developed, as seen in PIXIU [40] and FinGPT [35], and Chinese financial LLMs have advanced with DISC-FinLLM [4], CFGPT [16], and InvestLM [42]. Lince-zero [8] represents progress for Spanish LLMs. However, this progression still leaves a notable gap in open-sourced Spanish and bilingual financial LLMs, underscoring a critical area for the global financial industry.

3 METHOD

3.1 FIT-ES: Financial Instruction Tuning Dataset-Encompassing Spanish

In this section, we present the composition and development of our Financial Instruction Tuning dataset, FIT-ES, which is the foundation for our Spanish financial LLM. We detail the origins of the raw data, enumerate the specific tasks encompassed within FIT-ES, and describe the meticulous process employed to construct the dataset from this raw data. Unique among existing resources, FIT-ES distinguishes itself as the first instruction-tuning dataset specifically crafted for Spanish financial LLMs.

3.1.1 Raw Data. Our Spanish instruction tuning dataset, is developed from publicly available sources, encompassing a range of datasets for financial NLP tasks and examination content. This dataset is rooted in authentic finance-related scenarios and benefits from the high-quality annotations typically provided by domain experts in open-sourced data. As shown in Table 1, this dataset includes 15 datasets for 7 financial NLP and prediction tasks in both Spanish and English. This bilingual data selection strategy aims to enhance the cross-lingual capabilities of our financial language models and ensure a balanced representation of both languages in the training process.

Classification. The task integrates two distinct datasets, MultiFin [13] and Gold news headline [30], to assess the model's classification prowess across both Spanish and English financial texts. The MultiFin dataset, focusing on Spanish headlines, compiles 2,066 articles spanning six critical financial categories: "Business & Management", "Finance", "Government & Controls," "Industry," "Tax & Accounting," and "Technology," challenging the model to accurately categorize each headline into its respective sector. Concurrently, the Gold news headline dataset delves into English financial texts concerning gold, encompassing a period from 2000 to 2019. It features a detailed classification scheme with nine tags: "price or not," "price up," "price down," "price stable," "past price," "future price," "past general," "future general," and "asset comparison," aimed at dissecting the headlines into binary classifications based on their implied price movement or market sentiment. This comprehensive classification task not only tests the model's linguistic flexibility and sector-specific knowledge across two languages but also its ability to discern and predict market trends from textual data, reflecting its applicability in automated financial news analysis.

Question Answering. Our investigation extends into the domain of question answering (QA), a critical task for evaluating the model's comprehension and application of financial knowledge across both Spanish and English datasets. In the Spanish context, we utilize the EFP and EFPA datasets[5], which consist of questions derived from financial examinations provided by official examiner associations. The EFP dataset challenges the model with 37 questions, each offering three possible answers ("A," "B," or "C"), thereby testing the model's proficiency in accurately identifying the correct response based on the given financial scenario. The EFPA dataset further expands this challenge, presenting 228 questions with four answer choices ("A," "B," "C," or "D"), encompassing a wider array of financial topics, including economic knowledge, fundamental financial concepts, and detailed computations related to financial products. Transitioning to the English datasets, we employ FinQA [5] and ConvFinQA [6] to assess the model's QA capabilities in a different linguistic and financial context. The FinQA dataset, comprising 8,281 questions and answers extracted from earnings reports, necessitates the model to navigate through complex text and table data to derive accurate answers, reflecting its ability to handle multifaceted financial documents. Similarly, the ConvFinQA dataset, with 3,892 questions and answers based on earnings reports, challenges the model to understand and respond to queries within a conversational context, highlighting its capacity for nuanced language understanding and information retrieval within financial discussions.

Text Summarization. The task of text summarization within our study focuses on condensing voluminous financial documents into concise, informative abstracts, a critical capability for enhancing the accessibility and usability of financial information. This task employs the FNS-2023 dataset[6], which comprises a collection of 232 annual reports from various financial companies. These reports, rich in detailed financial data and narratives, present a unique challenge: to distill the essence of each document into a summary that captures the most crucial information while maintaining the factual integrity and coherence of the original text.

Financial Sentiment Analysis. This task delves into the intricate sentiment dynamics within financial texts, employing the TSA [25] and FinanceES [1] datasets for Spanish sentiment analysis, and the FPB and FiQA-SA datasets for English. The TSA dataset, comprising 3,892 entries from financial news and tweets, is meticulously annotated to reflect sentiments as positive, negative, or neutral, providing a nuanced spectrum of market emotions in Spanish. Similarly, the FinanceES dataset enriches this analysis with

[5]https://efpa-eu.org/
[6]https://wp.lancs.ac.uk/cfie/fns2023/

Table 1: The details of the raw data and instruction data.

