Plutus: Benchmarking Large Language Models in Low-Resource Greek Finance

Anonymous ACL submission

Abstract

Despite Greece's pivotal role in the global economy, large language models (LLMs) remain underexplored for Greek financial context due to the linguistic complexity of Greek and the scarcity of domain-specific datasets. While multilingual financial NLP has revealed large performance gaps across languages, no benchmarks or LLMs have been tailored for Greek financial tasks until now. To bridge this gap, we introduce **Plutus-ben**, the first Greek Financial Evaluation Benchmark, and **Plutus-8B**, the first financial LLM fine-tuned on Greek-specific financial data. Plutus-ben addresses five core tasks: numeric/textual named entity recognition, question answering, abstractive summarization, and topic classification. To support these tasks, we release three new expert-annotated Greek financial datasets and incorporate two existing resources. Our comprehensive evaluation of 22 LLMs reveals persistent challenges in Greek financial NLP, driven by linguistic complexity, domain terminology, and financial reasoning gaps. Experiment results underscore the limitations of cross-lingual transfer and the need for Greek-specific financial modeling. We publicly release Plutus-ben, Plutus-8B, and all associated datasets[1] to promote reproducible research and advance multilingual financial NLP.

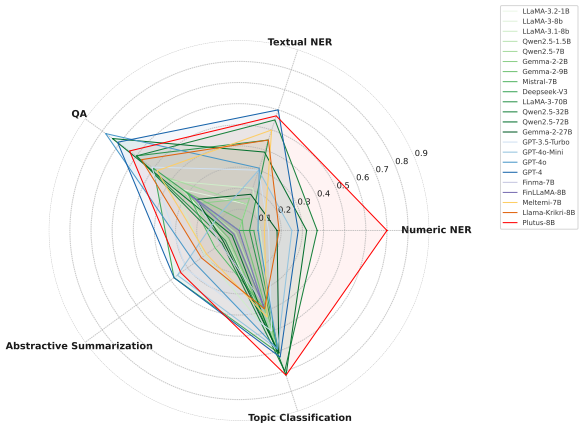

Figure 1: Radar graph of model performance on Plutus-ben, the first Greek financial benchmark. Plutus-8B achieves the best performance, surpassing GPT-4 by 15.38%, GPT-4o by 46.34%, and Deepseek-V3 by 93.55%.

1 Introduction

As an official language of the European Union[2] and the dominant language of Greece's merchant navy, which controls over 20% of the world's merchant fleet[3], Greek is central to international trade, banking, and regulatory affairs. Greek financial documents such as regulatory filings, maritime trade records, and economic reports hold substantial international relevance, yet their processing remains difficult (Esarey, 2020). Greek's complex morphology, inflectional system, and unique orthographic structures (Holton et al., 2012; Efthymiou and Koutsoukos) make it fundamentally different from high-resource financial languages such as English and Chinese. These linguistic complexities introduce challenges in financial information extraction, entity recognition, and numerical reasoning (Papantoniou and Tzitzikas, 2024).

Despite recent advancements in applying large language models (LLMs) to financial natural language processing (NLP) tasks, Greek remains largely unexplored. Extensive financial LLMs have been developed for English (Xie et al., 2024b; Wu et al., 2023; Xie et al., 2023a; Yang et al., 2023b,a), Chinese (Chen et al., 2023; Li et al., 2023), and Spanish (Zhang et al., 2024). Moreover, financial benchmarks have been established for En-

[1] We released all code in https://anonymous.4open.science/r/FinBen-379E/ and will release the datasets attached to the paper later.

[2] https://european-union.europa.eu/principles-countries-history/languages_en

[3] https://ugs.gr/en/greek-shipping-and-economy/greek-shipping-and-economy-2024/the-international-perspective/

glish (Xie et al., 2024a, 2023a; Shah and Chava, 2023), as well as for Chinese (Nie et al., 2024), Spanish (Zhang et al., 2024), and Japanese (Hirano, 2024). However, no dedicated benchmark exists for Greek, and while some multilingual evaluations include Greek (Bandarkar et al., 2024), they lack financial-specific datasets, making it difficult to assess LLMs' performance on Greek financial area. At the same time, Greek LLM research has largely overlooked finance. While Meltemi (Voukoutis et al., 2024a) is the first Modern Greek LLM, it lacks financial domain adaptation. Existing Greek datasets focus on general NLP tasks (Clark et al., 2018; Lin et al., 2022; Zellers et al., 2019), failing to capture the domain-specific terminology and numerical reasoning essential for financial applications.

In this work, we introduce **Plutus-ben**, the first Greek financial evaluation benchmark and **Plutus-8B**, the pioneering Greek financial LLM. Plutus-ben addresses the aforementioned gap by defining five core financial NLP tasks in Greek, including numeric and textual named entity recognition (NER), question answering (QA), abstractive summarization, and topic classification, establishing a foundation for systematic and reproducible assessments of LLMs in Greek financial area. Notably, tasks such as financial numeric NER and financial QA are introduced in Greek for the first time. To support these tasks, we develop three high-quality Greek financial datasets, including GRFinNUM, GRFinNER, and GRFinQA, each carefully annotated by expert native Greek speakers with deep financial and linguistic expertise. Annotations follow strict, standardized guidelines to ensure consistency, accuracy, and high inter-annotator agreement. These newly developed datasets are curated from authoritative financial sources, including Greek financial reports and university exams, and are further supplemented by two existing financial resources, GRFNS-2023 and GRMultiFin. Beyond benchmarking, we introduce Plutus-8B, the first Greek financial LLM fine-tuned on domain-specific data, demonstrating the impact of targeted adaptation in bridging performance gaps for Greek financial tasks.

We evaluate 22 representative LLMs, spanning English-centric and Greek models across general and financial domains in various sizes, alongside our Plutus-8B, and uncover fundamental limitations in handling Greek financial tasks. Despite their success in high-resource languages, top models like GPT-4o underperform on Greek financial text, while smaller open-source models (e.g., LLaMA-3.2-1B, Qwen2.5-1.5B, Mistral-7B) fail entirely on key tasks such as NER. The challenge extends beyond language: financial text introduces specialized terminology, numerical reasoning, and ambiguous context. English-trained financial models fail to transfer effectively to Greek, and Greek-oriented models like Meltemi-7B, though excelling in general tasks, lack domain-specific competence. Scaling models offers limited benefit (e.g., Qwen2.5-72B does not outperform its 32B variant), highlighting the limits of scale alone. In contrast, our domain-adapted Plutus-8B achieves the highest mean performance, demonstrating the effectiveness of fine-tuning on Greek financial data. Nonetheless, significant challenges remain, particularly in summarization, where all models including Plutus-8B struggle with long-form financial documents.

Our main contributions are: 1) We introduce **Plutus-ben**, the first comprehensive Greek financial evaluation benchmark covering five key tasks and **Plutus-8B**, the first Greek financial LLM that achieves state-of-the-art (SOTA) performance on the Plutus-ben benchmark. 2) We develop four new high-quality Greek financial datasets, annotated by expert native speakers, and augment them with two existing resources to enhance task coverage. 3) We evaluate 22 LLMs on Plutus-ben, revealing persistent challenges in Greek financial NLP due to linguistic complexity, domain-specific terminology, and financial reasoning gaps. Our findings underscore the limitations of cross-lingual transfer and the need for domain-adapted Greek models. 4) We release Plutus-ben, Plutus-8B, and all associated datasets to drive reproducible research and promote multilingual inclusivity in financial NLP.

2 Plutus-ben: the First Greek Financial Evaluation Benchmark

In this section, we introduce Plutus-ben, the first Greek financial evaluation benchmark. As shown in Table 1, Plutus-ben encompasses a wide range of tasks, including *numeric NER*, *textual NER*, *question answering*, *abstractive summarization*, as well as *topic classification*, enabling a comprehensive evaluation of models. To support these tasks, we developed three new high-quality Greek financial datasets from scratch, including GRFinNUM, GRFinNER, and GRFinQA. Additionally, we use

two established resources, GRFNS-2023 and GR-MultiFin, with examples provided in Table 5 [4]. These datasets were rigorously annotated by expert native Greek speakers with deep financial and linguistic expertise, following standardized guidelines to ensure consistency and accuracy.

2.1 Task Definition and Dataset Curation

2.1.1 Numeric NER

Numerals are crucial in financial narratives, conveying essential quantitative information and actionable insights (Chen et al., 2018). Accurate numeral recognition is vital for interpreting nuanced financial data, especially when various categories exist simultaneously, i.e, monetary values, timestamps, and quantities (Chen et al., 2019b; Yang et al., 2022).

