MULTIFINBEN: Benchmarking Large Language Models for Multilingual and Multimodal Financial Application

Xueqing Peng[1], Lingfei Qian[1], Yan Wang[1], Ruoyu Xiang[6], Yueru He[3], Yang Ren[1], Mingyang Jiang[1], Vincent Jim Zhang[1], Yuqing Guo[1], Jeff Zhao[1], Huan He[1], Yi Han[4], Yun Feng[17], Yuechen Jiang[7], Yupeng Cao[7], Haohang Li[7], Yangyang Yu[7], Xiaoyu Wang[6], Penglei Gao[8], Shengyuan Lin[20], Keyi Wang[3], Shanshan Yang[7], Yilun Zhao[2], Zhiwei Liu[10], Peng Lu[11], Jerry Huang[11], Suyuchen Wang[11], Triantafillos Papadopoulos[15,18], Polydoros Giannouris[10], Efstathia Soufleri[18], Nuo Chen[19], Zhiyang Deng[7], Heming Fu[1], Yijia Zhao[21], Mingquan Lin[13], Meikang Qiu[14], Kaleb E. Smith[5], Arman Cohan[2], Xiao-Yang Liu[3], Jimin Huang[1,10*], Guojun Xiong[12], Alejandro Lopez-Lira[9*], Xi Chen[6], Junichi Tsujii[16], Jian-Yun Nie[11], Qianqian Xie[1], Sophia Ananiadou[10,18]

[1]*The Fin AI*, [2]*Yale University*, [3]*Columbia University*, [4]*Georgia Institute of Technology*, [5]*NVIDIA*, [6]*New York University*, [7]*Stevens Institute of Technology*, [8]*Quantitative Health Sciences, Cleveland Clinic*, [9]*University of Florida*, [10]*University of Manchester*, [11]*Université de Montréal*, [12]*Harvard University*, [13]*University of Minnesota*, [14]*Augusta University*, [15]*Athens University of Economics and Business*, [16]*National Institute of Advanced Industrial Science and Technology*, [17]*Asian Development Bank*, [18]*Archimedes, Athena Research Center*, [19]*National University of Singapore*, [20]*Carnegie Mellon University*, [21]*University of Massachusetts Boston*

Correspondence: jimin.huang@thefin.ai, alejandro.lopez-lira@warrington.ufl.edu

Abstract

Real-world financial analysis involves information across multiple languages and modalities, from reports and news to scanned filings and meeting recordings. Yet most existing evaluations of LLMs in finance remain text-only, monolingual, and largely saturated by current models. To bridge these gaps, we present MULTIFINBEN, the first expert-annotated multilingual (five languages) and multimodal (text, vision, audio) benchmark for evaluating LLMs in realistic financial contexts. MULTIFINBEN introduces two new task families: multilingual financial reasoning, which tests cross-lingual evidence integration from filings and news, and financial OCR, which extracts structured text from scanned documents containing tables and charts. Rather than aggregating all available datasets, we apply a structured, difficulty-aware selection based on advanced model performance, ensuring balanced challenge and removing redundant tasks. Evaluating 21 leading LLMs shows that even frontier multimodal models like GPT-4o achieve only 46.01% overall, stronger on vision and audio but dropping sharply in multilingual settings. These findings expose persistent limitations in multilingual, multimodal, and expert-level financial reasoning. All datasets, evaluation scripts, and leaderboards are publicly released. [1]

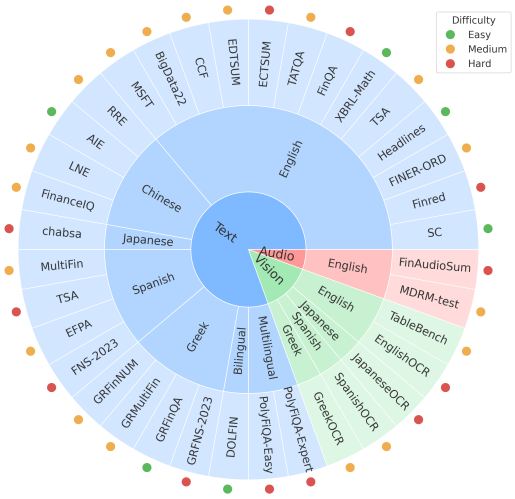


Figure 1: Overview of MULTIFINBEN.

Benchmark	EN	ZH	JA	ES	EL	BI	MU	Vis	OCR	Au	DA
Plutus (Peng et al., 2025)	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘	✘
CFinBench (Nie et al., 2024)	✘	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘
FLARE-ES (Zhang et al., 2024b)	✔	✘	✘	✔	✘	✘	✘	✘	✘	✘	✘
FinBen (Xie et al., 2024a)	✔	✘	✘	✔	✘	✘	✘	✘	✘	✘	✘
PIXIU (Xie et al., 2023)	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
DOLFIN (Nakhlé et al., 2025)	✔	✘	✘	✔	✘	✔	✘	✘	✘	✘	✘
FinMME (Luo et al., 2025)	✔	✘	✘	✘	✘	✘	✘	✔	✘	✘	✘
MME-Finance (Gan et al., 2024)	✔	✔	✘	✘	✘	✘	✘	✔	✘	✘	✘
FinAudio (Cao et al., 2025b)	✔	✘	✘	✘	✘	✘	✘	✘	✘	✔	✘
MULTIFINBEN (ours)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Table 1: Comparison with existing benchmarks. EN: English, ZH: Chinese, JA: Japanese, ES: Spanish, EL: Greek, BI: Bilingual, MU: Multilingual, Vis: Chart/Tabular, OCR: Optical character recognition, Au: Audio, DA: Difficulty-aware.

[1]Data: https://huggingface.co/collections/TheFinAI/multifinben
Code: https://github.com/xueqingpeng/MultiFinBen
Leaderboard: https://huggingface.co/spaces/TheFinAI/Open-FinLLM-Leaderboard

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16934–16963
July 2-7, 2026

1 Introduction

LLMs have demonstrated remarkable prowess in various tasks (Zhao et al., 2025; Chen et al., 2025b), yet their applications to finance (Liu et al., 2023b; Xie et al., 2023; Wu et al., 2023) require the continual push of the limits of model capabilities (Xiong et al., 2025). Although there are already several specialized benchmarks (Xie et al., 2024b; Li et al., 2024; Cao et al., 2025b) that evaluate LLM in core financial tasks, they still suffer from three critical limitations:

- *Monolingual and monomodal scope.* Prior benchmarks largely focus on English text only, whereas real-world financial applications demand cross-lingual understanding and multimodal reasoning across text, tables, charts, PDFs, and audio.

- *Imbalanced task difficulty.* Existing benchmarks are typically constructed by simple aggregation without difficulty calibration. This leads to an overrepresentation of trivial tasks. For example, in FinBen (Xie et al., 2024b), 7 out of 8 text classification datasets can already be solved above 60% accuracy in zero-shot, resulting in inflated scores and obscuring true model weaknesses.

- *Missing key financial scenarios.* No prior benchmark addresses multilingual financial reasoning that integrates heterogeneous sources, nor OCR tasks that require extracting and reasoning over scanned financial PDFs. These scenarios are ubiquitous in practice yet entirely absent in prior benchmarks.

Benchmark Design. To address these gaps, we introduce MULTIFINBEN, the first unified benchmark that spans three modalities (text, vision, and audio), three linguistic settings (monolingual, bilingual, and multilingual), and seven task categories across three difficulty tiers. In total, it comprises 36 datasets in five languages (English, Chinese, Japanese, Spanish, and Greek). Unlike prior benchmarks limited to monolingual or unimodal corpora, MULTIFINBEN introduces two novel multilingual financial reasoning datasets, *PolyFiQA-Easy* and *PolyFiQA-Expert*, requiring joint reasoning over mixed-language inputs. In addition, we present the first financial OCR datasets (*EnglishOCR*, *JapaneseOCR*, *SpanishOCR*, and *GreekOCR*) to evaluate

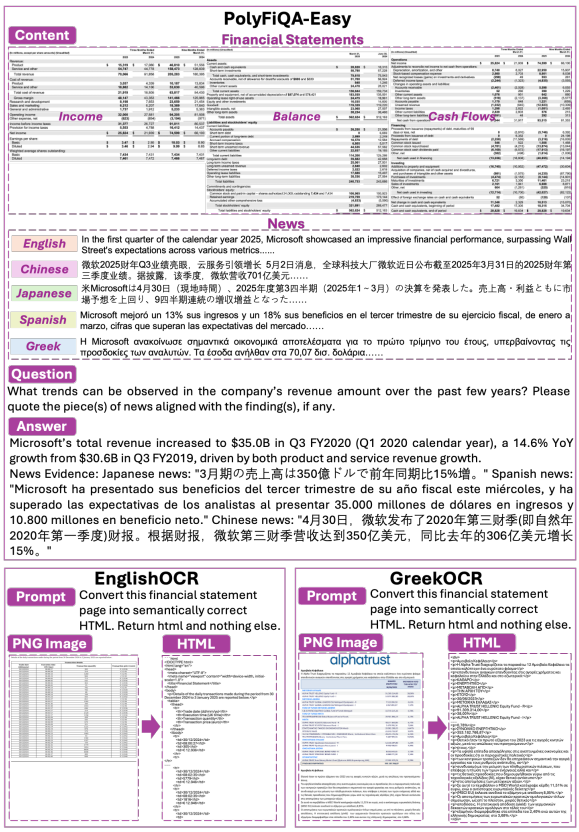


Figure 2: Representation examples of *PolyFiQA-Easy*, *EnglishOCR*, and *GreekOCR*.

models on document-image understanding, a setting ubiquitous in practice but overlooked in prior work. Finally, we propose a difficulty-aware selection mechanism that retains one dataset per modality–language–task tier with the largest inter-model gap, avoiding redundancy and ensuring meaningful evaluation.