Data	Task	Language	Raw	Instruction	Data Types	Modalities	License
MultiFin [13]	classification	Spanish	2,066	2,066	article headlines	text	MIT License
FNS-2023[2]	text summrization	Spanish	232	232	annual reports	text	Public
EFP[3]	question answering	Spanish	37	37	exam questions	text	Public
EFPA [4]	question answering	Spanish	228	228	exam questions	text	Public
TSA [25]	sentiment analysis	Spanish	3,892	3,892	news headlines	text	Public
FinanceES [1]	sentiment analysis	Spanish	7,980	7,980	news headlines	text	Public
FPB	sentiment analysis	English	4,845	48,450	news	text	CC BY-SA 3.0
FiQA-SA [20]	sentiment analysis	English	1,173	11,730	news headlines, tweets	text	Public
Headline [30]	news headline classification	English	11,412	11,412	news headlines	text	CC BY-SA 3.0
NER [2]	named entity recognition	English	1,366	13,660	financial agreements	text	CC BY-SA 3.0
FinQA [5]	question answering	English	8,281	8,281	earnings reports	text, table	MIT License
ConvFinQA [6]	question answering	English	3,892	3,892	earnings reports	text, table	MIT License
BigData22 [31]	stock movement prediction	English	7,164	7,164	tweets, historical prices	text, time series	Public
ACL18 [41]	stock movement prediction	English	27,053	27,053	tweets, historical prices	text, time series	MIT License
CIKM18 [37]	stock movement prediction	English	4,967	4,967	tweets, historical prices	text, time series	Public

7,980 Spanish financial news headlines, each labeled to indicate the underlying sentiment, thus offering a comprehensive base for evaluating the model's sentiment detection accuracy in Spanish financial discourse. Transitioning to English, the FPB [21] dataset introduces an expansive collection of 4,845 news items, with sentiment labels that challenge the model's ability to discern and classify sentiments in English financial news. Complementing this, the FiQA-SA [20] dataset includes 1,173 entries combining news headlines and tweets, each with sentiment annotations, further broadening the scope of sentiment analysis in English financial texts. This dataset not only tests the model's semantic understanding but also its capacity to navigate the subtleties of sentiment expression in diverse formats, from concise tweets to more detailed news articles.

English-Only Financial Tasks. Our study includes English-specific tasks leveraging datasets for named entity recognition (NER) [2], and stock movement prediction to evaluate the model's financial analysis capabilities. The NER task uses a dataset of 1,366 financial agreements to test entity identification within financial texts. For stock movement prediction, three datasets—BigData22 [31] with 7,164 entries, ACL18 [41] featuring 27,053 entries, and CIKM18 [37] comprising 4,967 entries—challenge the model to forecast stock prices based on textual and quantitative data. These tasks assess the model's proficiency in extracting critical information and predicting market trends, showcasing its utility in financial analytics and decision-making processes within the English financial domain.

3.1.2 Instruction Construction. We crafted financial instruction datasets from the raw data outlined in Table 1, with carefully designed instructions by domain experts who are proficient in both Spanish and English[7]. The construction of instruction tuning samples follows a general structured template:

Instruction: [task prompt] Text: [input] Response: [output]

which integrates human-designed instructions with input texts and their corresponding outputs. [task prompt] is the prompt designed for each data, [input text] is the input financial data from each data, e.g. the historical prices and tweets or headlines, [output] is the corresponding output for input text, e.g. sentiment label of input text from ["Positive", "Negative", "Neutral"] and ["positivo", "negativo", "neutral"] in Spanish.

MultiFin, FNS-2023, FinanceES, and TSA datasets adopted a unified approach due to their similar task structures and data types. However, the EFP and EFPA datasets required two distinct sets of instructions Customized to their respective answer choices.

3.2 FinMA-ES: Financial Large Language Model in Both English and Spanish

We propose the **FinMA-ES-Bilingual**, a bilingual financial large language model, through fine-tuning the LLaMA2-7B backbone model [34], specifically aimed at enhancing performance in both English and Spanish financial tasks based on FIT-ES. This fine-tuning involved 5 epochs using the AdamW optimizer [15], characterized by a batch size of 1, a learning rate of 3e-4, and a weight decay set to 1e-5. The entire process was executed on a robust computational framework provided by 2 NVIDIA HGX A100 SXM4 80GB GPUs. We also proposed the **FinMA-ES-Spanish** which is only finetuned with the Spanish data, for conducting the ablation study. This additional analysis aims to evaluate the contribution of Spanish training data towarding the overall model performance, highlighting the value of language-specific data in enhancing FinMA-ES's bilingual capabilities.