Task Definition: We introduced the first Greek financial numeric NER task, involving both number span identification and classification into fine-grained numeral types. Inspired by the English numeric NER framework FinNum (Chen et al., 2019a), we approach this task as a sequence labeling problem. Our task processes the input sentence $X = (x_1, x_2, \ldots, x_n)$ consisting of n tokens x_i to the output labels $Y = (y_1, y_2, \ldots, y_n)$ consisting of n labels y_i. The goal is to assign each token x_i a label y_i from the predefined set $\mathcal{C} =$ {MONETARY, PERCENTAGE, TEMPORAL, QUANTITY, OTHERS, O} , which includes specific numeric entity types and the "outside" label O. Among these categories, MONETARY includes financial amounts, such as prices, quotes, and changes, which are central to financial analysis. PERCENTAGE denotes ratios or relative changes, crucial for trend and growth tracking. TEMPORAL covers dates, times, and durations, integral to time-series analysis. QUANTITY captures measurable or countable values, such as inventory levels or investment positions. OTHERS encompasses numeric data not captured by the previous categories, leaving room for future exploration.

Data Source: To create our novel high-quality GRFinNUM dataset, we collected real-world, publicly available financial annual reports from Greek firms listed on the Athens Stock Exchange [5]. These reports include textual information and reviews provided by the firm's management and board of directors, offering rich, detailed financial data and narratives. We curated a dataset of 64 financial reports, each spanning 30 to 267 pages, with an average length of 105 pages or approximately 44,000 words per document. Due to their extensive length and inclusion of non-essential content, we meticulously filtered the text to extract sentences containing target entities. This rigorous selection process yielded a refined dataset of 500 sentences, ensuring relevance and quality for fine-grained numeral classification.

Expert Annotation: Rigorous annotation guideline (Appendix H) was developed for GRFinNUM, comprising both general rules for the overall task and specific rules tailored to each numeral category. These guidelines were iteratively refined through multiple rounds of pre-annotation and collaborative discussions, focusing on resolving ambiguous cases to ensure high consistency and accuracy across the dataset. To minimize annotator variability, only numbers, decimal points (.), and the percent sign (%) were included in annotated spans. To construct novel high-quality dataset, we enlisted three highly educated Greek native speakers with expertise in economics, business, and informatics from leading academic institutions (Appendix K). The annotation process was conducted using Label Studio platform (Tkachenko et al., 2020-2025) (Appendix L), ensuring a streamlined and reproducible workflow.

Quality Validation: To gauge the quality and reliability of our GRFinNUM annotation process, we utilized three key inter-annotator agreement metrics: F1 score (Goutte and Gaussier, 2005), Cohen's Kappa (Wongpakaran et al., 2013), and Krippendorff's Alpha (Hayes and Krippendorff, 2007) (Appendix N). F1 Score evaluated annotator consistency in span identification and classification. Cohen's Kappa adjusted for random agreement, while Krippendorff's Alpha addressed category distribution imbalances. The results demonstrated excellent inter-annotator agreement for the GRFinNUM dataset, with an F1 score of 0.988, a Cohen's Kappa of 0.979, and a Krippendorff's Alpha of 0.978 (Table 2). These high scores confirm the robustness and quality of our GRFinNUM dataset.

2.1.2 Textual NER

Identifying core financial entities, such as companies, is crucial for extracting meaningful insights from financial activities in the Greek financial domain. Unlike numeric NER, which focuses on recognizing numerical values, textual NER in

[4] More details in Appendix G.

[5] https://www.athexgroup.gr/el/web/guest/financial-statements-in-pdf-format

Table 1: Overview of the Plutus-ben benchmark. For each task, both raw data volume and processed size are listed, along with dataset source, split sizes for train/validation/test, evaluation metrics, licenses, and tested capabilities.

Task	Dataset	Raw	Processed	Source	Train	Valid	Test	Metrics	License	Tested Capabilities
Numeric NER	GRFinNUM	64	500	Annual Reports[1]	320	80	100	Entity F1	Public	Numeric information extraction
Textual NER	GRFinNER	64	500	Annual Reports[2]	320	80	100	Entity F1	Public	Textual information extraction
Question Answering	GRFinQA	540	540	Exam Questions	267	48	225	Acc	Public	Language comprehension and reasoning
Abstractive Summarization (Zavitsanos et al., 2023)	GRFNS-2023 (Zavitsanos et al., 2023)	262	262	Annual Reports	169	43	50	Rouge-1	CC-BY-4.0	Long-form financial document comprehension
Topic Classification (Jørgensen et al., 2023)	GRMultiFin (Jørgensen et al., 2023)	268	268	Article Headlines	171	43	54	Acc	CC BY-NC 4.0	Language comprehension and topical content categorizing

[1] https://www.athexgroup.gr/web/guest/company-fin.-statements/
[2] https://www.athexgroup.gr/web/guest/company-fin.-statements/

Table 2: Inter-annotator agreement metrics for human expert annotations on GRFinNUM and GRFinNER datasets.

Dataset	F1-score	Cohen's Kappa	Krippendorff's alpha
GRFinNUM	0.988	0.979	0.978
GRFinNER	0.974	0.993	0.948

Greek presents unique challenges due to the language's distinct expression patterns. For instance, long-form names with attribution, such as "George Demetriou of Konstantinos", should be treated as a single entity span.

Task Definition: To test LLMs' understanding of Greek financial entities, we introduce the first Greek financial textual NER task. Inspired by FinNER-ORD (Shah et al., 2023) and Farmakiotou et al. (Farmakiotou et al., 2000), our task involves span identification and classification of company-related information into three key entity types: Person, Location, and Organization. Our task processes the input sentence $X = (x_1, x_2, \ldots, x_n)$ consisting of n tokens x_i to the output labels $Y = (y_1, y_2, \ldots, y_n)$ consisting of n labels y_i. The goal is to assign each token x_i a label y_i from the predefined set $\mathcal{C} = \{\text{PERSON}, \text{LOCATION}, \text{ORGANIZATION}, O\}$, which includes specific textual entity types and the "outside" label O.

Data Source: We constructed the GRFinNER dataset using the same set of financial annual reports from Greek firms as in GRFinNUM. A total of 64 reports were collected. Similar sentences filtering is utilized for a different final dataset of 500 sentences with high relevance and quality for company-related entity classification.

Expert Annotation: Rigorous annotation guideline (Appendix I) was also iteratively developed for GRFinNER through multiple rounds of pre-annotation and collaborative discussions, consisting of general rules for the entire task, specific rules for each entity category, and distinct rules for handling ambiguous situations. The same three highly educated Greek native speakers (Appendix K) completed the annotation process. The entire annotation workflow was carried out using Label Studio platform (Appendix L).

Quality Validation: The inter-annotator agreement was meticulously assessed using the same rigorous framework: F1 score (Goutte and Gaussier, 2005), Cohen's Kappa (Wongpakaran et al., 2013), and Krippendorff's Alpha (Hayes and Krippendorff, 2007) (Appendix N). The GRFinNER task exhibited exceptional inter-annotator reliability, achieving an F1 score of 0.974, Cohen's Kappa of 0.993, and Krippendorff's Alpha of 0.948 (Table 2), ensuring the dataset's quality for application.

2.1.3 Question Answering

Effective financial decision-making and question answering require LLMs to comprehend and reason within financial contexts. The nuances of Greek financial terminology, combined with the complex morphology of the Greek language, pose unique challenges that demand rigorous assessment.

Task Definition: To evaluate LLMs' comprehension and reasoning capabilities in Greek financial contexts, we introduce the first Greek financial question-answering task. This task requires models to infer the correct answer using provided text under a multiple-choice format, testing their ability to process financial terminology, apply reasoning, and understand contextual nuances in Greek. Each question, along with its answer choices, is given as input, with the correct answer designated as the output. Our task processes the input question $Q = (q_1, q_2, \ldots, q_n)$ consisting of n tokens q_i and the possible choices $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ which is the set of k possible choices c_i. The task aims to map the question Q and choices $\mathcal{C}$ to the correct answer A, selected from $\mathcal{C}$.

Data Source: We propose the novel GRFinQA dataset which is the first in the Greek financial domain. It is comprised of 540 multiple-choice financial exam or revision questions sourced from Greek university courses and publicly available Greek finance, business and economics textbooks.

We collected the PDF files, and extracted the text that each question was grouped with it's appropriate choices and the correct choice.

Quality Validation: To ensure the quality of the dataset, we first identified three distinct types of questions present in the QA dataset: (1) right and wrong questions, which require a binary judgment on whether a statement is correct or incorrect; (2) fill-in-the-gap questions, where a missing word or phrase must be completed based on contextual understanding; and (3) generic multiple-choice questions, which present several answer options, with only one being correct. From this dataset, we selected a representative sample that included several questions from each category. The domain experts manually reviewed these questions to confirm that the designated correct answer was factually accurate. Following that, we used GPT-4o to process the questions, prompting it to read the text and explain its reasoning for selecting an answer. This helped us verify both the factual accuracy of the dataset's answers and the difficulty of questions.

2.1.4 Abstractive Summarization

The task of abstractive summarization originates from the Financial Narrative Summarization Shared Task (FNS 2023), which focuses on summarizing annual reports from the UK, Greece, and Spain (Zavitsanos et al., 2023). This task aims to test LLMs' abilities in understanding and reorganizing the given context. The challenge lies in condensing essential information while preserving factual accuracy and coherence. The structural and linguistic complexities of Greek financial texts further heighten this difficulty, requiring models to generate fluent, paraphrased summaries that remain faithful to the original content.