Evaluation and Findings. We benchmark 21 frontier models on MULTIFINBEN. Even the strongest model, GPT-4o, achieves only 46.01%, while monolingual and unimodal models fail dramatically on unsupported inputs, underscoring the necessity of cross-modal reasoning. Its significant low performance in multilingual scenario (9.79% and 5.31%) highlights persistent weaknesses in cross-lingual generalization. Our difficulty-aware design reveals steep drops from 32.73% (easy) to 7.20% (hard), exposing the gap between current model capabilities and real-world complexity. Crucially, our newly introduced datasets (*PolyFiQA-Easy/Expert*, *EnglishOCR*, and *JapaneseOCR*) emerge as among the hardest challenges, demonstrating how MULTIFINBEN systematically surfaces limitations and provides a roadmap for

developing more realistic financial evaluation settings.

Our Contributions. We summarize the main contributions of this work as follows:

▷ *First unified multilingual and multimodal benchmark for financial LLMs.* MULTIFINBEN spans five languages (EN, ZH, JA, ES, EL), three modalities (text, vision, audio), and seven task categories, offering the most comprehensive evaluation setting.

▷ *Novel and realistic datasets.* We introduce *PolyFiQA-Easy* and *PolyFiQA-Expert*, the first multilingual financial QA datasets requiring cross-lingual reasoning, as well as *EnglishOCR*, *JapaneseOCR*, *SpanishOCR*, and *GreekOCR*, the first financial OCR tasks targeting scanned financial documents.

▷ *Difficulty-aware evaluation framework.* We propose a principled dataset selection mechanism based on inter-model performance gaps, ensuring balanced coverage across easy to hard tiers, in contrast to prior aggregation-based benchmarks.

▷ *Comprehensive model evaluation and insights.* We evaluate 21 state-of-the-art models and reveal substantial performance gaps across modalities, languages, and difficulty levels, with even GPT-4o achieving only 46.01%. This demonstrates the necessity of our benchmark for stress-testing current models and guiding future development.

▷ *Open resources.* We release all datasets and code to foster transparency, reproducibility, and future extensions by the research community.

2 MULTIFINBEN Benchmark

In this section, we present MULTIFINBEN, the first unified benchmark for evaluating LLMs in the financial domain across diverse modalities, linguistic settings, and tasks (Table 2).

2.1 Overview

Built upon this foundation, MULTIFINBEN systematically organizes datasets along three structural dimensions (i.e., modality, language, and task) to enable comprehensive and interpretable evaluation. Rather than aggregating heterogeneous resources, it introduces a difficulty-aware benchmarking framework that selects representative datasets based on their discriminative power across models, ensuring balanced coverage and meaningful scalability. To address gaps in existing resources, we further design new datasets and tasks that capture underrepresented yet crucial financial scenarios. Together, these principles establish MULTIFINBEN as a unified, rigorous, and extensible framework for assessing financial reasoning across modalities, languages, and task types.

Modalities. Beyond text, it uniquely integrates visual (charts, tables, images) and audio (earnings calls) modalities into a unified financial benchmark, reflecting common formats of financial dissemination and enabling realistic assessment of multimodal understanding in financial contexts.

Linguistics. MULTIFINBEN is the first to evaluate LLMs under three linguistic settings (monolingual, bilingual, and multilingual) covering five typologically and economically diverse languages (English, Chinese, Japanese, Spanish, and Greek). This design spans major global financial regions, providing a realistic setting for evaluating cross-lingual financial understanding.

Task Categories. Inspired by FinBen (Xie et al., 2024a), our datasets are organized into seven core financial NLP tasks (Appendix B): Information Extraction (IE), Textual Analysis (TA), Question Answering (QA), Text Generation (TG), Risk Management (RM), Forecasting (FO), and Decision-Making (DM). This organization ensures broad task coverage and supports fine-grained evaluation of reasoning capabilities across diverse financial applications.

2.2 Structured Difficulty-Aware Benchmarking

Existing financial and cross-domain benchmarks often emphasize breadth, aggregating large numbers of datasets across tasks and modalities. While such aggregation maximizes coverage, it obscures the underlying sources of difficulty and makes model progress difficult to interpret. In contrast, MULTIFINBEN adopts a structured, difficulty-aware design that disentangles evaluation along three orthogonal dimensions, i.e., modality, language, and task, enabling controlled comparison and clear attribution of performance to specific reasoning or linguistic factors (Suzgun et al., 2022b; Glazer et al., 2024).

To quantify dataset difficulty in a reproducible and model-agnostic manner, we compute the mean standardized performance of two representative large language models, GPT-4o (Hurst et al., 2024) and LLaMA-3.1-70B-Instruct (Dubey et al., 2024),

Modality	Language	Task	Dataset	Source	Size	Metric	License	Difficulty
Text	English	IE	SC (Mariko et al., 2020)	Financial News	8,630	F1	CC BY 4.0	Easy
		IE	FinRED (Sharma et al., 2022)	Webhose Financial News and Earning Call Transcrip	1,068	F1	Public	Hard
		IE	FINER-ORD (Shah et al., 2023)	Financial News	1,080	Entity F1	CC BY-NC 4.0	Medium
		TA	Headlines (Sinha and Khandait, 2021)	Financial News	2,283	Avg F1	CC BY-SA 3.0	Easy
		TA	TSA (Cortis et al., 2017)	Financial Microblog	561	Accuracy	CC BY-NC-SA 4.0	Medium
		QA	XBRL-Math (Chen et al., 2022)	XBRL Agent	1,000	Accuracy	CDLA-Permissive 2.0	Easy
		QA	FinQA (Chen et al., 2021b)	Annual Reports	1,147	Accuracy	MIT License	Hard
		QA	TATQA (Zhu et al., 2021)	Annual Reports	1,668	Accuracy	MIT License	Medium
		TG	ECTSUM (Mukherjee et al., 2022)	Earnings Call Transcripts from the Motley Fool	495	ROUGE-1	Public	Hard
		TG	EDTSUM (Xie et al., 2023)	Financial News	2,000	ROUGE-1	Public	Medium
		RM	CCF (Feng et al., 2024)	Credit Card Transactions	2,278	MCC	DbCL v1.0	Medium
		FO	BigData22 (Soun et al., 2022)	Tweets, Historical Prices	1,470	MCC	Public	Medium
		DM	MSFT (Yu et al., 2024)	TradeTheEvent Dataset	222	SR	Open Source	Medium
	Chinese	IE	RRE (Huang et al., 2024a)	Regulatory Documents	117	Accuracy	CC BY-NC	Medium
		TA	AIE (Huang et al., 2024a)	Regulatory Documents	1,573	Accuracy	CC BY-NC	Easy
		TA	LNE (Huang et al., 2024a)	Regulatory Documents	218	Accuracy	CC BY-NC	Medium
		QA	FinanceIQ (Zhang and Yang, 2023)	Exam Questions	7,123	Accuracy	CC BY-NC-SA 4.0	Medium
	Japanese	TA	chabsa (Kubo et al., 2018)	Securities Reports	7,723	Macro F1	CC BY 4.0	Hard
	Spanish	TA	MultiFin (Jørgensen et al., 2023a)	Article Headlines	368	Accuracy	MIT License	Medium
		TA	TSA (Pan et al., 2023)	News Headlines	726	Accuracy	Public	Hard
		QA	EFPA [5]	Exam Questions	50	Accuracy	Public	Medium
		TG	FNS-2023 [6]	Annual Reports	50	ROUGE-1	Public	Hard
	Greek	IE	GRFinNUM (Peng et al., 2025)	Annual Reports	100	Entity F1	Public	Medium
		TA	GRMultiFin (Jørgensen et al., 2023b)	Article Headlines	54	Accuracy	CC BY-NC 4.0	Medium
		QA	GRFinQA (Peng et al., 2025)	Exam Questions	225	Accuracy	Public	Easy
		TG	GRFNS-2023 (Zavitsanos et al., 2023)	Annual Reports	50	ROUGE-1	CC BY 4.0	Hard
	Bilingual	TG	DOLFIN (Nakhlé et al., 2025)	Fundinfo Financial Documents	1,932	Comet-da-22	MIT License	Easy
	Multilingual	QA	**PolyFiQA-Easy**	Financial Reports[1]and News[7]	204	ROUGE-1	Public	Hard
		TG	**PolyFiQA-Expert**	Financial Reports[1]and News[7]	204	ROUGE-1	Public	Hard
Vision	English	IE	**EnglishOCR**	SEC EDGAR Company Filings[1]	7,961	ROUGE-1	Public	Hard
		QA	TableBench (Xie et al., 2024c)	SynthTabNet (Fintabnet and Marketing categories)	450	Accuracy	Public	Medium
	Japanese	IE	**JapaneseOCR**	FSA White Paper[2]	17,586	ROUGE-1	Public	Hard
	Spanish	IE	**SpanishOCR**	Regulatory Documents[3]	12,819	ROUGE-1	Public	Medium
	Greek	IE	**GreekOCR**	Annual Company Filings on Athens Stock Exchange[4]	6,533	ROUGE-1	Public	Medium
Audio	English	TG	MDRM-test (Qin and Yang, 2019)	Earnings Conference Calls	22,208	WER	Public	Medium
		TG	FinAudioSum (Cao et al., 2025b)	Earnings Conference Calls	64	ROUGE-L	Public	Hard

[1] https://www.sec.gov/search-filings
[2] https://www.fsa.go.jp/en/
[3] https://www.bvl.com.pe/en/home-general
[4] https://www.athexgroup.gr/en/market-data/issuers
[5] https://efpa-eu.org/
[6] https://wp.lancs.ac.uk/cfie/fns2023/
[7] Please check Appendix H.2 for more details.