3.3 FLARE-ES: Financial Evaluation Benchmark on Spanish and English

We further propose the FLARE-ES evaluation benchmark with 21 datasets from 11 tasks, to holistically evaluate the capabilities of LLMs in the financial domain, covering both English and Spanish languages. Each task is designed to probe different aspects of financial data understanding and generation, and the financial prediction, utilizing specific metrics for a detailed assessment, as shown in Table 2 and Figure 1.

3.3.1 Evaluation Tasks and Datasets. To thoroughly assess models' performance, we incorporate datasets from the same sources used in training, as well as additional 6 datasets and 2 tasks not

[7]For detailed instruction, please see Appendix A

Table 2: The details of our evaluation datasets and evaluation metrics. In order to compare performance across different models, such as GPT-4 and ChatGPT, we maintain consistency by using identical datasets with the same data distributions across all models during training.

Data	Task	Language	Valid	Test	Evaluation
MultiFin [13]	classification	Spanish	230	368	F1, Accuracy
FNS-2023[8]	text summarization	Spanish	36	50	rouge1, rouge2, rougeL
EFP[9]	question answering	Spanish	5	210	F1, Accuracy
EFPA[10]	question answering	Spanish	35	50	F1, Accuracy
TSA [25]	sentiment analysis	Spanish	200	726	F1, Accuracy
FinanceES [1]	sentiment analysis	Spanish	1,272	1,621	F1, Accuracy
FPB [21]	sentiment analysis	English	775	970	F1, Accuracy
FiQA-SA [20]	sentiment analysis	English	188	235	F1
Headlines [30]	news headline classification	English	1,141	2,283	Avg F1
NER [2]	named entity recognition	English	103	980	Entity F1
FinQA [5]	question answering	English	883	1,147	EM Accuracy
ConvFinQA [6]	question answering	English	2,210	1,490	EM Accuracy
BigData22 [31]	stock movement prediction	English	798	1,470	Accuracy, MCC
ACL18 [41]	stock movement prediction	English	2,560	3,720	Accuracy, MCC
CIKM18 [37]	stock movement prediction	English	431	1,140	Accuracy, MCC
FiNER-ORD [28]	named entity recognition	English	-	1080	Entity F1
ECTSum [23]	text summarization	English	-	495	ROUGE, BERTScore, BARTScore
EDTSum [44]	text summarization	English	-	2000	ROUGE, BERTScore, BARTScore
German [12]	credit scoring	English	-	1000	F1, MCC
Australian [26]	credit scoring	English	-	690	F1, MCC
FOMC [27]	hawkish-dovish classification	English	-	496	F1, Accuracy

employed during the training phase. This approach is designed to rigorously evaluate the generalization capabilities of various models in financial tasks.

1) Sentiment analysis. We utilize the TSA and FinanceES [1] datasets for Spanish sentiment analysis, alongside the FPB [21] and FiQA-SA [20] datasets for English, for evaluating models' abilities to discern and categorize sentiments as positive, negative, or neutral. Following previous works [39, 41], performance across these datasets is meticulously quantified using Accuracy (ACC) and the F1 Score.

2) Classification. We leverage the MultiFin dataset for Spanish, containing 230 validation and 368 test samples, alongside the English Headlines dataset, which includes 1,141 validation and 2,283 test samples. These datasets challenge LLMs to accurately classify financial news articles into predefined categories reflecting the model's understanding of domain-specific content. Performance is evaluated using the F1 Score and Accuracy metrics.

3) Text sumarization. This task utilizes the Spanish FNS-2023 dataset[11], consisting of 36 validation and 50 test samples, alongside English datasets including ECTSum and EDTSum derived from additional sources not utilized in training. To quantitatively measure the quality of the model-generated summaries against reference summaries, we employ ROUGE scores (Recall-Oriented Understudy for Gisting Evaluation), BERTScore, an automated evaluation metric for text generation, and BARTScore, which is a superior text generation evaluation metric, excelling in 16 of 22 tests, with accessible code and interactive leaderboard for comprehensive assessment.). ROUGE metrics provide a comprehensive analysis of summarization performance by measuring unigram overlap (rouge1), bigram overlap (rouge2), and the longest common subsequence (rougeL). BERTScore, utilizing BERT's contextual embeddings, further refines this evaluation by comparing the semantic similarity between generated and reference texts, allowing for a nuanced understanding of content quality beyond mere textual overlap.

4) Question answering. This task utilizes the EFP and EFPA datasets for Spanish, featuring 5 and 35 validation samples, and 210 and 50 test samples, respectively, along with English datasets FinQA and ConvFinQA. This task assesses models on their ability to retrieve or generate accurate answers to financial questions based on provided information. Performance is evaluated through F1 Score and Accuracy for Spanish datasets [39, 41]. For English datasets, EM (Exact Match) Accuracy is used, focusing on the model's ability to produce answers that exactly match the gold standard responses.