Task Definition: To evaluate LLMs' abilities of understanding the Greek financial contexts, we adopt the abstractive summarization task from FNS 2023 (Zavitsanos et al., 2023). This task involves generating concise summaries of Greek financial annual reports, emphasizing both informativeness and readability while preserving key details. The task processes the input document $D = (d_1, d_2, \ldots, d_n)$ consisting of n tokens d_i to the abstractive summary $S = (s_1, s_2, \ldots, s_m)$ consisting of m tokens s_i. The goal is to map the document D to a concise summary S that conveys the essential information in natural language, which is paraphrased or restructured rather than directly copied from D.

Data Source: The FNS 2023 shared task (Zavitsanos et al., 2023) comprises UK, Greek, and Spanish financial annual reports. The dataset includes narrative sections from finanical annual reports, each paired with both a short and long gold summary. For GRFNS-2023, we focus solely on the Greek portion, using the short gold summary as our target. As the original authors did not release a test set, we repurposed their validation set as our test set and split the training data to create our training and validation sets.

2.1.5 Topic Classification

The topic classification task is derived from MultiFin (Jørgensen et al., 2023), and it focuses on categorizing financial news headlines into predefined financial topics. This task is particularly challenging due to the brevity and ambiguity characteristic of financial news headlines. Furthermore, financial categories often exhibit thematic and lexical overlaps, demanding that models discern the appropriate category from limited context and shared terminology.

Task Definition: To improve LLMs' comprehension of Greek financial topics, we incorporated the Greek financial topic classification task adapted from MultiFin (Jørgensen et al., 2023). This task requires assigning financial article headlines to one of six predefined thematic categories. The objective is to evaluate models' proficiency in distinguishing between overlapping topics and extracting significant insights from brief and ambiguous texts. Our task processes the input document $D = (d_1, d_2, \ldots, d_n)$ consisting of n tokens d_i and the possible topics $\mathcal{C} = \{\text{Topic}_1, \text{Topic}_2, \ldots, \text{Topic}_k\}$ which is the set of k possible topics. The goal is to map the input document D to the correct topic T from $\mathcal{C}$, based on the content of D.

Data Source: The dataset utilized for this task is the MultiFin dataset (Jørgensen et al., 2023). It comprises 10,048 financial article headlines in 15 languages, each reflecting diverse language families and writing systems. These headlines are categorized into one of six classes: Business & Management, Tax & Accounting, Finance, Technology, Government & Controls, and Industry. For our specific analysis, we extracted the Greek subset to create the GRMultiFin dataset.

2.2 Evaluation

To optimize task-specific performance, facilitate effective benchmarking, and support instruction

fine-tuning for the Greek financial LLM, we converted our raw datasets into structured instruction datasets[6]. Task-specific prompts were thoughtfully crafted by Greek domain experts, as shown in Table 6[7]. We partitioned our dataset into training, validation, and test subsets, as detailed in Table 1. To comprehensively assess model performance, we conducted both automated metrics and human evaluations.

Automatic Evaluation We adopt the same metrics following previous studies in financial NLP tasks (Zhang et al., 2024; Xie et al., 2024a). The Entity F1 score (Derczynski, 2016) is applied to numeric and textual NER tasks due to its balance of precision and recall, crucial for accurate entity identification. Accuracy (Acc) (Makridakis, 1993) is used for QA and topic classification tasks as it straightforwardly measures the correctness of predictions. Rouge-1 (Lin, 2004) is employed for abstractive and extractive summarization tasks to assess the overlap in content between gold-standard and generated summaries focusing on unigram comparison.

Human Evaluation Beyond automated metrics, we implement a human evaluation to rigorously assess the quality of outputs from LLMs. This evaluation specifically concentrates on abstractive summarization task. We selected four representative models, including GPT-4, FinLLaMA-8B, Meltemi-7B, and Plutus-8B. Expert native Greek speakers with deep financial and linguistic expertise[8] compare the model-generated summaries against gold standard summaries following a rigorous, standardized annotation guideline[9] using Label Studio platform[10]. The evaluation focuses on three critical dimensions: **(1) Language Appropriate Fluency (Fluency):** This dimension assesses the readability and naturalness of the summaries, emphasizing grammatical correctness, lexical accuracy, absence of repetition, and the use of domain-specific terminology, all within the context of Greek's linguistic intricacies. **(2) Coherence:** We examine the logical progression and structural consistency of the summaries, vital for maintaining integrity in financial narratives. **(3) Factuality:** This dimension verifies the factual accuracy of summaries against the original financial content, ensuring reliability and trustworthiness.

2.3 Model Evaluation

We conduct a comprehensive evaluation of 22 prominent LLMs encompassing[11]: (1) **4 proprietary models** from OpenAI (Brown et al., 2020; OpenAI et al., 2024; Hurst et al., 2024; Achiam et al., 2023), (2) **13 open-source general-purpose models** with both large size and small size models from Mistral, Qwen, Gemma, LLaMA, and DeepSeek (Mistral AI team, 2023; Dubey et al., 2024; Yang et al., 2025a; Team et al., 2024; Liu et al., 2024), (3) **2 financial domain-specific models** including FinMA and OpenFinLLM (Xie et al., 2023b, 2024c), and (4) **2 Greek general models** including Meltemi-7B (Voukoutis et al., 2024b) and Llama-Krikri-8B-Base[12].

For evaluation integrity, we develop our own benchmark suites based on LM Evaluation Harness (Gao et al., 2024). Models such as GPT and DeepSeek, are interfaced via their own APIs. In-house evaluation of open-source models is conducted using a cluster of four A100 GPUs, each equipped with 80GB memory. We standardize the maximum generation token length to 8192 tokens for abstractive summarization and 1024 tokens for other tasks.

3 Plutus-8B: the First Greek Financial LLM

To investigate the impact of fine-tuning on Greek financial data on enhancing model performance across various tasks, and to determine its effectiveness in addressing the challenges posed by low-resource language conditions and domain-specific complexities, we developed Plutus-instruction, the first instruction dataset tailored to the Greek financial domain. As shown in Table 1, we adopted GRFinNUM, GRFinNER, GRFNS-2023, and GR-MultiFin. Specifically, the GRFinQA dataset is withheld to evaluate the generalization performance of the trained model.

Based on the instruction dataset, we selected Llama-Krikri-8B-Instruct for further instruction-tuning[13], as this model performs best on the benchmark compared to other models of similar size. This is due to its training on extensive Greek texts,

[6] More details in Appendix B.
[7] More details in Appendix G
[8] More details in Appendix K
[9] More details in Appendix J
[10] More details in Appendix L
[11] More details in Appendix C
[12] https://huggingface.co/ilsp/Llama-Krikri-8B-Base
[13] For training details, please see Appendix D.

Table 3: LLM performance on the Plutus-ben benchmark, evaluated across multiple Greek financial NLP tasks. Bold values denote the highest scores, while underlined values indicate the second-highest scores in each column.

Model	GRFinNUM	GRFinNER	GRFinQA	GRFNS-2023	GRMultiFin	Mean
	Entity F1	Entity F1	Acc	Rouge-1	Acc	
Open-source Small Models						
LLaMA-3.2-1B	0.00	0.00	0.29	0.14	0.39	0.16
LLaMA-3-8b	0.00	0.13	0.33	0.07	0.70	0.25
LLaMA-3.1-8b	0.10	0.21	0.40	0.20	0.54	0.29
Qwen2.5-1.5B	0.00	0.00	0.36	0.02	0.31	0.14
Qwen2.5-7B	0.00	0.13	0.43	0.07	0.54	0.23
Gemma-2-2B	0.00	0.16	0.22	0.03	0.41	0.16
Gemma-2-9B	0.02	0.05	0.31	0.06	0.61	0.21
Mistral-7B	0.00	0.00	0.30	0.14	0.39	0.17
Open-source Large Models						
Deepseek-V3	0.07	0.00	0.50	**0.38**	0.61	0.31
LLaMA-3-70B	0.05	0.45	0.60	0.08	0.61	0.36
Qwen2.5-32B	0.37	0.55	0.60	0.10	0.70	0.47
Qwen2.5-72B	0.32	0.39	0.74	0.04	**0.72**	0.44
Gemma-2-27B	0.18	0.18	0.25	0.09	0.61	0.26
Proprietary Models						
GPT-3.5-Turbo	0.14	0.30	0.51	0.31	0.50	0.35
GPT-4o-Mini	0.25	0.30	0.12	0.36	0.59	0.32
GPT-4o	0.09	0.31	**0.78**	0.26	0.59	0.41
GPT-4	0.28	**0.60**	0.71	**0.38**	0.63	0.52
English Financial Models						
Finma-7B	0.00	0.00	0.25	0.11	0.35	0.14
FinLLaMA-8B	0.00	0.00	0.28	0.03	0.38	0.14
Greek General Models						
Meltemi-7B	0.12	0.50	0.48	0.19	0.43	0.34
Llama-Krikri-8B	0.19	0.45	0.57	0.22	0.39	0.36
Greek Financial Models						
Plutus-8B	**0.70**	0.57	0.64	0.34	**0.72**	**0.60**

as well as its inclusion of code and mathematical data to enhance its mathematical reasoning abilities. We further evaluate our Plutus-8B model in Plutus-ben and compare it with all evaluated models[14].