Table 2: Overview of MULTIFINBEN.

defined as

$$\bar{s}_d = \frac{1}{2}\left(s_{\text{GPT-4o},d} + s_{\text{LLaMA3.1},d}\right). \quad (1)$$

These models are chosen for their complementary characteristics: GPT-4o provides frontier-level reasoning and instruction alignment, while LLaMA-3.1-70B-Instruct serves as an open, transparent baseline. Their differences in architecture, scale, and training corpus yield a balanced estimate of intrinsic dataset complexity, independent of individual model bias.

We categorize datasets into three tiers, *easy* ($\bar{s}_d > 60$), *medium* ($20 \leq \bar{s}_d \leq 60$), and *hard* ($\bar{s}_d < 20$), corresponding to reliable competence, transitional reasoning, and consistent failure, respectively (Figure 3; Table 5; Appendix D). Within each modality–language–task configuration, we retain the dataset exhibiting the largest inter-model divergence, as such cases most effectively reveal capability boundaries. In the event of ties, the lower-performing dataset is selected to preserve headroom for future progress. This principled selection process yields a compact yet diagnostic benchmark that balances interpretability with challenge, supporting layered evaluation across easy, medium, and hard regimes while maintaining scalability and discriminative power.

2.3 Benchmark Composition

MULTIFINBEN is organized around two complementary sources of datasets, aligning with our three evaluation dimensions: modality, language, and task. (*i*) Newly introduced resources expand the benchmark's task coverage, including two multilingual financial QA datasets (*PolyFiQA-Easy* and *PolyFiQA-Expert*; Section 2.4.1) and four OCR-based datasets for document-image understanding (*EnglishOCR*, *JapaneseOCR*, *SpanishOCR*, and *GreekOCR*; Section 2.4.2). (*ii*) Existing financial benchmarks are integrated through our structured, difficulty-aware selection framework, ensuring balanced representation across text, vision, and audio modalities and across five languages. Together, these resources define the complete task suite of MULTIFINBEN, supporting evaluation of financial reasoning, comprehension, and generation across multilingual and multimodal settings.

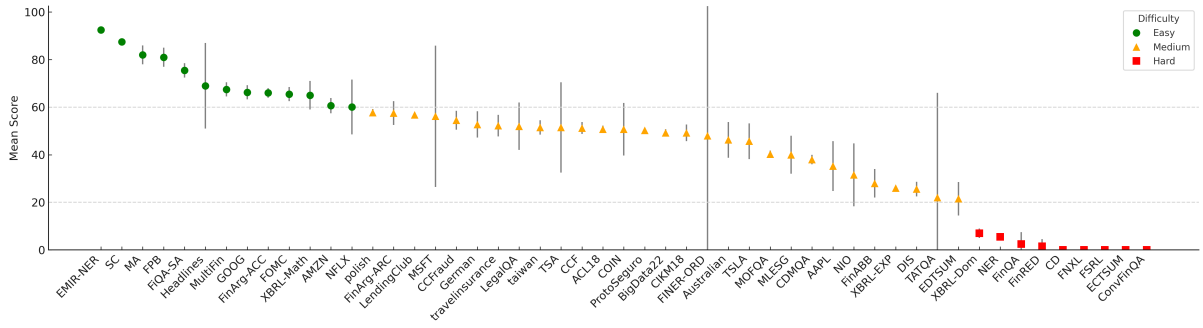


Figure 3: Structured difficulty-aware benchmarking of English datasets.

Text. In the monolingual setting, 26 datasets across five languages are selected from 68 available candidates following our difficulty calibration (Appendix E). In the bilingual setting, we include the English–Spanish subset of *DOLFIN* (Nakhlé et al., 2025), a document-level financial translation benchmark. As no public datasets exist for multilingual financial reasoning, we introduce two new multilingual QA datasets, *PolyFiQA-Easy* and *PolyFiQA-Expert*, designed to evaluate cross-lingual comprehension and reasoning.

Vision. For the visual modality, we integrate two English QA datasets from Open-FinLLMs (Xie et al., 2024c). Among these, *TableBench* is retained due to its greater inter-model performance variance, offering richer signal for difficulty assessment. We further contribute four newly constructed OCR-based datasets (Section 2.4.2) to evaluate document-level text extraction and structural understanding across languages.

Audio. For the audio modality, we initially consider five English text-generation datasets from FinAudio (Cao et al., 2025b). After applying our structured filtering, two datasets are retained: *MDRM* (Qin and Yang, 2019), a medium-tier ASR task featuring short earnings call clips, and *FinAudioSum* (Cao et al., 2025b), a hard-tier summarization task comprising long-form financial recordings (Appendix F).

Together, these resources enable systematic evaluation of financial understanding across text, vision, and audio modalities under consistent difficulty calibration.

2.4 Novel Tasks and Datasets

To close critical gaps in multilingual and visual financial reasoning, we introduce two new task families within MULTIFINBEN: (1) Multilingual Financial Question Answering (i.e., *PolyFiQA*), the first task for multilingual cross-document reasoning grounded in native financial sources; and (2) Financial OCR, the first OCR task for structured extraction from scanned financial documents. Together, these tasks extend financial evaluation beyond text-based English corpora, bringing cross-lingual reasoning and document-level visual understanding into one unified framework.

2.4.1 Multilingual Financial QA (*PolyFiQA*)

Unlike prior work relying on translation-based alignment, *PolyFiQA* draws directly from native-language financial disclosures and contemporaneous news, preserving authentic phrasing, cultural framing, and domain complexity. This design enables realistic evaluation of LLMs in real-world multilingual financial contexts, where decision-making increasingly spans languages and jurisdictions.

Task Definition. Provided a multilingual context

$$C = \{R, N_{\text{en}}, N_{\text{zh}}, N_{\text{ja}}, N_{\text{es}}, N_{\text{el}}\}, \qquad (2)$$

where R denotes a financial report (10-K or 10-Q) and N_{lang} represents contemporaneous news articles in five languages. Given a carefully designed natural language question q and the associated multilingual context C, the model must generate an answer a grounded in integrated multilingual information. This task challenges models to perform *multilingual cross-document reasoning*, mirroring how analysts synthesize signals from diverse linguistic sources under real financial conditions.

Data Construction. We release two complementary datasets: *PolyFiQA-Easy* and *PolyFiQA-Expert*, targeting different reasoning depths (Appendix H.4 and H.5). Financial filings are col-

lected from SEC EDGAR [2] and paired with temporally aligned multilingual news (Appendix H.2). To maintain focus and reduce noise, we extract three core statements from lengthy financial reports: Comprehensive Income, Consolidated Balance Sheets, and Cash Flows. Low-resource cases are supplemented with expert-authored, native-language news verified by bilingual financial professionals (Appendix H.3).

To ensure domain fidelity, we adopt an ***expert-in-the-loop*** pipeline (Figure 5, Appendix H.1): three financial experts (Appendix H.7) curated, annotated, and validated all data, spending over 130 hours in Label Studio (Appendix H.8). Tier-specific guidelines (Appendix H.6) and iterative pilots ensured consistency. Each question was authored and scored for two dimensions, i.e., Relevance (1–4) and Consistency (1–3), by multiple annotators following a structured validation protocol (Appendix H.9). Only instances with cumulative scores above 5 were retained, yielding inter-annotator agreement above 89% for both tiers (Appendix H.10). The raw datasets were further converted into structured instruction datasets with task-specific prompts (Appendix H.11). This rigorous process results in a high-quality, auditable dataset for fine-grained evaluation of multilingual financial reasoning. More details can be found in Appendix H.1.

Evaluation Metrics. We adopt ROUGE-1 (Lin, 2004) to measure unigram overlap between model predictions and references, offering a proxy for content coverage and factual alignment in multilingual QA and TG tasks. Further evaluation results with complementary metrics, including BERTScore (Zhang et al., 2020) and numeric-consistency checks (Chen et al., 2021a), are provided in Table 9 (Appendix H.12).

2.4.2 Optical Character Recognition (*OCR*)

While financial communication heavily relies on PDFs, existing benchmarks (Xie et al., 2024b) focus on visual QA or chart reasoning rather than document-structured text recovery. We introduce the first multilingual financial OCR task, targeting end-to-end extraction of structured content from scanned filings, regulatory documents, and financial white papers. This task bridges the gap between visual understanding and textual reasoning, enabling evaluation of multimodal systems on real-world document pipelines.

Task Definition. The OCR task is defined as a structured information extraction problem from document images. Each financial PDF document is segmented into a set of page-level images $\{I_1, I_2, \ldots, I_n\}$, where each image I_i corresponds to a single page. The model processes each image individually and generates a corresponding HTML-formatted text sequence T_i, such that $T_i = \text{OCR}(I_i)$, preserving both textual content and document structure (e.g., headings, tables, and paragraphs). The goal is to evaluate structural fidelity and semantic recovery, not just text recognition.

Data Construction. Following the task design, we construct four novel datasets in a multilingual setting: *EnglishOCR*, *JapaneseOCR*, *SpanishOCR*, and *GreekOCR* (Table 2). The *EnglishOCR* dataset is built using U.S. SEC EDGAR filings[3], which are primarily distributed in HTML format. For documents without native PDF versions, we use `wkhtmltopdf` to render PDF files from the corresponding HTML sources. Each PDF is segmented into page-level PNG images aligned with its corresponding HTML content. To match each image with the most relevant HTML snippet, we compute cosine similarity between OCR-extracted text and HTML sentences using a Sentence-BERT (SBERT) model, resulting in 7,961 aligned image–HTML pairs. The *JapaneseOCR*, *SpanishOCR*, and *GreekOCR* datasets are constructed using source PDFs from Japanese Financial Services Agency (FSA) white papers[4], Peruvian public regulatory documents[5], and Athens Stock Exchange annual company filings[6], respectively. Each PDF is decomposed into page-level PNG images, and the corresponding HTML content is generated by applying OCR to extract content and wrapping it in HTML tags to preserve document structure. We further apply quality-control by removing pages with excessive OCR errors and post-processing malformed outputs, such as repeated or garbled characters. After filtering, these datasets contain 17,586, 12,819, and 6,533 aligned image–HTML pairs, respectively. Detailed statistics are provided in Table 7 (Appendix G.1). The

[2] https://www.sec.gov/search-filings

[3] https://www.sec.gov/search-filings

[4] https://www.fsa.go.jp/en/

[5] https://www.bvl.com.pe/en/home-general

[6] https://www.athexgroup.gr/en/market-data/issuers

raw datasets were further converted into structured instruction datasets with task-specific prompts (Appendix G.3).