5) Named entity recognition. This task is to identify essential entities in the financial domain, namely individuals, companies, and geographic locations. These entities play a crucial role in constructing comprehensive financial knowledge graphs and we utilized the FIN dataset [2], which comprises sentences extracted from publicly available financial agreements found in U.S. Security and Exchange

[11] https://wp.lancs.ac.uk/cfie/fns2023/

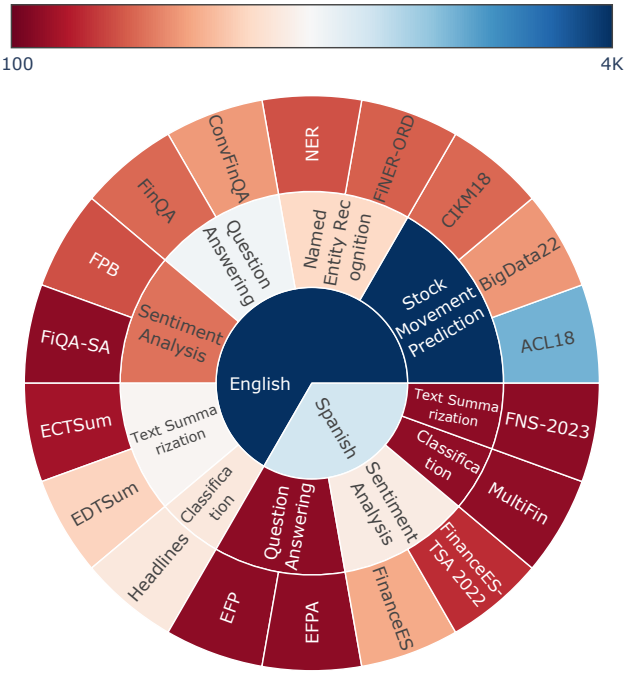

Figure 1: An overview of bilingual benchmark FLARE-ES.

Commission (SEC) filings. Additionally, we have manually annotated the entity types, categorizing them into LOCATION (LOC), ORGANISATION (ORG), and PERSON (PER) classifications.

6) Credit scoring. This task focuses on evaluating performance using specific datasets related to German and Australian Credit Scoring. It aims to assess how well models perform in predicting credit scores based on these datasets, employing metrics such as F1 Score and Matthews Correlation Coefficient (MCC) [39] for evaluation.

7) Hawkish-dovish classification. The Hawkish-Dovish classification distinguishes sentences in financial texts as "hawkish" or "dovish", requiring intricate comprehension of monetary policy language beyond conventional sentiment analysis. This approach utilizes the FOMC dataset [27], where Federal Open Market Committee (FOMC) meeting sentences are carefully annotated to reflect their economic stance. Similary to sentiment analysis, we use F1 and Accuracy metrics for evaluation.

4 EXPERIMENTS

The proposed FIT-ES and FLARE-ES allow us to train, select the model, and evaluate the performance of LLMs on financial understanding and predictions. This section investigates how powerful our fine-tuned models and other LLMs are on FLARE-ES. We compare FinMA-ES with the following baselines:

(1) GPT-4 [24]. A powerful instruction following large language model with around 1T parameters proposed by OpenAI.
(2) ChatGPT[12]. An instruction following large language model with 175B parameters from OpenAI.
(3) LLaMA2-7B [34]. An open-sourced large language model by META with 7B parameters.
(4) Falcon-7B[13]. A causal decoder-only model built by TII with 7B parameters.
(5) Bloomz-7B1-mt [22]. An open-access multilingual large language model with 7B parameters.
(6) Lince-zero[14]. An open-sourced Spanish-instruction tuned large language model based on Falcon-7B.
(7) FinMA-7B-full [40]. An open-sourced English financial instruction tuned large language model based on LLaMA2 with 7B parameters.
(8) FinMA-30B-nlp [40]. An open-sourced English financial large language model fine-tuned using only financial NLP tasks based on LLaMA2 with 30B parameters.

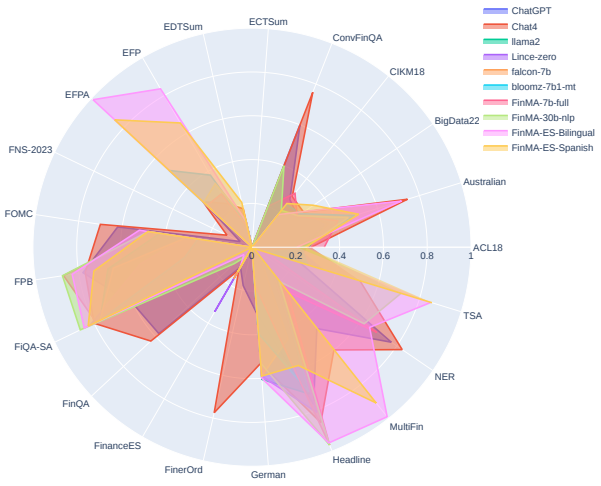

Figure 2: The radar graph of the performance for all methods on FLARE-ES.