4 Results

In this section, we present the results of evaluated models on the Plutus-ben benchmark, addressing: (i) how current models handle Greek financial tasks under low-resource, linguistically complex, and domain-specific conditions; and (ii) whether fine-tuning on Greek financial data mitigates those challenges.

4.1 Main Results

Table 3[15] and Figure 1 summarize the performance of various LLMs on our Greek-oriented financial benchmark, Plutus-ben. Overall, results confirm that both linguistic and domain-specific limitations significantly hinder LLM performance.

Most models struggle with Greek's rich morphology and inflectional structure, particularly in NER. Smaller open-source models (e.g., LLaMA-3.2-1B, Qwen2.5-1.5B, Mistral-7B) perform poorly across all tasks, often scoring near zero on GRFinNER and GRFinNUM. Even larger models like LLaMA-3-70B and Gemma-2-27B offer limited improvement, especially in numeric comprehension. Proprietary models such as GPT-4 achieve higher mean scores (up to 0.52) but still underperform compared to their performance on English benchmarks (Xie et al., 2023a, 2024a).

Financial text introduces additional challenges, including specialized terminology, complex numeric formats, and context-dependent semantics. English-trained financial models (e.g., Finma-7B, FinLLaMA-8B) fail to transfer effectively to Greek, scoring only 0.14 on average and failing on NER tasks. Even GPT-4o, while better on GRFinNER (0.31), performs poorly on GRFinNUM (0.09), highlighting limitations in adapting to Greek-specific financial numeracy. Greek-centric models (e.g., Meltemi-7B, Llama-Krikri-8B) show better linguistic adaptation, outperforming their backbone models. For instance, Meltemi-7B achieves a mean score of 0.34 (vs. 0.17 for Mistral-7B), and Llama-Krikri-8B reaches 0.36 (vs. 0.29 for LLaMA-3.1-8B). However, both underperform in GRFinNUM (0.12 and 0.19, respectively), despite strong GRFinNER scores, indicating that linguistic adaptation alone is insufficient for financial reasoning.

While larger models generally perform better, gains from scaling plateau quickly. Qwen2.5-32B outperforms Qwen2.5-72B on multiple tasks despite its smaller size, and LLaMA-3-70B struggles with numeric tasks (GRFinNUM = 0.05). GPT-4o (mean = 0.41) offers only marginal improvements over GPT-3.5-Turbo (0.35). These results suggest that scale alone does not ensure better performance without financial and linguistic adaptation.

Finally, **fine-tuning on a dedicated Greek financial corpus significantly enhances model performance but also reveals explicit bottlenecks that require further improvements.** Our model, Plutus-8B, fine-tuned exclusively on Greek financial data, achieves the highest mean score (0.60), surpassing all baselines. It particularly excels in GRFinNUM (0.70), demonstrating strong numeric reasoning capabilities. Plutus-8B also performs well on GRFinNER and GRMultiFin, highlighting the benefits of targeted fine-tuning. On GRFinQA (held out during fine-tuning), Plutus-8B achieves 0.64, outperforming Meltemi-7B (0.48) and Llama-Krikri-8B (0.57), indicating strong generalization. However, performance on GRFNS-2023 remains modest due to the difficulty of modeling long-range dependencies in financial documents. These find-

[14] For demo, please see Appendix F.

[15] Ranked results are visualized on our leaderboard. For more details, refer to Appendix E.

Table 4: Human evaluation results assessing fluency, coherence, and factuality of representative LLMs, evaluated on the GRFNS-2023 dataset within the Plutus-ben benchmark.

Domain	Model	Fluency	Coherency	Factuality
English general model	**GPT-4**	**4.97**	**4.33**	**3.06**
English financial model	**FinLLaMA-8B**	2.09	1.48	1.54
Greek general model	**Meltemi-7B**	3.99	1.49	1.60
Greek financial model	**Plutus-8B**	3.90	3.51	2.93

ings highlight the critical role of domain-specific pretraining, especially for tasks requiring numeric reasoning, while also indicating areas for further improvement.

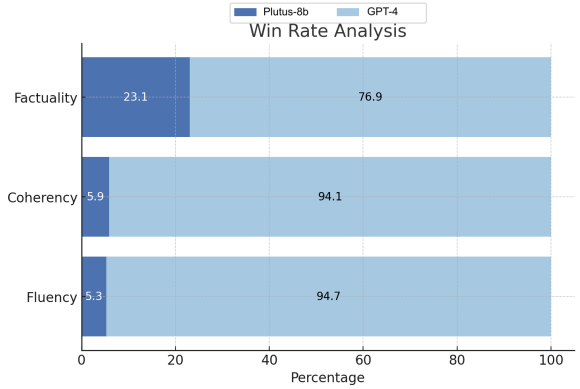

Figure 2: Comparison of model win rates in fluency, coherence, and factuality between Plutus-8B and GPT-4, evaluated on the GRFNS-2023 dataset within the Plutus-ben benchmark.

4.2 Human Evaluation

To complement automatic metrics, we conducted a human evaluation of selected models on Greek financial tasks (Appendix J). Results in Table 4 show that while GPT-4 leads in fluency, our domain-specific Plutus-8B outperforms similarly sized models in coherency (3.51) and factuality (2.93), underscoring the benefits of domain-aware fine-tuning. These findings highlight the need to strengthen both linguistic and domain-specific capabilities when adapting general-purpose LLMs to specialized, low-resource settings. The strong performance of Plutus-8B, especially compared to models like FinLLaMA-8B which is trained on English financial data, demonstrates the limitations of cross-lingual transfer and the importance of in-language financial supervision. Notably, Meltemi-7B, trained for general Greek tasks, ranks second in fluency (3.99) but lags in coherency (1.49) and factuality (1.60), suggesting that fluency benefits from Greek-specific training, whereas factual consistency requires domain-specific grounding.

We further compare Plutus-8B and GPT-4 using a pairwise win-rate evaluation on long-context processing (Figure 2), focusing on GRFNS-2023, a long-form dataset derived from financial reports averaging 60 pages (31.5k words). Due to its larger size and more advanced architecture, GPT-4 outperforms Plutus-8B across most metrics. However, Plutus-8B achieves a 23.1% win rate in factuality and closes the performance gap with a factuality score of 2.93 vs. GPT-4's 3.06. These results suggest that Plutus-8B benefits from instruction tuning with financial disambiguation patterns and Greek-specific numerical structures, enhancing its reliability in financial summarization. Although it struggles with long-context inputs compared to GPT-4, Plutus-8B demonstrates that targeted fine-tuning significantly improves domain-specific performance in low-resource languages.

Overall, Plutus-8B's domain-aware fine-tuning equips it to better navigate financial contexts—narrowing the gap with larger, general-purpose models like GPT-4. This highlights the critical role of combining linguistic and domain-specific training to enhance LLM performance in non-English, domain-focused tasks.

5 Conclusion

In this study, we introduced **Plutus-ben**, the first Greek financial evaluation benchmark, and **Plutus-8B**, the first Greek financial LLM. Addressing a critical resource gap, **Plutus-ben** includes five key NLP tasks, numeric and textual NER, QA, abstractive summarization, and topic classification. To support these tasks, we develop and release three novel datasets, GRFinNUM, GRFinNER, and GRFinQA, carefully annotated by expert native Greek speakers, establishing the first high-quality resources for Greek financial NLP. Our evaluation of 22 models, including a detailed human study, demonstrates that current LLMs face significant challenges due to linguistic complexity, domain-specific requirements, and cross-lingual transfer. Plutus-8B achieves SOTA results across most tasks and demonstrates strong factuality in long-context evaluation, underscoring the importance of domain-aware, language-specific adaptation. By releasing Plutus-ben, Plutus-8B, and associated datasets, we aim to advance research in Greek financial NLP, promote multilingual inclusivity, and encourage further innovation.

Limitations

While this study offers valuable insights, it is important to acknowledge the following limitations: (1) **Parameter Restriction**: Plutus-8B is currently limited to a size of 8B parameters, and future work should explore both smaller models for efficiency and larger models for enhanced performance. (2) **Limited Evaluation Benchmark**: The datasets available in Plutus-ben are limited in size, which may impede the model's ability to understand financial contexts comprehensively and generalize effectively across diverse scenarios. Plutus-8B exhibits varied performance on Plutus-ben, particularly struggling with summarizing long-form financial documents. (3) **Limited Application Scope**: The design and instructional approach of Plutus-8B may constrain its utility across different bilingual contexts. This specific tailoring could limit its generalizability to other linguistic or cultural 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 Plutus-8B primarily for scholarly research, mindful of these ethical aspects.