Evaluation Metrics. We employ ROUGE-1 (Lin, 2004) to measure lexical and structural overlap between predicted and reference HTML sequences. Further evaluation results with BERTScore (Zhang et al., 2020), numeric-consistency checks (Chen et al., 2021a), and edit distance (Wei et al., 2025) are provided in Table 8 (Appendix G.2).

3 Experimental Results

To rigorously evaluate the effectiveness of MULTIFINBEN in exposing model limitations and guiding future development, we structure our experiments around four key research questions (RQs): **RQ1:** How severe are the performance deficits of current LLMs when confronted with the multilingual and multimodal scenarios ubiquitous in real-world finance? **RQ2:** Are there significant biases or trade-offs in model performance across different modalities? **RQ3:** Can the model's monolingual capabilities be effectively generalized to complex multilingual tasks that require cross-lingual comprehensive reasoning? **RQ4:** How does our difficulty-aware framework help analyze the strengths and weaknesses of the model in more detail?

Evaluation Models. We evaluate 21 models (Table 4, Appendix C) spanning text, vision, audio, and multimodal modalities, considering their openness under the Model Openness Framework (MOF, Appendix J) (White et al., 2024b).

Implementation Details. To ensure evaluation integrity and consistency, we customize our evaluation pipeline based on the LM Evaluation Harness (Gao et al., 2024). OpenAI and TogetherAI-hosted models (DeepSeek, Llama-4, and Gemma-3) are accessed via their official APIs with temperature set to 0. All other open-source models are deployed and evaluated locally using vLLM (Kwon et al., 2023) on GPUs. Including the OpenAI and TogetherAI API costs, the total expenditure amounts to approximately $80,000.

Key Results. Table 3 presents model performance on the MULTIFINBEN benchmark[7]. To ensure fair comparison, we additionally report a modality-balanced overall score for each model. The major findings are summarized as follows.

Finding 1 for RQ1: Cross-modal and cross-lingual reasoning in MULTIFINBEN *exposes clear limits of current LLMs.* Building upon our structured, difficulty-aware benchmarking framework, MULTIFINBEN consists of 6 easy, 18 medium, and 12 hard datasets spanning three modalities and five economically critical languages across three linguistic settings. All multilingual, vision, and audio datasets fall into the medium or hard tiers, posing nontrivial challenges even for the most advanced models. The leading model, GPT-4o, achieves an overall score of only 46.01%, revealing clear limitations even at the frontier of multimodal and multilingual reasoning. The next best models, Qwen2.5-Omni and Llama-4, reach 33.53% and 20.89%, respectively, underscoring the steep difficulty gradient established by our benchmark. Notably, all three top-performing systems are multimodal and multilingual, reinforcing the necessity of these capabilities for robust financial reasoning. In contrast, monomodal and monolingual models perform considerably worse. The strongest text-only model, Llama-3.1-70B, attains merely 14.07% overall (with a best text-modality score of 42.20%). Likewise, modality- or language-specific models such as Whisper-V3 (audio-only, 17.19%), LLaVA-1.6 (vision-only, 5.15%), and FinMA-7B (English-only, 8.94%) trail far behind. These disparities highlight the inherent limitations of models lacking cross-modal and cross-lingual reasoning, emphasizing the critical need for integrated multimodal and multilingual understanding in financial AI, precisely the kind of comprehensive evaluation that MULTIFINBEN is designed to provide.

Finding 2 for RQ2: Multimodal models exhibit trade-offs across modalities. Figure 4a illustrates the stratified performance across modalities. While GPT-4o leads overall performance on MULTIFINBEN, it is surpassed by the text-specialized Llama-3.1-70B in text-only tasks (42.20% vs. 40.98%). Similarly, other leading multimodal models such as Qwen2.5-Omni and Llama-4 fall short of their text-only counterparts Qwen2.5-32B (38.07% vs. 33.06%) and Llama-3.1-70B (42.20% vs. 39.50%) in textual evaluations. This trade-off becomes more pronounced in mid-tier models like Gemma-3-4B and Gemma-3-27B, which show comparable vision-task performance (18.88% and 26.80%) to vision-only models (LLaVA-1.6: 15.45%, Deepseek-VL: 14.14%) but

[7]Results are also visualized on our leaderboard. For more details, refer to Appendix K.

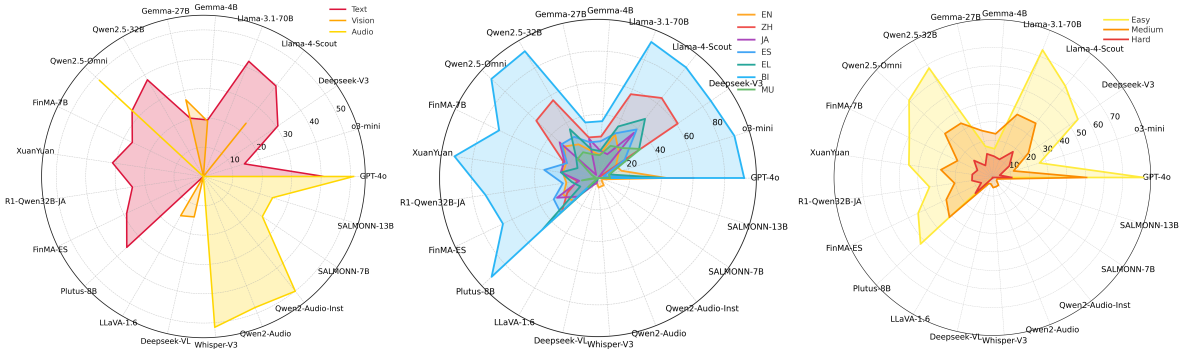
Modality	Language	Task	Dataset	GPT-4o	o3-mini	Deepseek-V3	Llama-4-Scout	Llama-3.1-70B	Gemma-4B	Gemma-27B	Qwen2.5-32B	Qwen2.5-Omni	FinMA-7B	XuanYuan	R1-Qwen32B-JA	FinMA-ES	Plutus-8B	LLaVA-1.6	Deepseek-VL	Whisper-V3	Qwen2-Audio	Qwen2-Audio-Inst	SALMONN-7B	SALMONN-13B
Text	EN	IE	SC	88.00	0.00	0.00	20.73	87.00	0.69	0.00	22.28	18.61	56.62	24.09	15.92	52.70	19.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		IE	FinRED	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.09	0.00	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		IE	FINER-ORD	78.00	9.58	0.18	2.23	18.00	0.00	0.00	28.30	8.30	0.04	0.00	0.00	0.00	5.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TA	Headlines	78.00	0.00	47.32	84.33	60.00	0.00	0.00	85.42	82.23	97.08	85.10	82.21	94.69	71.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TA	TSA	61.00	0.00	0.85	63.82	42.00	32.34	32.34	42.98	43.40	81.70	85.11	60.00	86.38	54.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		QA	XBRL-Math	68.00	68.89	67.78	27.78	62.00	0.00	11.11	64.44	44.44	7.78	3.33	2.22	6.67	14.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		QA	FinQA	5.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.41	0.00	0.00	0.00	1.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		QA	TATQA	0.00	0.00	0.00	0.36	44.00	0.00	0.00	0.05	1.73	4.14	0.00	0.00	0.00	15.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TG	ECTSUM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TG	EDTSUM	25.00	19.13	16.80	16.59	18.00	0.98	0.10	20.16	23.89	19.92	12.49	8.06	2.06	13.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		RM	CCF	52.50	50.00	50.62	51.34	50.00	50.93	50.00	52.94	50.31	50.05	50.00	50.00	51.18	50.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		FO	BigData22	48.50	50.00	50.93	46.91	50.00	50.75	50.00	49.89	51.82	50.80	53.12	50.00	52.12	50.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		DM	MSFT	41.32	65.06	0.00	0.00	72.25	74.03	79.97	49.32	0.00	0.00	68.81	74.50	66.53	65.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ZH	IE	RRE	63.25	0.00	67.52	54.70	46.15	36.75	36.75	8.55	7.69	0.85	2.56	0.85	0.85	2.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TA	AIE	82.26	0.00	82.01	80.99	76.80	33.82	33.82	83.03	80.17	40.81	10.04	4.32	21.55	54.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TA	LNE	63.30	0.00	58.72	55.50	41.28	9.17	9.17	57.80	59.17	29.82	22.48	12.84	32.11	26.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		QA	FinancialIQ	32.53	0.00	35.52	66.83	62.71	25.19	25.20	77.09	65.32	26.21	57.07	34.70	31.48	40.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	JA	TA	chabsa	0.00	0.00	0.00	48.43	32.17	8.98	23.96	4.54	44.35	46.94	47.59	23.96	57.36	34.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ES	TA	MultiFin	61.74	0.00	53.91	62.17	48.26	22.17	22.17	46.52	46.96	43.04	31.74	12.61	44.78	51.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TA	TSA	0.39	0.00	29.17	52.29	24.29	63.04	63.46	31.63	46.46	31.03	68.19	63.38	16.64	51.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		QA	EFPA	31.14	0.00	18.86	67.54	66.67	25.44	25.44	65.79	55.70	32.46	65.79	25.44	91.67	48.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TG	FNS-2023	25.94	18.11	0.00	9.61	12.14	0.00	0.00	5.93	7.50	1.64	5.71	10.62	1.65	9.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	EL	IE	GRFinNUM	9.18	20.98	7.43	49.12	46.34	0.00	0.00	36.77	0.40	0.00	0.00	0.00	0.00	70.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		QA	GRFinQA	78.22	0.00	50.00	74.22	64.44	22.67	22.67	60.44	48.89	25.33	57.78	28.44	23.11	64.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TG	GRFNS-2023	25.50	16.95	37.72	16.90	13.61	0.24	0.21	9.71	5.60	11.20	6.48	14.45	3.56	34.46	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		TA	GRMultiFin	59.26	0.00	61.11	55.56	50.00	38.89	38.89	70.37	38.89	35.19	53.70	40.74	35.19	72.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	BI	TG	DOLFIN	92.29	90.13	86.26	89.17	92.13	35.92	35.92	92.29	91.80	69.24	91.60	71.81	66.57	91.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MU	QA	PolyFiQA-Easy	9.79	9.56	34.72	27.73	25.04	15.02	14.74	19.34	18.81	2.44	2.40	11.63	0.63	7.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		QA	PolyFiQA-Expert	5.31	4.85	30.35	20.60	18.56	13.83	16.01	18.17	16.35	6.38	0.71	8.80	0.00	9.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Average	40.98	14.59	30.61	39.50	42.20	19.34	20.41	38.07	33.06	26.83	31.24	24.40	28.95	35.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vision	EN	IE	EnglishOCR	21.38	0.00	0.00	12.39	0.00	10.70	11.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		QA	TableBench	66.70	0.00	0.00	32.30	0.00	28.60	60.90	0.00	74.90	0.00	0.00	0.00	0.00	0.00	59.30	57.30	0.00	0.00	0.00	0.00	0.00
	JA	IE	JapaneseOCR	21.63	0.00	0.00	24.52	0.00	25.82	26.72	0.00	21.59	0.00	0.00	0.00	9.70	6.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ES	IE	SpanishOCR	78.55	0.00	0.00	4.12	0.00	5.60	4.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	EL	IE	GreekOCR	41.86	0.00	0.00	42.51	0.00	23.69	30.60	0.00	0.00	0.00	0.00	0.00	8.25	6.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Average	46.02	0.00	0.00	23.17	0.00	18.88	26.80	0.00	19.30	0.00	0.00	0.00	0.00	0.00	15.45	14.14	0.00	0.00	0.00	0.00	0.00
Audio	EN	TG	MDRM-test	95.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	96.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	97.86	96.03	95.32	48.48	49.17
		TG	FinAudioSum	6.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.30	0.00	4.80	0.00	0.00
			Average	51.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	48.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	51.58	48.02	50.06	24.24	24.59
			Modality-Balanced Average	46.01	4.86	10.20	20.89	14.07	12.74	15.74	12.69	33.53	8.94	10.41	8.13	9.65	11.84	5.15	4.71	17.19	16.01	16.69	8.08	8.20