4.1 Results

4.1.1 Overall Performance. Table 3 and Figure 2 presents a detailed comparative performance analysis of our FinMA-ES models against other leading large language models (LLMs) on the FLARE-ES benchmark. In the realm of Spanish financial tasks, the FinMA-ES-Bilingual and FinMA-ES-Spanish models, both with a 7-billion parameter count, demonstrate superior performance against other state-of-the-art (SOTA) models, including significantly larger models like GPT-4, in four out of six datasets. These datasets include MultiFin, EFP, EFPA, and TSA. The standout performance of our models on these tasks highlights the substantial impact of instruction fine-tuning, which is specifically tailored to enhance

[12]https://openai.com/blog/chatgpt
[13]https://huggingface.co/tiiuae/falcon-7b
[14]https://huggingface.co/clibrain/lince-zero

Table 3: The performance of different LLMs on the FLARE-ES benchmark. Results of ChatGPT, GPT-4, LLaMA2, Lince-zero, FinMA-7B-full, Falcon-7B, Bloomz-7B1-mt, FinMA-7B-full, and FinMA-30B-nlp. Test datasets were built to be the same data on every LLM.

Dataset	Metrics	Chat GPT	GPT 4	LLaMA 2-7B	Lince-zero	Falcon-7B	Bloomz-7B1-mt	FinMA-7B-full	FinMA-30B-nlp	FinMA-ES-Bilingual	FinMA-ES-Spanish
MultiFin	Acc	0.48	0.6	0.23	0.22	0.05	0.23	0.25	0.21	**0.99**	0.91
	F1	0.47	0.6	0.11	0.1	0.07	0.16	0.27	0.19	**0.99**	0.91
EFP	Acc	0.30	0.27	0.27	0.27	0.27	0.38	0.35	0.38	**0.84**	0.65
	F1	0.26	0.19	0.12	0.12	0.12	0.38	0.21	0.29	**0.83**	0.66
EFPA	Acc	0.31	0.34	0.26	0.25	0.26	0.51	0.35	0.34	**0.99**	0.86
	F1	0.25	0.27	0.1	0.1	0.10	0.52	0.21	0.26	**0.99**	0.85
FNS-2023	rouge1	0.02	**0.19**	0	0	0	0	0.01	0	0	0
	rouge2	0.04	**0.06**	0	0	0	0	0	0	0	0
	rougeL	0.12	**0.13**	0	0	0	0	0	0	0	0
TSA	Acc	0.21	0.47	0.07	0.32	0.06	0.22	0.04	0.67	0.85	**0.86**
	F1	0.24	0.56	0.04	0.36	0.10	0.32	0.07	0.76	**0.86**	0.86
FinanceES	Acc	0.13	0.15	0.14	**0.39**	0.15	0.03	0.02	0.03	0.11	0.13
	F1	0.08	0.09	0.13	**0.29**	0.18	0.04	0.03	0.06	0.11	0.13
FPB	Acc	0.78	0.76	0.68	0.51	0.64	0.39	0.87	**0.87**	0.83	0.73
	F1	0.78	0.78	0.65	0.52	0.64	0.23	0.87	**0.88**	0.83	0.73
FiQA-SA	F1	0.6	0.8	0.77	0.82	0.77	0.77	0.79	**0.87**	0.85	0.83
Headline	AvgF1	0.77	0.86	0.72	0.81	0.45	0.71	0.97	**0.97**	0.96	0.58
NER	EntityF1	0.77	**0.83**	0	0	0	0	0.69	0.62	0.65	0.01
FinQA	EmAcc	0.58	**0.63**	0	0	0.002	0	0.04	0.11	0.05	0
ConvFinQA	EmAcc	0.60	**0.76**	0	0	0	0	0.20	0.40	0	0
BigData22	Acc	0.53	0.54	0.51	0.55	0.55	0.55	0.49	0.47	0.48	**0.57**
	MCC	-0.025	0.03	0.030	0.000	0.000	-0.007	0.010	0.040	0.100	**0.110**
ACL18	Acc	0.50	0.52	0.51	0.47	0.51	0.50	**0.56**	0.49	0.49	0.50
	MCC	0.005	0.020	0.010	-0.060	-0.004	-0.040	**0.100**	0.000	-0.080	-0.010
CIKM18	Acc	0.55	**0.57**	0.47	0.43	0.44	0.55	0.53	0.43	0.42	0.55
	MCC	0.005	0.020	-0.070	0.010	-0.010	-0.050	**0.100**	0.000	-0.040	-0.040
FinerOrd	EntityF1	0.28	**0.77**	0	0	0	0	0	0	0	0
	F1	0.08	**0.78**	0	0	0	0	0	0	0	0
ECTSum	rouge1	0	0	0	0	0	0	0	0	0	0
	rouge2	0	0	0	0	0	0	0	0	0	0
	rougeL	0	0	0	0	0	0	0	0	0	0
EDTSum	rouge1	0.17	0.2	0.13	0.07	0.15	0.12	0.13	0.17	0.15	**0.26**
	rouge2	0.08	**0.19**	0.06	0.03	0.06	0.06	0.06	0.08	0.07	0.14
	rougeL	0.13	0.15	0.12	0.07	0.13	0.12	0.10	0.14	0.14	**0.23**
German	Acc	0.2	0.55	0.61	**0.66**	0.66	0.39	0.17	0.53	0.60	0.66
	F1	0.41	0.513	**0.60**	0.52	0.52	0.40	0.17	0.53	0.60	0.52
Australian	Acc	0.41	**0.74**	0.43	0.43	0.47	0.57	0.41	0.46	0.72	0.56
	F1	0.26	**0.75**	0.26	0.26	0.26	0.41	0.41	0.46	0.71	0.51
FOMC	Acc	0.6	**0.69**	0.50	0.33	0.30	0.30	0.46	0.43	0.55	0.50
	F1	0.64	**0.71**	0.35	0.28	0.30	0.20	0.49	0.53	0.49	0.46