Ethical Statement

The authors take full responsibility for the development and dissemination of Plutus-ben and Plutus-8B, ensuring that all raw data used are publicly available, devoid of personal information, and conform to established ethical guidelines. The data are shared under the MIT license, requiring users to adhere to its terms. This manuscript, including large language models, source codes, and datasets, is intended for academic and educational purposes only and is not a substitute for professional advice. While efforts have been made to ensure its accuracy, the authors and their institutions disclaim liability for any outcomes arising from its use. Users agree to take responsibility for ethical and lawful use and to indemnify the authors and their affiliates against any claims or damages resulting from reliance on this Material.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 749–775. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143. IEEE.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019a. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, pages 19–27.

Chung-Chi Chen, Hen-Hsen Huang, Chia-Wen Tsai, and Hsin-Hsi Chen. 2019b. Crowdpt: Summarizing crowd opinions as professional analyst. In *The World Wide Web Conference*, WWW '19, page 3498–3502, New York, NY, USA. Association for Computing Machinery.

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. *CoRR*, abs/2310.15205.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Leon Derczynski. 2016. Complementarity, F-score, and NLP evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266, Portorož, Slovenia. European Language Resources Association (ELRA).

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Angeliki Efthymiou and Nikos Koutsoukos. Inflectional and semantic properties of verbal pairs in modern greek.

Sharman Esarey. 2020. Lessons from financial assistance to greece–technical appendix.

Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D Spyropoulos, and Panagiotis Stamatopoulos. 2000. Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, pages 75–78.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Masanori Hirano. 2024. Construction of a japanese financial benchmark for large language models. *Preprint*, arXiv:2403.15062.

David Holton, Peter Mackridge, Irene Philippaki-Warburton, and Vassilios Spyropoulos. 2012. *Greek: A comprehensive grammar of the modern language*. Routledge.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Preprint*, arXiv:2311.11944.

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*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Koduvayur Subbalakshmi, Guojun Xiong, Jimin Huang, Lingfei Qian, Xueqing Peng, Qianqian Xie, and Jordan W. Suchow. 2024. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. *Preprint*, arXiv:2412.18174.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. CFGPT: chinese financial assistant with large language model. *CoRR*, abs/2309.10654.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Spyros Makridakis. 1993. Accuracy measures: theoretical and practical concerns. *International journal of forecasting*, 9(4):527–529.

Mistral AI team. 2023. Mistral 7b in short.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, Yunhe Wang, and Dacheng Tao. 2024. Cfinbench: A comprehensive chinese financial benchmark for large language models. *Preprint*, arXiv:2407.02301.

OpenAI, Josh Achiam, and Steven Adler etal. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Katerina Papantoniou and Yannis Tzitzikas. 2024. Nlp for the greek language: A longer survey. *Preprint*, arXiv:2408.10962.

Agam Shah and Sudheer Chava. 2023. Zero is not hero yet: Benchmarking zero-shot performance of llms for financial tasks. *Preprint*, arXiv:2305.16633.

Agam Shah, Abhinav Gullapalli, Ruchit Vithani, Michael Galarnyk, and Sudheer Chava. 2023. Finerord: Financial named entity recognition open research dataset. *arXiv preprint arXiv:2302.11157*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024a. Meltemi: The first open large language model for greek. *CoRR*, abs/2407.20743.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024b. Meltemi: The first open large language model for greek. *Preprint*, arXiv:2407.20743.

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

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13:1–7.

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.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024a. Finben: A holistic financial benchmark for large language models. *Preprint*, arXiv:2402.12659.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023a. PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023b. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *Preprint*, arXiv:2306.05443.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, Shunian Chen, Yifei Zhang, Lihang Shen, Daniel Kim, Zhiwei Liu, Zheheng Luo, Yangyang Yu, Yupeng Cao, Zhiyang Deng, and 20 others. 2024b. Open-finllms: Open multimodal large language models for financial applications. *Preprint*, arXiv:2408.11878.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, and 1 others. 2024c. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025a. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

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

Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. *Preprint*, arXiv:2201.01770.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. *CoRR*, abs/2309.13064.

Yuzhe Yang, Yifei Zhang, Yan Hu, Yilin Guo, Ruoli Gan, Yueru He, Mingcong Lei, Xiao Zhang, Haining Wang, Qianqian Xie, Jimin Huang, Honghai Yu, and Benyou Wang. 2025b. Ucfe: A user-centric financial expertise benchmark for large language models. *Preprint*, arXiv:2410.14059.

Elias Zavitsanos, Aris Kosmopoulos, George Giannakopoulos, Marina Litvak, Blanca Carbajo-Coronado, Antonio Moreno-Sandoval, and Mo El-Haj. 2023. The financial narrative summarisation shared task (fns 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2890–2896.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

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 2024, Barcelona, Spain, August 25-29, 2024*, pages 6236–6246. ACM.

A Related Work

A.1 Financial and Greek LLMs

In recent years, an increasing number of LLMs have been tailored to financial applications. Most existing work is English-centric, such as FinLLaMA (Xie et al., 2024b), BloombergGPT (Wu et al., 2023), PIXIU (Xie et al., 2023a), InvestLM (Yang et al., 2023b), and FinGPT (Yang et al., 2023a), leveraging domain-specific financial corpora for tasks. In parallel, recent research in Chinese (DISC-FinLLM (Chen et al., 2023) and CFGPT (Li et al., 2023) and bilingual financial LLMs (FinMA-ES (Zhang et al., 2024) for Spanish and English) extend these efforts by covering related non-English and bilingual finance tasks. Despite these notable advancements, there is a conspicuous absence of specialized Greek financial LLMs. Existing Greek open-source LLMs, such as Meltemi (Voukoutis et al., 2024a) and Llama-Krikri[16], do not include finance-oriented training data, which highlights the critical need for developing a financial model specifically tailored to the Greek context.

A.2 Financial Benchmarks

Numerous financial benchmarks have been developed for evaluating LLMs' capabilities in financial domain. Though FinBen (Xie et al., 2024a), INVESTORBENCH (Li et al., 2024), PIXIU (Xie et al., 2023a), UCFE (Yang et al., 2025b), FinanceBench (Islam et al., 2023), and FinGPT (Wang et al., 2023) provide wide-ranging evaluations, covering comprehensive financial tasks and experiment settings, they are predominantly in English. Efforts to move beyond English have resulted in benchmarks covering Spanish (Zhang et al., 2024), Chinese (Nie et al., 2024), and Japanese (Hirano, 2024), underscoring the the value of linguistic and cultural diversity in financial tasks. While Greek mentioned in a few multilingual benchmarks like the Belebele benchmark (Bandarkar et al., 2024), there is no dedicated Greek financial benchmark, making it difficult to rigorously assess LLMs in Greek finance-specific contexts.

B Instruction Data Conversion

To optimize task-specific performance, facilitate effective benchmarking, and support instruction fine-tuning for the Greek financial LLM, we converted our raw datasets into structured instruction datasets. Task-specific prompts were thoughtfully crafted by Greek domain experts, as shown in Table 6[17]. Each prompt adheres to the standardized template as outlined below:

Task Instruction

```
{Task Specific Instruction} Text: {Input} Answer: {Output}
```

In this template, task specific instruction refers to the unique prompt designed for each task. The "`Input`" denotes the input financial data from each dataset, such as a Greek annual report, while "`Output`" represents the corresponding output for the input text, such as a summary of the Greek annual report.

[16] https://huggingface.co/ilsp/Llama-Krikri-8B-Base

[17] More details in Appendix G

C Model Evaluation

We conduct a comprehensive evaluation of 22 prominent LLMs encompassing:

- **Proprietary Models**: close source APIs, including GPT-3.5-Turbo (Brown et al., 2020), GPT-4o-Mini (OpenAI et al., 2024), GPT-4o (Hurst et al., 2024), and GPT-4 (Achiam et al., 2023).
- **Open-source General Small Models**: publicly available models with less than 10B parameters, including Mistral-7B (Mistral AI team, 2023), LLaMA-3.2-1B (Dubey et al., 2024), LLaMA-3-8B (Dubey et al., 2024), LLaMA-3.1-8B (Dubey et al., 2024), Qwen2.5-1.5B (Yang et al., 2025a), Qwen2.5-7B (Yang et al., 2025a), Gemma-2-2B (Team et al., 2024), and Gemma-2-9B (Team et al., 2024).
- **Open-source General Large Models**: publicly available models with more than 20B parameters, including Deepseek-V3 (Liu et al., 2024), LLaMA-3-70B (Dubey et al., 2024), Qwen2.5-32B (Yang et al., 2025a), and Qwen
  2.5-72B (Yang et al., 2025a), and Gemma-2-27B (Team et al., 2024).
- **English Financial Models**: publicly available models continual trained with English financial corpus, including Finma-7B (Xie et al., 2023b) and FinLLaMA-8B (Xie et al., 2024c).
- **Greek General Models**: publicly available models continual trained with Greek general corpus, including Meltemi-7B (Voukoutis et al., 2024b) and Llama-Krikri-8B[18].