Table 3: Standardized performance of evaluated LLMs on the MULTIFINBEN. The final row presents the modality-balanced average, computed as the mean of text, vision, and audio averages.

substantially lower text-task performance (12.74% and 15.74%). These trends highlight the difficulty of preserving text-specific optimization when expanding to multimodal capabilities. In contrast, multimodal models decisively outperform unimodal baselines in vision and audio tasks, i.e, GPT-4o achieves 46.02% in vision and 51.04% in audio, surpassing modality-specific counterparts. The asymmetric performance reveals that while text tasks benefit from data maturity, vision and audio gain more from multimodal integration, making unified models essential for complex financial real-world applications.

Finding 3 for RQ3: Scaling monolingual ability does not yield multilingual reasoning. Figure 4b illustrates the stratified performance across linguistic settings. Although GPT-4o consistently ranks among the top performers across individual monolingual settings, it performs poorly on multilingual tasks (9.79% on *PolyFiQA-Easy* and 5.31% on *PolyFiQA-Expert*; see detailed error analysis in Appendix I). A similar pattern holds for Qwen2.5-Omni (18.81% and 16.35%), suggesting that strong monolingual proficiency does not necessarily translate into multilingual reasoning. In contrast, models with lower overall benchmark performance, such as Deepseek-V3 (34.72% and 30.35%) and Llama-3.1-70B (25.04% and 18.56%), achieve markedly better results on multilingual tasks. These findings indicate that multilingual reasoning constitutes a fundamentally different challenge from monolingual evaluation, one that cannot be addressed simply by scaling monolingual capacity but instead requires dedicated modeling for multilingual understanding and reasoning. Given the increasing demand for simultaneous multilingual comprehension in global financial contexts, MULTIFINBEN, particularly through its multilingual component (*PolyFiQA-Easy* and *PolyFiQA-Expert*, the first multilingual financial datasets), provides a critical stress test for this capability, capturing dimensions of multilingual reasoning that isolated monolingual evaluations fail to reveal.

Finding 4 for RQ4: Difficulty-aware benchmarking enables stratified, fine-grained, and dynamic evaluation. Figure 4c illustrates the stratified performance across difficulty levels. MULTIFINBEN introduces a three-tier structure (easy, medium, and hard) enabling layered evaluation of model capabilities. Performance converges at the hard tier, where all models struggle, whereas the easy and medium tiers yield clearer distinctions in capability. Except for Gemma-3-4B and Gemma-3-27B, which outperform on medium datasets due to



(a) Performance across modalities: Text, Vision, Audio.

(b) Performance across languages: EN, ZH, JA, ES, EL, BI, MU.

(c) Performance across difficulty levels: Easy, Medium, Hard.

Figure 4: Radar charts comparing model performance across (a) modalities, (b) languages, and (c) difficulty levels. The figures demonstrate diverse strengths and limitations of models in various dimensions.

their multimodal strengths, most models follow the monotonic trend easy>medium>hard. General-purpose models such as GPT-4o, Llama-4, and Qwen2.5-Omni demonstrate strong performance on easy and medium tiers but face substantial challenges on hard tasks. These hard-tier datasets (*PolyFiQA-Easy*, *PolyFiQA-Expert*, OCR-based, and other hard-tier datasets) require multilingual, multi-document reasoning and the joint interpretation of textual and structural information, posing a fundamental challenge for current models. By contrast, domain-specific models such as o3-mini, XuanYuan, and R1-Qwen32B-JA underperform on easy datasets, suggesting that heavy specialization may compromise generalization, a consideration for future training paradigms. Through this difficulty-aware design, MULTIFINBEN enables stratified and interpretable evaluation, revealing nuanced model strengths and weaknesses across complexity levels. As model capabilities advance, these tiers may shift dynamically, medium datasets may become easy, allowing the benchmark to evolve alongside model progress and real-world financial reasoning demands.

4 Conclusion

We present MULTIFINBEN, the first unified multilingual and multimodal benchmark for evaluating real-world financial applications under difficulty-aware settings. Our results show persistent performance gaps across languages, modalities, and difficulty levels, with even GPT-4o exhibiting substantial limitations in complex scenarios. The newly introduced datasets (*PolyFiQA* and *OCR*) constitute the core contribution of MULTIFINBEN by addressing critical gaps in multilingual reasoning and document understanding from scanned financial reports. Together, they reveal weaknesses in cross-lingual and cross-modal reasoning. MULTIFINBEN provides a comprehensive foundation for developing inclusive, multilingual, and multimodal financial LLMs capable of realistic expert-level reasoning.

Limitations

Despite its contributions, this study has several limitations. *(1) Limited availability of open datasets.* The financial domain lacks truly open-source datasets, and many publicly released resources have restrictive or ambiguous licenses, limiting redistribution and standardized benchmarking. As a result, our coverage may not fully reflect real-world task diversity. *(2) Modest scale of new datasets.* The newly introduced datasets remain relatively small compared to large-scale NLP benchmarks, due to the high cost and expertise required for financial annotation. Expanding both scale and diversity will improve robustness and coverage. *(3) Limited multilingual audio coverage.* Audio evaluation is currently restricted to English earnings calls, as multilingual data is constrained by legal and copyright barriers. While the U.S. market provides a representative testbed, future work will extend to multilingual settings. *(4) Incomplete model coverage.* We evaluate a representative but limited set of models due to computational and access constraints. Broader inclusion of emerging and region-specific models would strengthen future versions. *(5) No baseline-adjusted normalization.* Our scoring does not subtract random or majority baselines, which may inflate performance on simpler or imbalanced tasks. Future work will incorporate baseline-adjusted metrics for better comparability.

Potential Risks

Potential Positive Societal Impacts. This work contributes a dynamic, multilingual, and multimodal benchmark for evaluating large language models in the financial domain. By enabling more rigorous and transparent comparison of models across diverse languages, modalities, and task types, this benchmark: (1) Promotes equitable access to financial AI by supporting underrepresented languages and non-English financial datasets. (2) Encourages the development of safer, more reliable financial LLMs that can assist in decision-making, regulatory compliance, and financial literacy across global markets. (3) Provides tools for better understanding model capabilities and limitations in high-stakes, real-world applications such as financial question answering, document analysis, and OCR. By emphasizing open evaluation and releasing datasets and evaluation code, this work also fosters reproducibility and responsible innovation in financial AI.

Potential Negative Societal Impacts. Despite its contributions, this benchmark carries several potential risks: (1) Misuse or Overreliance: Improved performance on benchmark tasks may be mistaken for robust real-world competence. Overreliance on high-scoring models without thorough safety, fairness, or legal audits could lead to financial misinformation, biased decision-making, or consumer harm. (2) Representation Bias: While we aim for multilingual coverage, some languages and financial contexts remain underrepresented. This may amplify biases or exclude important financial narratives from less-resourced regions. (3) Data Privacy and Licensing Concerns: Although we carefully selected datasets for inclusion, ambiguities in licensing and potential data leakage from pretraining corpora may introduce ethical or legal risks if reused without careful scrutiny. Ongoing community engagement and transparent documentation are essential to mitigate these concerns as the benchmark evolves.