the models' understanding and generation of Spanish financial language nuances.

Despite the strong results in most datasets, the FinanceES dataset presents a more competitive scenario, where our models perform

on par with SOTA LLMs like GPT-4. However, they do not outperform Lince-zero, which benefits from instruction tuning with a large corpus of Spanish data. This suggests that while our models are highly competitive, there is a unique advantage inherent to models fine-tuned on extensive general domain Spanish data, indicating room for further optimization in future model iterations. The FNS-2023 dataset, focusing on text summarization, reveals a universal challenge for most models, with only GPT-4 achieving notable success. This might indicate an area where specialized training or model architectures are required to handle the complexities of summarization tasks effectively. Our models exhibit their robustness within the English datasets, achieving the best results in the FinanceES dataset and comparable performance to SOTA models such as FinMA-30B-nlp and GPT-4 in eight others.

4.1.2 Generalization Ability. Notably, in three of the six leave-out English datasets—German, Australian, and FOMC—our models emerge as the top performers among all open-source LLMs. Moreover, they maintain competitive performance with renowned models like GPT-4 and ChatGPT in four datasets. This underscores the adaptability and generalization capacity of our models across diverse financial contexts and languages. For the FinerORD and ECTSum datasets, we implemented a complex prompt design inspired by the Pixiu paper [40] which involves direct generation of label sequences. This advanced methodological approach did not yield the desired performance for any model except GPT-4 in the FinerOrd dataset, and none of the models managed to effectively tackle the ECTSum dataset. These results underscore the challenges in generating accurate label sequences and point towards the need for more innovative approaches or specialized training to overcome these hurdles.

4.1.3 Language Disparity. Table 3 elucidates a marked language disparity when evaluating the proficiency of existing LLMs, including the renowned GPT-4, across Spanish and English financial tasks. This table starkly underscores the challenges LLMs face with Spanish financial tasks, underscoring a considerable performance gap that persists despite advancements in the field. The data reveals that while these LLMs are adept at handling English financial tasks, they falter significantly when it comes to Spanish. The underperformance of these models in Spanish is indicative of the fact that while LLMs have made significant strides, their proficiency is unevenly distributed across languages, with a clear bias towards English. The performance of our FinMA-ES models shows a significant improvement over other LLMs in Spanish tasks, which can be attributed to the instruction tuning performed with datasets in the target language. This suggests that, to address the language disparity in LLMs effectively, a dedicated effort to develop and fine-tune models with a rich and diverse dataset in the target language is essential.

4.1.4 Ablation study. The ablation study focuses on comparing the performance of monolingual and bilingual models across different datasets. From Table 3, we can see the FinMA-ES-Bilingual model, which leverages both Spanish and English data, exhibits a distinct advantage over its monolingual counterpart, the FinMA-ES-Spanish model. This superior performance is evidenced in three out of six datasets focused on Spanish financial tasks. The bilingual model's proficiency suggests that exposure to English domain-specific datasets not only reinforces but also enhances its performance in Spanish financial contexts. Further analysis shows that the FinMA-ES-Bilingual model outperforms the monolingual model in six out of nine English datasets and in half of the leave-out English datasets. The underlying reason for this better performance is attributed to the bilingual model's integration of additional English financial instruction tuning data, which is not utilized by the FinMA-ES-Spanish model. This strategic use of cross-lingual data emphasizes the importance of diverse linguistic training in the development of more adaptable and proficient LLMs.