Notably, LLaMA-3-8B, Mistral-7B, and LLaMA-3.1-8b serve as the core foundational models for FinLLaMA-8B, Meltemi-7B, and Llama-Krikri-8B, respectively.

D Training Details

To efficiently adapt the model parameters, we employ Low-Rank Adaptation (LoRA) (Dettmers et al., 2023) with a rank of $r = 16$, a scaling factor of $\alpha = 32$, and no dropout. We applied `int4` quantization to reduce memory overhead while preserving model expressiveness. Fine-tuning is conducted with a block size of 4,096 tokens, while allowing sequences to extend to 42k tokens to accommodate the complex structure and extensive length of financial and legal documents. To ensure better optimization, we leveraged the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $5e-4$ and a cosine learning rate schedule over 3 epochs. Additionally, we use gradient accumulation with a step size of 4 to mitigate the constraints of batch size 1, leveraging mixed-precision training with `bf16` for improved numerical stability.

[18] https://huggingface.co/ilsp/Llama-Krikri-8B-Base

E Open Greek Financial LLM Leaderboard

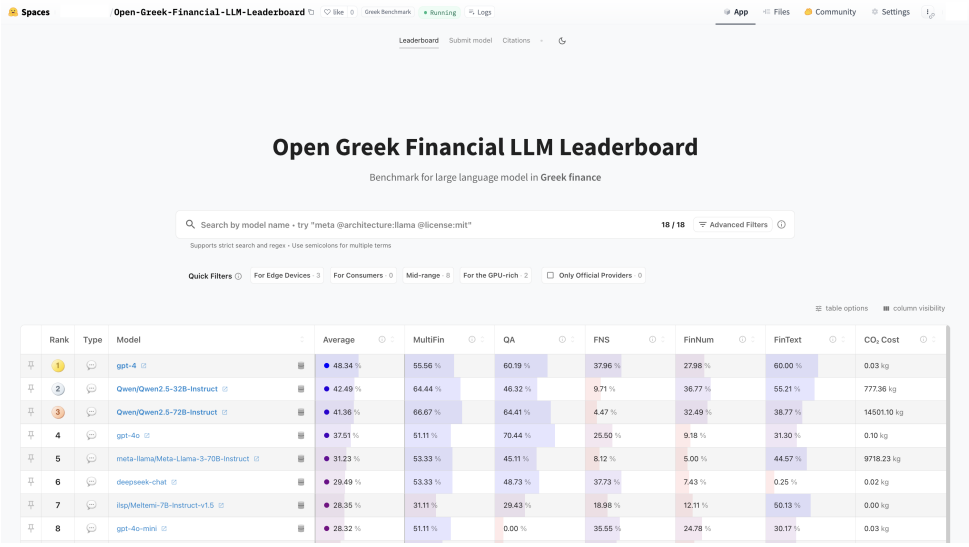

Figure 3: The Plutus-ben interface.

F Plutus-8B-instruct

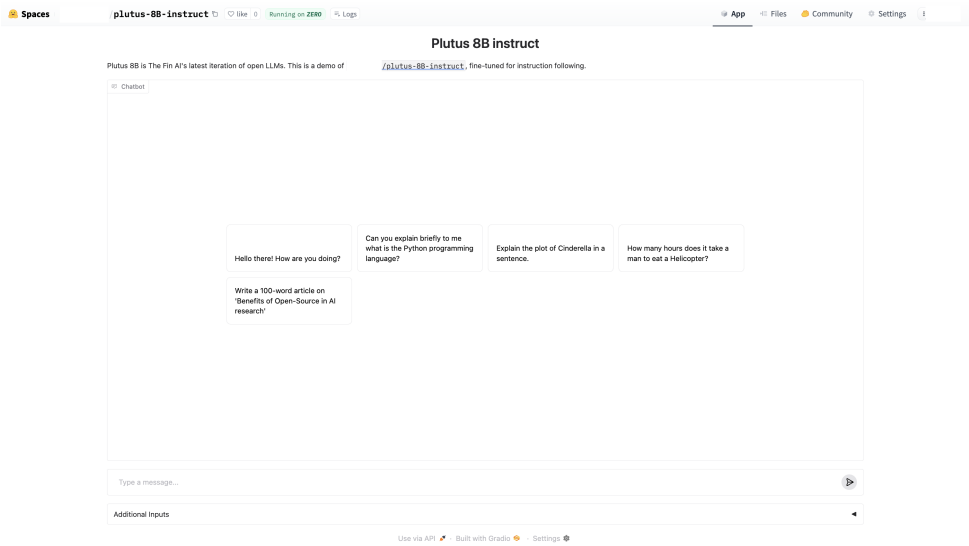

Figure 4: The Demo of Plutus-8B-instruct.

G Dataset Curation and Conversion

Table 5: Datasets included in the Plutus-ben benchmark, presented in both the original Greek and their English translations.

Dataset	Version	Input	Output
GRFinNUM	Greek Original	Σε επίπεδο ομίλου τα κέρδη ανα μετοχή είναι αυξημένα + 10,11% λόγω της επίδρασης λειτουργίας της Cosmokid AE που ξεκίνησε ουσιαστικά το Β εξάμηνο του 2008.	10,11%, ΠΟΣΟΣΤΑ 2008, ΧΡΟΝΙΚΑ
	English Translation	At the group level, earnings per share are increased by +10.11% due to the impact of Cosmokid AE's operations, which started in the second half of 2008.	10.11%, PERCENTAGE 2008, TEMPORAL
GRFinNER	Greek Original	Στις 08.11.2019, η ΟΠΑΠ INVESTMENT LTD ήρθε σε συμφωνία με την Εταιρεία για την πώληση του συνόλου των μετοχών που κατέχει στην ΙΠΠΟΔΡΟΜΙΕΣ Α.Ε., έναντι συνολικού τιμήματος € 10.411.	ΟΠΑΠ INVESTMENT LTD, ΟΡΓΑΝΙΣΜΟΣ ΙΠΠΟΔΡΟΜΙΕΣ Α.Ε., ΟΡΓΑΝΙΣΜΟΣ
	English Translation	On 08.11.2019, OPAP INVESTMENT LTD reached an agreement with the Company for the sale of all the shares it holds in HIPPODROMIES S.A., for €10,411.	OPAP INVESTMENT LTD, ORGANIZATION HIPPODROMIES S.A., ORGANIZATION
GRFinQA	Greek Original	Βραχυχρονίως, μία αύξηση των δημοσίων δαπανών Πιθανές απαντήσεις: Α) αυξάνει το επίπεδο τιμών αλλά όχι το πραγματικό ΑΕΠ Β) αυξάνει το πραγματικό ΑΕΠ αλλά όχι το επίπεδο τιμών Γ) αυξάνει το πραγματικό ΑΕΠ και το επίπεδο τιμών Δ) δεν αυξάνει ούτε το πραγματικό ΑΕΠ ούτε το επίπεδο τιμών	Γ
	English Translation	In the short term, an increase in public spending Possible answers: A) Increases the price level but not real GDP B) Increases real GDP but not the price level C) Increases both real GDP and the price level D) Increases neither real GDP nor the price level	C
GRFNS-2023	Greek Original	Τα μέλη του Διοικητικού Συμβουλίου της ΚΑΠΝΟΒΙΟΜΗΧΑΝ (...TRUNCATED)	Ετήσια Οικονομική Έκθεση της Χρήσης ΔΩΔΕΚΑΜΗΝΗ ΠΕΡΙΟ (...TRUNCATED)
	English Translation	The members of the Board of Directors of TOBACCO INDUSTRY (...TRUNCATED)	Annual Financial Report for the TWELVE-MONTH PERIOD (...TRUNCATED)
GRMultiFin	Greek Original	Αναστολή συμβάσεων εργασίας Αυγούστου	Επιχειρήσεις & Διοίκηση
	English Translation	Suspension of employment contracts in August	Business & Administration

Table 6: Conversion prompts for instruction data, presented with original Greek prompts alongside their English translations.