Ethical Considerations and Usage Disclaimer

The authors assume responsibility for the development and release of the MultiFinBen benchmark. All included datasets are publicly available and do not contain personal or sensitive information. We ensure that the dataset construction and release comply with relevant ethical standards and privacy guidelines. All datasets are released under the MIT License, and users are expected to comply with its terms.

The provided materials (including code, datasets, and documentation) are intended for academic and educational purposes only and do not constitute financial, legal, or investment advice. While efforts have been made to ensure accuracy, no warranties are provided regarding completeness or suitability for any use. The authors and their affiliated institutions disclaim liability for any damages arising from the use of these materials. Users are encouraged to seek professional advice for real-world decision-making. By accessing or utilizing the Material, users agree to indemnify, defend, and hold harmless the authors and their affiliated organizations from any claims, damages, or liabilities arising from such use.

Acknowledgments

The authors acknowledge The Fin AI community for its research support, feedback, and collaborative environment that contributed to this work. This research was supported by the NVIDIA Academic Grant Program using 32K A100 GPU-hours on Brev. This work has 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. Shengyuan Colin Lin, Keyi Wang, and Xiao-Yang Liu Yanglet acknowledge the support from Columbia's SIRS and STAR Program, The Tang Family Fund for Research Innovations in FinTech, Engineering, and Business Operations. Shengyuan Colin Lin and Xiao-Yang Liu Yanglet acknowledge the support from NSF IUCRC CRAFT center research grant (CRAFT Grant 22017) for this research. The opinions expressed in this publication do not necessarily represent the views of NSF IUCRC CRAFT. Xiao-Yang Liu Yanglet acknowledges the JPMorganChase Faculty Research Award. The views expressed are those of the authors and do not necessarily reflect those of JPMorganChase or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC and does not constitute investment advice, a recommendation, or a solicitation or offer in any jurisdiction.

References

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Accessed: 2025-05-09.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. FinTral: A family of GPT-4 level multimodal financial large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13064–13087, Bangkok, Thailand. Association for Computational Linguistics.

Jialun Cao, Yuk-Kit Chan, Zixuan Ling, Wenxuan Wang, Shuqing Li, Mingwei Liu, Chaozheng Wang, Boxi Yu, Pinjia He, Shuai Wang, and 1 others. 2025a. How should i build a benchmark? *arXiv preprint arXiv:2501.10711*.

Yupeng Cao, Haohang Li, Yangyang Yu, Shashidhar Reddy Javaji, Yueru He, Jimin Huang, Zining Zhu, Qianqian Xie, Xiao-Yang Liu, Koduvayur Subbalakshmi, Meikang Qiu, Sophia Ananiadou, and Jian-Yun Nie. 2025b. FinAudio: A benchmark for audio large language models in financial applications. *Preprint*, arXiv:2503.20990.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, and 2 others. 2025b. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1).

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021a. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and 1 others. 2021b. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

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. *Preprint*, arXiv:2210.03849.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535.

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).

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*.

Duanyu Feng, Yongfu Dai, Jimin Huang, Yifang Zhang, Qianqian Xie, Weiguang Han, Zhengyu Chen, Alejandro Lopez-Lira, and Hao Wang. 2024. Empowering many, biasing a few: Generalist credit scoring through large language models. *Preprint*, arXiv:2310.00566.

Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. 2024. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *Preprint*, arXiv:2411.03314.

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.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, and 5 others. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *Preprint*, arXiv:2411.04872.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Maria Cristina Hinojosa Lee, Johan Braet, and Johan Springael. 2024. Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores. *Applied Sciences*, 14(21):9863.

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Jiajia Huang, Haoran Zhu, Chao Xu, Tianming Zhan, Qianqian Xie, and Jimin Huang. 2024a. AuditWen: An open-source large language model for audit. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1351–1365, Taiyuan, China. Chinese Information Processing Society of China.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024b. AudioGPT: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

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*.

Ryosuke Ishigami. 2025. Deepseek-r1-distill-qwen-32b-japanese.

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

Igor Ivanov and Dmitrii Volkov. 2025. Resurrecting saturated llm benchmarks with adversarial encoding. *arXiv preprint arXiv:2502.06738*.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition (ICDAR)*.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023a. Multifin: A dataset for multilingual financial nlp. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023b. 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.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2023. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*.

Takahiro Kubo, Hiroki Nakayama, and Junya Kamura. 2018. chabsa: Aspect based sentiment analysis dataset in japanese.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yang Lei, Jiangtong Li, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv preprint arXiv:2311.05812*.

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.

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.

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*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023b. FinGPT: Democratizing internet-scale data for financial large language models. *Workshop on Instruction Tuning and Instruction Following, NeurIPS*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, and Yike Guo. 2025. FinMME: Benchmark dataset for financial multi-modal reasoning evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29465–29489, Vienna, Austria. Association for Computational Linguistics.

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

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and 1 others. 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*, pages 10893–10906.

Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. 2025. Dolfin–document-level financial test set for machine translation. *arXiv preprint arXiv:2502.03053*.

Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. 2025. Dolfin – document-level financial test set for machine translation. *Preprint*, arXiv:2502.03053.

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.

NVIDIA Research. 2024. nvingest: High-performance multimodal table extraction in the wild. `https://github.com/NVIDIA/nv-ingest`. Accessed May 2025.

OpenAI. 2025. Openai o3-mini: Pushing the frontier of cost-effective reasoning. Accessed: 2025-05-09.

Ronghao Pan, José Antonio García-Díaz, Francisco Garcia-Sanchez, and Rafael Valencia-García. 2023. Evaluation of transformer models for financial targeted sentiment analysis in spanish. *PeerJ Computer Science*, 9:e1377.

Seungryong Park, Seung Shin, Byoungjip Lee, Sangdoo Lee, Junhwa Lee, and In So Kweon. 2019. Cord: A consolidated receipt dataset for post-ocr parsing. In *Document Intelligence Workshop at NeurIPS*.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. 2008. An empirical analysis of word error rate and keyword error rate. In *Interspeech*, volume 2008, pages 2070–2073.

Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. 2025. Plutus: Benchmarking large language models in low-resource greek finance. *arXiv preprint arXiv:2502.18772*.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401.

Yu Qiu, Venkata C Duvvuri, Pratibha Yadavalli, and Neal Prasad. 2024. Evaluation of generative ai q&a chatbot chained to optical character recognition models for financial documents. In *Proceedings of the 2024 8th International Conference on Machine Learning and Soft Computing*, pages 101–110.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. 2024. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *arXiv preprint arXiv:2411.12990*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2023. AudioPALM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

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*.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference 2022*, pages 595–597.

William F Sharpe. 1998. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, 3:169–85.

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*, pages 589–601. Springer.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. 2025. Linguistic generalizability of test-time scaling in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14333–14368, Vienna, Austria. Association for Computational Linguistics.

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)*, pages 1691–1700. IEEE.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and 1 others. 2022a. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022b. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Duxiaoman DI Team. 2024. Du xiaoman-xuanyuan. Accessed: 2025-05-09.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, and 1 others. 2024a. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Matt White, Ibrahim Haddad, Cailean Osborne, Xiao-Yang Yanglet Liu, Ahmed Abdelmonsef, Sachin Varghese, and Arnaud Le Hors. 2024b. The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence. *arXiv preprint arXiv:2403.13784*.

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*.

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, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024b.

Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

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 preprint arXiv:2306.05443*.

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*.

Guojun Xiong, Zhiyang Deng, Keyi Wang, Yupeng Cao, Haohang Li, Yangyang Yu, Xueqing Peng, Mingquan Lin, Kaleb E Smith, Xiao-Yang Liu, Jimin Huang, Sophia Ananiadou, and Qianqian Xie. 2025. Flagtrader: Fusion llm-agent with gradient-based reinforcement learning for financial trading. *Preprint*, arXiv:2502.11433.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yiheng Xu, Tengchao Xu, Lei Cui, Guoxin Wang, Shaohan Huang, Furu Wei, and Ming Zhou. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

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. 2025. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Preprint*, arXiv:2407.06567.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

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. 2024a. 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, page 6236–6246, New York, NY, USA. Association for Computing Machinery.

Xiao Zhang, Ruoyu Xiang, Chenhan Yuan, Duanyu Feng, Weiguang Han, Alejandro Lopez-Lira, Xiao-Yang Liu, Meikang Qiu, Sophia Ananiadou, Min Peng, and 1 others. 2024b. 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*, pages 6236–6246.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4435–4439, New York, NY, USA. Association for Computing Machinery.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2023. Docmath-eval: Evaluating math reasoning capabilities of llms in understanding long and specialized documents. *arXiv preprint arXiv:2311.09805*.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. *Preprint*, arXiv:2501.12380.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Related Work

A.1 Trend of Benchmark Development

With the rapid development of large language models, many traditional benchmark datasets have become insufficiently challenging to effectively distinguish between models of varying capacities (Ivanov and Volkov, 2025). Even relatively small-scale models now achieve comparable performance to much larger ones on these benchmarks, despite notable disparities in their real-world capabilities (Hsieh et al., 2023). The FinBen benchmark (Xie et al., 2024b) also reveals that certain datasets are unable to clearly distinguish differences in model capabilities by conducting thorough experiments and evaluates LLMs across diverse financial tasks.

This performance saturation has motivated increasing efforts to design more difficult and nuanced evaluation tasks across various domains (He et al., 2024; Wang et al., 2024; White et al., 2024a). In the general domain, the BIG-Bench dataset (Suzgun et al., 2022a) compiles challenging tasks to stress-test LLMs. Subsequent work examines whether prompting techniques can enhance performance in such tasks. In the medical domain, Chen et al. (Chen et al., 2025a) developed QA tasks requiring clinical reasoning and explanation, providing a more rigorous evaluation of LLMs' capabilities in healthcare. In the financial domain, Zhao et al. (Zhao et al., 2023) introduced datasets of varying difficulty. For simpler tasks, small models perform comparably to larger ones. However, for complex benchmarks like Complong, involving multi-table reasoning and long document processing, smaller models struggle while larger models maintain a lead. This highlights the need for more discriminative benchmarks to expose nuanced differences in model capabilities.