A pivotal aspect of this study involves the comparison between the FinMA-ES-Spanish model and its foundational model, LLamA2 13B. Intriguingly, the FinMA-ES-Spanish model surpasses LLamA2 13B across all English datasets. This finding is particularly enlightening as it illustrates that fine-tuning LLMs with low-resource language data does not hinder but actually enhances their performance in high-resource language datasets within the financial domain.

Moreover, a comparative review of Lince-zero and Falcon-7B reveals that Lince-zero, which is fine-tuned on a broader set of general domain Spanish data, surpasses Falcon-7B across nearly all Spanish datasets. This reinforces the hypothesis that general domain data in Spanish can significantly enhance model performance on Spanish financial tasks. However, the scenario reverses when it comes to English datasets, where Falcon-7B tends to perform better. This dichotomy underscores a potential trade-off, suggesting that while general domain Spanish data is beneficial for Spanish task performance, it may inadvertently impair the model's effectiveness in English financial tasks.

5 CONCLUSION

This paper addresses the linguistic disparity in the financial area by introducing the first bilingual LLMs framework for Spanish and English. Through the meticulous curation of over 144K bilingual instruction samples and the development of FinMA-ES, a model fine-tuned for bilingual financial analysis, we bridge a crucial gap in the field. Our efforts culminate in the FLARE-ES benchmark, a novel benchmark for comprehensive cross-lingual evaluations, which exposes significant performance gaps and biases in current LLMs. Notably, FinMA-ES demonstrates superior performance over existing SOTA LLMs, including GPT-4, in Spanish financial tasks by effectively leveraging strategic instruction tuning and a diverse dataset for cross-linguistic transfer. By releasing our datasets, models, and benchmarks, we seek to encourage further exploration into complex bilingual scenarios, aiming for more inclusive and effective financial NLP solutions. Looking ahead, we aim to further refine our models and evaluation methods by integrating datasets in additional languages, such as Japanese and Greek. We plan to collaborate with international partners and language experts to ensure the quality and representativeness of these datasets. This will allow us to explore broader applications and improve performance in complex bilingual scenarios.

6 LIMITATIONS

Despite the positive contributions of this study, we recognize the following limitations: 1) **Parameter Restriction**: FinMA-ES is developed with a cap of 7B parameters, a constraint dictated by our available computational resources, which has implications for its depth of training and overall efficacy. 2) **Evaluation Benchmark Diversity**: The model demonstrates a limited range in its evaluation benchmarks, particularly affecting its capability in tasks like financial summarization. 3) **Scope of Application**: The specific design and instructional approach of FinMA-ES might limit its applicability across varied bilingual scenarios. 4) **Ethical and Practical Concerns**: We must consider the potential for negative outcomes, such as disseminating inaccurate financial information or improper market influence. Therefore, we recommend utilizing FinMA-ES primarily for scholarly research, mindful of these ethical aspects. 5) **Languages**: In this study, we primarily focused on the bilingual performance of FinMA-ES in Spanish and English, which limits the discussion within these two languages.

ACKNOWLEDGMENTS

We would like to thank all the anonymous reviewers and area chairs for their comments. This work is supported by the project JPNP20006 from New Energy and Industrial Technology Development Organization (NEDO). This work has also been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. This work is also supported by National Science and Technology Major Project (No.2021ZD0113304), National Natural Science Foundation of China (U23A20316), Key R&D Project of Hubei Province (2021BAA029), General Program of Natural Science Foundation of China (NSFC) (Grant No.62072346), and founded by Joint&Laboratory on Credit Technology.

REFERENCES

[1] 2023. IBERLEF 2023 Task - FinancES. Financial Targeted Sentiment Analysis in Spanish.
[2] Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*. 84–90.
[3] Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063* (2019).
[4] Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. 2023. DISC-FinLLM: A Chinese Financial Large Language Model based on Multiple Experts Fine-tuning. arXiv:2310.15205 [cs.CL]
[5] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3697–3711.
[6] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849* (2022).
[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/
[8] clibrain.com. 2023. LINCE-ZERO: Llm for Instructions from Natural Corpus en Español. (2023).
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
[10] Weiguang Han, Boyi Zhang, Qianqian Xie, Min Peng, Yanzhao Lai, and Jimin Huang. 2023. Select and Trade: Towards Unified Pair Trading with Hierarchical Reinforcement Learning. *arXiv preprint arXiv:2301.10724* (2023).
[11] Erika Hoff and Krystal M Ribot. 2017. Language growth in English monolingual and Spanish-English bilingual children from 2.5 to 5 years. *The Journal of pediatrics* 190 (2017), 241–245.
[12] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.
[13] Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A Dataset for Multilingual Financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 894–909. https://doi.org/10.18653/v1/2023.findings-eacl.66
[14] George Julian. 2020. What are the most spoken languages in the world. *Retrieved May* 31, 2020 (2020), 38.
[15] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). https://api.semanticscholar.org/CorpusID:6628106
[16] Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. CFGPT: Chinese Financial Assistant with Large Language Model. arXiv:2309.10654 [cs.CL]
[17] Xianzhi Li, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? An Examination on Several Typical Tasks. *arXiv preprint arXiv:2305.05862* (2023).
[18] Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *arXiv preprint arXiv:2304.07619* (2023).
[19] Dakuan Lu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, Hengkui Wu, and Yanghua Xiao. 2023. BBT-Fin: Comprehensive Construction of Chinese Financial Domain Pre-trained Language Model, Corpus and Benchmark. *arXiv preprint arXiv:2302.09432* (2023).
[20] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*. 1941–1942.
[21] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65, 4 (2014), 782–796.
[22] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786* (2022).