Dataset	Original Greek Prompt	English Translated Prompt
GRFinNUM	Στις παρακάτω προτάσεις που προέρχονται από οικονομικές εκθέσεις ελληνικών εταιρειών, αναγνώρισε αριθμητικές οντότητες που ανήκουν στις εξής κατηγορίες: χρηματικά ποσά (ΧΡΗΜΑΤΑ), ποσοστά (ΠΟΣΟΣΤΑ), χρονικές τιμές (ΧΡΟΝΙΚΑ), ποσότητες (ΠΟΣΟΤΗΤΕΣ) και άλλες αριθμητικές τιμές (ΑΛΛΑ). Η απαιτούμενη μορφή απάντησης είναι 'όνομα οντότητας, τύπος οντότητας'. Κείμενο: {Input} Απάντηση:	In the following sentences which originate from Greek Company filings, recognize the numeric entities which correspond to the following categories: monetary values (MONETARY), percentages (PERCENTAGES), temporal values (TEMPORAL), quantities (QUANTITIES) and other numeric values (OTHER). The required answer format is: "entity name, entity type". Text: {Input} Answer:
GRFinNER	Στις παρακάτω προτάσεις που προέρχονται από οικονομικές εκθέσεις ελληνικών εταιρειών, αναγνώρισε τις οντότητες που αντιπροσωπεύουν ένα πρόσωπο (ΠΡΟΣΩΠΟ), έναν οργανισμό (ΟΡΓΑΝΙΣΜΟΣ) ή μία τοποθεσία (ΤΟΠΟΘΕΣΙΑ). Η απαιτούμενη μορφή είναι: 'όνομα οντότητας, τύπος οντότητας'. Κείμενο: {Input} Απάντηση:	In the following sentences which originate from Greek Company filings, recognize the entities which correspond to a person ("Person"), an organization ("Organisation") or a location ("Location"). The required answer format is: "entity name, entity type". Text: {Input} Answer:
GRFinQA	Διάβασε προσεκτικά την παρακάτω ερώτηση και τις πιθανές απαντήσεις. Επίλεξε το γράμμα που αντιστοιχεί στη σωστή απάντηση. Ερώτηση: {Input} Απάντηση:	Read the following question and the possible answers carefully. Choose the letter which corresponds to the correct answer. Question: {Input} Answer:
GRFNS-2023	Σε παρακαλώ διάβασε το παρακάτω κείμενο και συνόψισε το σύντομα και με ακρίβεια. {Input}	Please read the following text and summarize it briefly and accurately. {Input}
GRMultiFin	Διάβασε το κείμενο προσεκτικά και επέλεξε την σωστή κατηγοριά για το κείμενο από τις κατηγορίες Φορολογία & Λογιστική, Επιχειρήσεις & Διοίκηση, Οικονομικά, Βιομηχανία, Τεχνολογία, Κυβέρνηση & Έλεγχοι. Κείμενο: {Input} Απάντηση:	Read the text carefully and choose the correct category for the text from the categories "Tax & Accounting", "Business & Management", "Finance", "Industry", "Technology", "Government & Controls". Text: {Input} Answer:

H GRFinNUM Annotation Guideline

To ensure consistent annotation of numerical entities in financial texts, we define the following annotation guidelines.

H.1 Entity Categories

We annotate five types of numerical entities:

- **Monetary**
- **Percentage**
- **Temporal**
- **Quantity**
- **Others**

H.2 General Annotation Rules

1. **Only numbers are annotated**: Include only numerical digits, decimal points (“.”), and the percent sign (“%”).
2. **Decimal delimiter exclusion**: When a decimal point is used as a delimiter (e.g., 2024.11.26), annotate each component separately as 2024, 11, and 26.
3. **Exclusion of textual numbers**: Text-based numbers (e.g., two weeks) are excluded, but numeric equivalents (e.g., 2 weeks) are included.
4. **Exclusion of non-numeric symbols**: Symbols such as “$” are not included.

H.3 Specific Entity Annotation Rules

H.3.1 Monetary

Numbers related to money, including explicit currencies or monetary values.

- Include: The numeric value in “$50” and “100 euros” $\rightarrow$ annotate as “50” and “100”.

H.3.2 Percentage

Numbers representing percentages, “%” symbol as part of the percentage.

- Include: “45%”, “0.5%”.

H.3.3 Temporal

Numbers related to time, such as years, dates, and durations.

- Include: only numbers in “2024”, “12.25”, “12/25”, “2 weeks”, “1 year” and “3 hours” should be included.
- Exclude: Words such as “two weeks”, where the number is not explicitly written in numeric form.

H.3.4 Quantity

Numbers representing measurable or countable quantities, excluding monetary values.

- Include: only numbers in “5 items” and “100 shares”.

H.3.5 Others

Numbers that do not fit into the above categories, such as identifiers, version numbers, numerical codes, or numeric positions.

- Include: only "3" in "3rd place", "2" and "1" in "v2.1", and "202" in "model 202".
- Exclude: "second investor" (textual ordinal numbers).

H.4 Annotation Examples

Text	Annotated Entity
"$50 was paid."	'50' (Monetary)
"45% of users agreed."	'45%' (Percentage)
"The event happened in 2024."	'2024' (Temporal)
"5 items were sold."	'5' (Quantity)
"Version v2.1 is released."	'2', '1' (Others)

Table 7: Examples of annotated numerical entities.

I GRFinNER Annotation Guideline

To ensure consistent annotation of named entities in financial texts, we define the following annotation guidelines.

I.1 Entity Categories

We annotate three types of named entities:

- **Person**
- **Location**
- **Organization**

I.2 General Annotation Rules

1. **Abbreviations**: Annotate them together if they appear together; otherwise, annotate them as two entities.
   - Include: "World Health Organization (WHO)" as one span.
2. **Ambiguous Terms**: Resolve ambiguity using context.
   - Include: "Amazon" as a company.
   - Exclude: "Amazon" as a river.
3. **General Terms Exclusion**: Exclude generic terms.
   - Exclude: "the professor", "downtown", "north", "the team".
4. **Definite Articles**: Exclude "the" from entity spans.
   - Exclude: "the" in "the WHO".
5. **Consecutive Entities**: When two entities are consecutive, annotate them separately except postal addresses.
   - Include separately: "London" and "United Kingdom" in "London United Kingdom".
   - Include separately: 'street Egnatias 127" and "Thessaloniki" in "street Egnatias 127 in Thessaloniki (Postal Code 54 635)".
   - Include separately: "Acharnes Attica" and "Parnithos Avenue" in "municipality of Acharnes Attica, 15 km Parnithos Avenue".
   - Include as one span: "5900 Penn Avenue, Pittsburgh".

I.3 Specific Entity Annotation Rules

I.3.1 Person

Names of individual people. Include real people, fictional characters, and usernames. Exclude animal names. Exclude titles that are not part of the legal name.

- Include: “Marie Curie”, “George Demetriou of Konstantinos”.
- Include only ‘John” in ‘Dr. John”.
- Exclude: “the professor”.

I.3.2 Location

Names of geographical places, such as cities, countries, natural landmarks, and fictional locations.

- Include: ‘Paris”, ‘Mount Everest”.
- Exclude: ‘downtown”, ‘north”.

I.3.3 Organization

Names of companies, institutions, and formal groups. Including words like “company”, “association”, “Inc.”, “Co.”, and “Ltd.”.

- Include: “World Health Organization”, “Tesla Inc.”, “WHO”, “OPAP Association”.
- Exclude: “the team”.

I.4 Special Cases

1. **Organizations with Location Names**: If the location refers to a specific organization, annotate both; otherwise, only annotate the location.
    - Include: Only ‘Cypriot” in ‘the Cypriot company”.
2. **Organizations Representing Administrative Units or Sports Teams**: Annotate as **Organization**.
    - Include: “Baltimore” and “Indianapolis” in “Baltimore lost to Indianapolis last weekend” as Organizations.

J Human Evaluation Annotation Guideline

To ensure consistent annotation of summarization quality in financial texts, we define the following annotation guidelines.

J.1 Evaluation Categories

We evaluate summaries based on three criteria:

- **Language Appropriate Fluency**
- **Coherence**
- **Factuality**

J.2 General Annotation Rules

1. **Language Appropriate Fluency (Fluency)**: Measures how well the summary aligns with the expected language fluency and domain-specific terminology.
    - 1 (Bad): Response is entirely in the wrong language (e.g., English instead of Greek).
    - 2 (Poor): Response is a mixture of English and Greek.
    - 3 (Okay): Response is fully in Greek but contains grammatical or lexical errors or repetition.
    - 4 (Good): Response is entirely in fluent Greek without grammatical or lexical errors or repetition.
    - 5 (Excellent): Response is entirely in fluent Greek with appropriate domain-specific terminology.
2. **Coherence**: Evaluates the logical progression and structure of ideas in the text.
    - 1 (Bad): The text is disorganized, with sentences or paragraphs lacking logical flow.
    - 2 (Poor): The text attempts structure but has logical leaps, disjoint ideas, and is confusing.
    - 3 (Okay): The text is mostly coherent, with a general structure and minor logical errors or awkward transitions.
    - 4 (Good): The text flows well, with clear progression and only minor errors.
    - 5 (Excellent): The text flows naturally and consistently, with smooth transitions between ideas.
3. **Factuality**: Evaluates whether the summary is factually consistent with the original content.
    - 1 (Bad): Multiple factual inaccuracies, such as misrepresented company names, locations, or numerical data.
    - 2 (Poor): Some factual errors with key points missing or distorted.
    - 3 (Okay): Fairly accurate, with only minor omissions or discrepancies.
    - 4 (Good): Accurate, with only a few minor omissions or discrepancies.
    - 5 (Excellent): Entirely accurate, with all facts presented as found in the source document.