Building on these observations and the proliferation of financial datasets, some recent work has proposed criteria to filter and prioritize representative datasets for benchmarking (Reuel et al., 2024; Cao et al., 2025a). These efforts assess benchmarks based on implementation, design, documentation, maintenance, and construction to determine their effectiveness. However, these approaches do not systematically evaluate the impact of dataset difficulty or establish objective criteria for filtering datasets based on complexity.

To address this gap, we propose a strategy that dynamically filters datasets and prioritizes those that are both challenging and effective in evaluating the performance of different LLMs. This approach enables more targeted benchmarking by focusing on datasets that can better reveal meaningful differences between models.

A.2 Multilingual Textual Financial Benchmarks

Despite the growing number of financial benchmarks developed to assess the capabilities of large language models in domain-specific tasks, most existing benchmarks remain monolingual, predominantly focusing on English or Chinese. Notable examples include PIXIU (Xie et al., 2023), CFBenchmark (Lei et al., 2023), FinanceBench (Islam et al., 2023), and BizBench (Koncel-Kedziorski et al., 2023), each offering tasks such as question answering, text classification, and numerical reasoning, all rooted in financial documents or news within a single language setting. Some efforts have extended financial benchmarking to additional languages. For instance, (Peng et al., 2025) and (Hirano, 2024) introduced benchmarks that encompass Greek and Japanese, broadening the linguistic focus beyond English and Chinese. Although works like (Xie et al., 2024b) and (Zhang et al., 2024b) expand beyond English by including tasks in both English and Spanish, they treat the two languages as separate tasks rather than integrating them into a unified multilingual evaluation. This approach limits the ability to assess cross-lingual generalization within financial contexts.

Some recent studies have attempted to expand financial benchmarks by incorporating bilingual tasks. For example, DOLFIN (Nakhlé et al., 2025) includes English-Chinese financial QA and reasoning tasks. However, its focus remains limited to translation-based tasks, which do not fully reflect real-world multilingual applications. Recent work shows that even advanced reasoning strategies such as test-time scaling fail to generalize effectively across languages, highlighting the need for explicit cross-lingual evaluation (Son et al., 2025).

Currently, there is a lack of comprehensive financial benchmarks that simultaneously encompass monolingual, bilingual, and multilingual settings. Most existing resources either focus solely on monolingual tasks or limited bilingual tasks, without integrating them into a unified multilingual framework. This gap highlights the need for more inclusive and versatile evaluation resources that can assess model performance across diverse linguistic settings. Our work addresses this lim-

itation by proposing a benchmark that not only evaluates LLMs across a wider range of financial tasks, but also considers dataset value filtering and cross-lingual robustness, setting the stage for more comprehensive and discriminative benchmarking in the financial domain.

A.3 Multimodal Financial Benchmarks

Most financial benchmarks focus solely on text, with limited exploration of visual and audio data. Recent works like (Qiu et al., 2024) address financial document parsing through OCR, while FinAudio (Cao et al., 2025b) targets financial audio content such as earnings calls. However, a unified evaluation framework that integrates all three modalities remains absent. Our work introduces a benchmark that incorporates text, visual, and audio data, emphasizing two key directions: (1) dynamic dataset filtering to prioritize challenging content and (2) promoting multilingual and multimodal benchmarks to reflect real-world financial complexities.

For visual modality, existing financial datasets such as CORD (Park et al., 2019) and FUNSD (Jaume et al., 2019) focus on semi-structured documents but lack the complexity of financial statements, FinMME (Luo et al., 2025), MME-Finance (Gan et al., 2024), and FinTral (Bhatia et al., 2024) focus on image understanding rather than the complex structure of document-based financial information. Applied systems like nvIngest (NVIDIA Research, 2024) extract tabular data from PDFs but are not research benchmarks. Modern multimodal LLMs, such as LayoutLMv2 (Xu et al., 2021), effectively combine text, layout, and visual cues but focus mainly on entity extraction rather than complete document conversion. Our benchmark targets the end-to-end transformation of scanned documents into structured formats like HTML, aligning with practical financial analytics needs.

AudioLLMs such as AudioGPT (Huang et al., 2024b) and AudioPaLM (Rubenstein et al., 2023) excel in spoken content processing. Financial-specific benchmarks are scarce, with FinAudio (Cao et al., 2025b) being the first to focus on financial audio but lacking integration with text and visual data. Our benchmark unifies text, visual, and audio modalities, enabling comprehensive evaluation of multimodal financial reasoning with metrics that assess both accuracy and semantic consistency across modalities.

B Task Categories

Inspired by FinBen (Xie et al., 2024a), we organize candidate datasets under a unified taxonomy of seven core financial NLP tasks:

- *Information Extraction (IE)* focuses on converting unstructured financial text into structured outputs.
- *Textual Analysis (TA)* assesses a model's ability to interpret sentiment, topic, or tone in financial discourse.
- *Question Answering (QA)* evaluates comprehension of financial content through question answering.
- *Text Generation (TG)* focuses on producing coherent, informative, and factually accurate financial text.
- *Risk Management (RM)* targets detection or analysis of risk-related signals.
- *Forecasting (FO)* measures a model's ability to predict market trends or investor behavior.
- *Decision-Making (DM)* simulates complex financial decision processes.

C Evaluation Models

Access	Modality	Language	Domain	Target Language	Model	MOF Class
Closed	Multimodal	Multilingual	General	English	GPT-4o (Hurst et al., 2024)	Class III - Open Model - In progress (17%)
	Text	Multilingual	General	English	o3-mini (OpenAI, 2025)	Class III - Open Model - In progress (0%)
Open	Multimodal	Multilingual	General	English	meta-llama/Llama-4-Scout-17B-16E-Instruct (AI, 2025)	Class III - Open Model - In progress (0%)
			General	English	google/gemma-3-4b-it (Team, 2025)	Class III - Open Model - In progress (17%)
			General	English	google/gemma-3-27b-it (Team, 2025)	Class III - Open Model - In progress (17%)
			General	English	Qwen/Qwen2.5-Omni-7B (Xu et al., 2025)	Class III - Open Model - In progress (67%)
	Text	Multilingual	General	English	meta-llama/Llama-3.1-70B-Instruct (Dubey et al., 2024)	Class III - Open Model - In progress (0%)
			General	English	Deepseek-V3 (Liu et al., 2024)	Class III - Open Model - In progress (17%)
			General	English	Qwen/Qwen2.5-32B-Instruct (Yang et al., 2025)	Class III - Open Model - In progress (33%)
			General	Japanese	cyberagent/DeepSeek-R1-Distill-Qwen-32B-Japanese (Ishigami, 2025)	Class III - Open Model - In progress (50%)[1]
			Financial	Chinese	Duxiaoman-DI/Llama3.1-XuanYuan-FinX1-Preview (Team, 2024)	Class III - Open Model - In progress (0%)[1]
			Financial	Spanish	TheFinAI/FinMA-ES-Bilingual (Zhang et al., 2024a)	Class III - Open Model - In progress (50%)[1]
			Financial	Greek	TheFinAI/plutus-8B-instruct (Peng et al., 2025)	Class III - Open Model - In progress (17%)[1]
		English	Financial	English	TheFinAI/finma-7b-full (Xie et al., 2023)	Class III - Open Model - In progress (67%)[1]
	Vision	Multilingual	General	English	Deepseek-VL-7B-Chat (Lu et al., 2024)	Class III - Open Model - In progress (17%)
		English	General	English	llava-hf/llava-v1.6-vicuna-13b-hf (Liu et al., 2023a)	Class III - Open Model - Qualified & Class II - Open Tooling - In progress (80%)[1]
	Audio	Multilingual	General	English	Whisper-V3 (Radford et al., 2023)	Class III - Open Model - In progress (67%)
			General	English	Qwen2-Audio-7B (Chu et al., 2024)	Class III - Open Model - In progress (67%)
			General	English	Qwen2-Audio-7B-Instruct (Chu et al., 2024)	Class III - Open Model - In progress (67%)
			General	English	SALMONN-7B (Tang et al., 2024)	Class III - Open Model - In progress (50%)[1]
			General	English	SALMONN-13B (Tang et al., 2024)	Class III - Open Model - In progress (50%)[1]

[1] The model is fine-tuned from other organizations' models, and its MOF class is evaluated only on the fine-tuned portion.

Table 4: Overview of evaluated models with access type, modality, language scope, and MOF class.