[23] Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, et al. 2022. ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 10893–10906.
[24] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
[25] Ronghao Pan, José Antonio García-Díaz, Francisco García-Sánchez, and Rafael Valencia-García. 2023. Evaluation of transformer models for financial targeted sentiment analysis in Spanish. *PeerJ Computer Science* 9 (2023). https://api.semanticscholar.org/CorpusID:258596166
[26] Ross Quinlan. 1987. Statlog (Australian Credit Approval). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C59012.
[27] Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.
[28] Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157* (2023).
[29] Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain. *arXiv preprint arXiv:2211.00083* (2022).
[30] Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*. Springer, 589–601.
[31] Yejun Soun, Jaemin Yoo, Minyong Cho, Jihyeong Jeon, and U Kang. 2022. Accurate Stock Movement Prediction with Self-supervised Learning from Sparse Noisy Tweets. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 1691–1700.
[32] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.
[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[34] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv* abs/2307.09288 (2023). https://api.semanticscholar.org/CorpusID:259950998
[35] Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets. arXiv:2310.04793 [cs.CL]
[36] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).
[37] Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1627–1630.
[38] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).
[39] Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023. The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over MultiModal Stock Movement Prediction Challenges. *arXiv preprint arXiv:2304.05351* (2023).
[40] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. arXiv:2306.05443 [cs.CL]
[41] Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1970–1979.
[42] Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning. *arXiv preprint arXiv:2309.13064* (2023).
[43] Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, Zhoufan Zhu, Anbo Wu, Xin Guo, and Yun Chen. 2023. FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models. arXiv:2308.09975 [cs.CL]
[44] Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2114–2124.

A INSTRUCTIONS

Table 4: The example prompts for all Spanish datasets along with their corresponding prompts in Spanish and English translations. MultiFin is a classification task and includes article headlines as its text data type. Also, FNS-2023 is a text summarization task that includes annual reports text data type. EFP and EFPA are both question-answering tasks from financial exams in Spanish. TSA and FinanceES are both sentiment analysis tasks with {category}: negative, positive, neutral in English and {category}:negativo, positivo, neutral in Spnaish.

Dataset	Spanish Prompts	English Translations
FinanceES	"¿Cuál es el sentimiento de esta oración? Responde solo negativo, positivo o neutral. {category}:positivo, negativo, or neutral?"	What is the sentiment of this sentence? Answer only negative, positive or neutral.
TSA	"¿Cuál es el sentimiento de esta oración? Responde solo negativo, positivo o neutral. {category}:positivo, negativo, or neutral?"	What is the sentiment of this sentence? Answer only negative, positive or neutral.
FNS-2023	"Por favor, lea el texto con atención y resuma su contenido de forma breve y precisa."	Please read the text carefully and summarize the content of the text accurately and birefly.
EFP	"Lea cuidadosamente las preguntas y respuestas, y elija la que considere apropiada entre las tres opciones A, B y C."	Read the questions and answers carefully, and choose the option that you think is appropriate from the three options A, B and C.
EFPA	"Lea cuidadosamente las preguntas y respuestas, y elija la que considere apropiada entre las tres opciones A, B ,C y D. "	Read the questions and answers carefully, and choose the option that you think is appropriate from the three options A,B, C and D.
MultiFin	"Lee el texto cuidadosamente y elige la etiqueta adecuada para el texto de las etiquetas de 'Negocios y Gestión', 'Finanzas', 'Gobierno y Control', 'Industria', 'Impuestos y Contabilidad', 'Tecnología' {category}:'Negocios y Gestión', 'Finanzas', 'Gobierno y Control', 'Industria', 'Impuestos y Contabilidad', 'Tecnología' ".	Read the text carefully and choose one appropriate label for the text from the labels of 'Business and Management', 'Finance', 'Government and Controls', 'Industry', 'Tax and Accounting', 'Technology'.