K Annotator Demography

Our benchmark construction relies on the expertise of a team of highly qualified annotators, who are native Greek speakers with diverse backgrounds in computer science, mathematics, statistics, and finance. Their combined knowledge ensures the high-quality annotation of financial texts, contributing to the robustness and reliability of our dataset.

One annotator, currently pursuing a Ph.D. in Computer Science at a leading Greek university, has a strong foundation in both mathematics and statistics, complemented by industry experience as a credit risk analyst. This background provides valuable information on financial knowledge, risk assessment, and statistical modeling, which are essential to annotate our benchmark dataset.

Another annotator, a Ph.D. student in Computer Science at a major UK institution, holds an Integrated Master's degree in Electrical and Computer Engineering. Their expertise in computer science enhances the annotation process by ensuring precision and alignment with modern NLP techniques.

The team is further strengthened by a postdoctoral researcher with an interdisciplinary background spanning electrical and computer engineering, computer science, and mathematics. Having obtained a Ph.D. from a prestigious U.S. university, this annotator brings extensive research experience and a deep understanding of theoretical and applied aspects of financial computing, making them instrumental in refining annotation guidelines and resolving complex cases.

The collective expertise of our annotators is critical to the development of our Greek financial benchmark. Their deep familiarity with the Greek financial ecosystem, combined with strong computational and analytical skills, ensures that our dataset accurately reflects domain-specific nuances while maintaining linguistic and terminological precision. By leveraging their diverse backgrounds, we are able to construct a high-quality resource that will serve as a foundation for advancing NLP research in financial applications.

L Annotation Process

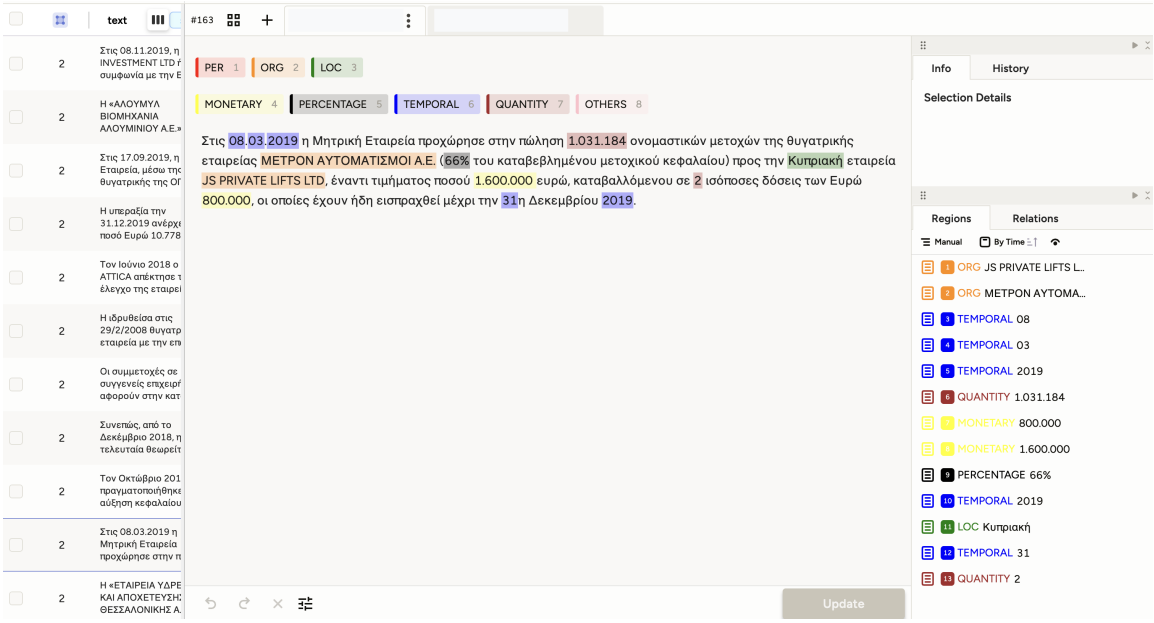

Figure 5: The Label Studio interface of the NER annotation process.

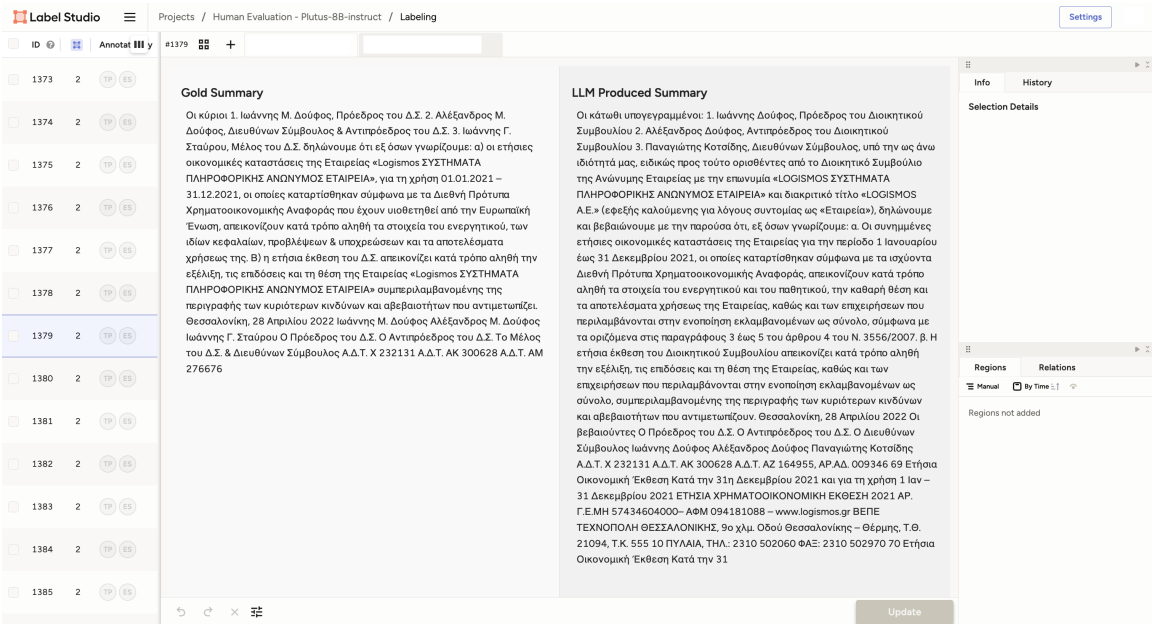

Figure 6: The Label Studio interface of the human evaluation process.

M Evaluation Metrics

The Entity F1 is the harmonic mean of Precision and Recall, calculated as follows.

$$P_{entity} = \frac{TP}{TP + FP} \tag{1}$$

$$R_{entity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Entity F1} = 2 \times \frac{P_{entity} \times R_{entity}}{P_{entity} + R_{entity}} \tag{3}$$

where P_{entity} and R_{entity} denote the Precision and Recall of entity prediction, respectively. TP (True Positive) represents the number of actual entities correctly identified. In contrast, FP (False Positive) refers to the number of non-entities incorrectly predicted as entities. FN (False Negative) denotes the number of entities that were not correctly predicted.

Accuracy Acc measures the proportion of correct predictions made by the model and is defined as follows.

$$\mathrm{Acc} = \frac{\text{Number of correct predictions}}{\text{Total number of Predictions}} \tag{4}$$

Rouge-1 is primarily used to compute the unigram-level (word-level) overlap between the generated summary and the reference summary, and is defined as follows:

$$P_{rouge1} = \frac{\text{Number of overlapping unigrams in generated and reference summary}}{\text{Total unigrams in generated summary}} \tag{5}$$

$$R_{rouge1} = \frac{\text{Number of overlapping unigrams in generated and reference summary}}{\text{Total unigrams in reference summary}} \tag{6}$$

$$\text{Rouge-1 F1} = 2 \times \frac{P_{rouge1} \times R_{rouge1}}{P_{rouge1} + R_{rouge1}} \tag{7}$$

where P_{rouge1} and R_{rouge1} denote the Precision and Recall of Rouge-1, respectively. Rouge-1 F1 is the final Rouge-1 score that calculates the unigram (single-word) matches without considering word order.

N Dataset Quality Validation

The F1-score, Cohen's Kappa, and Krippendorff's alpha were calculated to measure the agreement of annotators for data quality control purposes.

The F1-score is a performance metric for classification models that combines Precision and Recall using their harmonic mean as shown in the equation (8).

$$F1 - scores = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{8}$$

where *Precision* measures how many of the samples predicted as positive are actually positive; *Recall* measures the proportion of actual positive samples that the model correctly identifies.

Cohen's Kappa measures the agreement between two annotators on a classification task, accounting for the possibility of random agreement, as shown in equation (9).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{9}$$

where P_o means the observed agreement and P_e is the expected agreement.

Krippendorff's alpha is a general measure of inter-rater reliability applicable to categorical, ordinal, interval, or ratio data, as shown in equation (10).

$$\alpha = 1 - \frac{D_o}{D_e} \tag{10}$$

where D_o is the total disagreement observed among annotators, and D_e is the total disagreement expected by chance.