D Structured Difficulty-Aware Benchmarking of English Datasets

Difficulty	Task	Dataset	GPT-4o	LLaMA3.1-70B-Instruct	Mean	Variance
Easy	IE	EMIR-NER	93.00	92.00	92.50	1.00
	IE	SC	88.00	87.00	87.50	1.00
	TA	MA	80.00	84.00	82.00	4.00
	TA	FPB	83.00	79.00	81.00	4.00
	TA	FiQA-SA	77.00	74.00	75.50	3.00
	TA	Headlines	78.00	60.00	69.00	18.00
	TA	MultiFin	66.00	69.00	67.50	3.00
	DM	GOOG	67.76	64.79	66.27	2.97
	TA	FinArg-ACC	67.00	65.00	66.00	2.00
	TA	FOMC	67.00	64.00	65.50	3.00
	QA	XBRL-Math	68.00	62.00	65.00	6.00
	DM	AMZN	62.22	59.01	60.62	3.22
	DM	NFLX	54.32	65.86	60.09	11.54
Medium	RM	polish	58.50	57.00	57.75	1.5
	TA	FinArg-ARC	60.00	55.00	57.50	5.00
	RM	LendingClub	57.00	56.50	56.75	0.50
	DM	MSFT	41.32	71.05	56.18	29.73
	RM	CCFraud	52.50	56.50	54.50	4.00
	RM	German	50.00	55.50	52.75	5.50
	RM	travelinsurance	54.50	50.00	52.25	4.50
	QA	LegalQA	57.00	47.00	52.00	10.00
	RM	taiwan	50.00	53.00	51.50	3.00
	TA	TSA	61.00	42.00	51.50	19.00
	RM	CCF	52.50	50.00	51.25	2.50
	FO	ACL18	50.00	51.50	50.75	1.50
	DM	COIN	45.18	56.25	50.71	11.07
	RM	ProtoSeguro	50.00	50.50	50.25	0.50
	FO	BigData22	48.50	50.00	49.25	1.50
	FO	CIKM18	47.50	51.00	49.25	3.50
	IE	FINER-ORD	78.00	18.00	48.00	60.00
	RM	Australian	50.00	42.50	46.25	7.50
	DM	TSLA	49.47	41.92	45.69	7.54
	QA	MOFQA	39.50	41.00	40.25	1.50
	TA	MLESG	36.00	44.00	40.00	8.00
	QA	CDMQA	37.00	39.00	38.00	2.00
	DM	AAPL	40.51	30.03	35.27	10.49
	DM	NIO	24.96	38.18	31.57	13.22
	TA	FinABB	31.00	25.00	28.00	6.00
	TG	XBRL-EXP	26.00	26.00	26.00	0.00
	DM	DIS	27.11	24.04	25.57	3.07
	QA	TATQA	0.00	44.00	22.00	44.00
	TG	EDTSUM	25.00	18.00	21.50	7.00
Hard	QA	XBRL-Dom	6.00	8.00	7.00	2.00
	IE	NER	6.00	5.00	5.50	1.00
	QA	FinQA	5.00	0.00	2.50	5.00
	IE	FinRED	3.00	0.00	1.50	3.00
	IE	CD	0.00	0.00	0.00	0.00
	IE	FNXL	0.00	0.00	0.00	0.00
	IE	FSRL	0.00	0.00	0.00	0.00
	TG	ECTSUM	0.00	0.00	0.00	0.00
	QA	ConvFinQA	0.00	0.00	0.00	0.00

Table 5: Structured difficulty-aware benchmarking of English datasets. Datasets are ranked by their mean standardized performance.

E Textual Monolingual Benchmarks

E.1 English

To comprehensively assess the understanding capabilities of LLMs in English regulatory and financial contexts, we collected 13 datasets for 7 tasks including information extraction (IE), question answering (QA), text analysis (TA), text generation (TG), risk management (RM), forecasting (FO), and decision-making (DM). All task data are sourced from diverse real-world scenarios, including the U.S. Securities and Exchange Commission (SEC), regulatory reporting, financial reporting, financial news, earnings call transcripts, and microblog.

For the IE task, it focuses on extracting structured financial information from unstructured text. We evaluate this capability using three datasets: SC (Mariko et al., 2020), FinRED (Sharma et al., 2022), and FINER-ORD (Shah et al., 2023). The **SC** (Mariko et al., 2020) dataset assesses the detection of causal relationships in financial news and SEC filings, using F1 (Sokolova and Lapalme, 2009) as the evaluation metric. The **FinRED** (Sharma et al., 2022) dataset targets relation extraction, identifying relationships such as "product/material produced" and "manufacturer" in financial news and earnings call transcripts, also evaluated using F1 (Sokolova and Lapalme, 2009). The **FINER-ORD** (Shah et al., 2023) dataset involves named entity recognition, extracting entities like organizations, locations, and persons from financial agreements and SEC filings, with performance measured by entity F1 (Derczynski, 2016). Overall, the IE task provides a foundation for evaluating the model's understanding of financial knowledge.

For the TA task, it encompasses tasks such as sentiment classification, topic detection, and event identification from financial texts. In this work, we focus on two representative financial datasets: Headlines (Sinha and Khandait, 2021) and TSA (Cortis et al., 2017). The **Headlines** (Sinha and Khandait, 2021) dataset targets the extraction of actionable signals—such as price movements—from financial news headlines, with performance measured by the average F1 score (avg F1). The **TSA** (Cortis et al., 2017) dataset centers on sentiment analysis, where the goal is to classify textual segments into positive, negative, or neutral sentiment categories, evaluated using the accuracy (Makridakis, 1993) metric. This task is designed to assess the capability of LLMs in performing fine-grained understanding and reasoning over domain-specific financial language.

For the QA task, it involves answering specific regulatory and financial questions by leveraging the inherent knowledge of LLMs. Both **FinQA** (Chen et al., 2021b) and **TATQA** (Zhu et al., 2021) assess models' numerical reasoning abilities using financial reports, tables, and contexts, whereas **XBRL-Math** (Chen et al., 2022) focus on equation inference. All are evaluated by accuracy (Makridakis, 1993).

For the TG task, it involves summarizing or explaining specific content within regulatory or financial contexts, evaluating models' ability to produce coherent and informative text. **EDTSUM** (Xie et al., 2023) targets the summarization of financial news articles into concise and informative summaries. In addition, we utilize the **ECTSUM** (Mukherjee et al., 2022) dataset, which focuses on summarizing earnings call transcripts. Model performance on these summarization tasks is evaluated using ROUGE-1 (Lin, 2004).

For the RM task, it focuses on identifying, extracting, and analyzing risk-related information, interpreting numerical data, and evaluating a model's ability to understand complex relationships. Specifically, we use the CCF (Feng et al., 2024) datasets to assess LLMs' ability in our paper. The **CCF** (Feng et al., 2024) dataset is suitable for detecting fraud, which identifies whether transactions are "fraudulent" or "non-fraudulent". This dataset is evaluated by Matthews Correlation Coefficient (MCC) (Matthews, 1975).

For the FO task, it evaluates a model's ability to predict future market trends and investor behavior based on patterns in financial data. We focus on stock movement prediction, aiming to forecast whether a stock's price will rise or fall using historical price data and social media signals from the **BigData22** (Soun et al., 2022) dataset. Model performance is assessed using MCC (Matthews, 1975).

For the DM task, it evaluates an LLM's ability to synthesize diverse financial information to formulate and execute trading strategies—an inherently challenging task even for human experts. We refer to **MSFT** (Yu et al., 2024) dataset, which consists of data over one year, simulating real-world trading scenarios using historical prices, news articles, and sentiment analysis. Model performance is assessed using Sharpe Ratio (SR) (Sharpe, 1998), providing a comprehensive evaluation of profitability, risk

management, and strategic decision-making.

E.2 Chinese

To address domain-specific challenges in Chinese financial and regulatory contexts, we include 4 datasets from 3 core task types including IE, TA, and QA. These tasks require nuanced understanding of financial contexts, audit-related concepts, legal references, and inter-entity relationships, and serve to assess models' capabilities in financial reasoning, governance interpretation, and terminology grounding:

In IE task, **RRE** (Huang et al., 2024a) is a relation extraction task that identifies the semantic relation between two audit entities in a sentence. Relations span 8 predefined categories including *audit issue, audit item, audit basis, audit method, audit institution, audit outcome, audited entity, and related sector or domain*. This task tests a model's ability to infer structured relationships in financial text.

For TA task, **AIE** (Huang et al., 2024a) involves categorizing audit-target entities into 7 regulatory domains (*fiscal audit, infrastructure audit, customs audit, financial audit, social security audit, tax audit, and environmental audit*). This task evaluates regulatory disambiguation within Chinese financial texts. **LNE** (Huang et al., 2024a) is a legal grounding task where models must align a given audit scenario with the most relevant legal or regulatory category (*financial regulations, fiscal law, personal income tax, financial supervision, labor and employment, listed companies, social security, industry regulation, value-added tax, tax administration, asset evaluation law, resource tax, integrated governance, and comprehensive tax policy*). It requires both regulatory reasoning and knowledge of Chinese legal terminology.

In QA task, **FinanceIQ** (Zhang and Yang, 2023) is a multi-choice classification task where models need to answer Chinese-language financial domain-specific questions. The task evaluates an LLM's understanding of domain-specific vocabulary, foundational financial knowledge, and contextual reasoning in Chinese.

All datasets are evaluated using classification accuracy (Makridakis, 1993).

E.3 Japanese

To capture the unique challenges posed by Japanese in financial contexts, we only find one public TA dataset from a Japanese Financial Benchmark (Hirano, 2024) specifically designed to evaluate Japanese financial language understanding.

For TA task, we adopt the **chabsa** dataset (chABSA: Aspect-Based Sentiment Analysis dataset in Japanese) (Kubo et al., 2018), a binary classification task evaluates the sentiment polarity (positive or negative) of financial terms within securities reports. It is derived from publicly disclosed annual securities filings in Japan. The dataset includes annotated sentiment terms and is scored using the macro F1 (Hinojosa Lee et al., 2024) metric, excluding neutral terms for evaluation consistency.

E.4 Spanish

To comprehensively evaluate the understanding capabilities of LLMs in Spanish regulatory and financial contexts, we included four domain-specific datasets from FLARE-ES (Zhang et al., 2024a) across three tasks: TA, QA, and TG.

For TA task, text classification and sentiment analysis were included. In text classification task, the **MultiFin** (Jørgensen et al., 2023a) dataset evaluates the model's ability to categorize Spanish financial texts. This dataset, which centers on Spanish headlines, contains 2,066 articles spanning six key financial categories: *Business & Management, Finance, Government & Controls, Industry, Tax & Accounting, and Technology*. The task challenges the model to accurately assign each headline to its respective sector, thereby testing its linguistic adaptability and domain-specific knowledge in Spanish. In sentiment analysis task, the **TSA** (Pan et al., 2023) dataset includes 3,892 entries from financial news and tweets, annotated to classify sentiments as *positive, negative, or neutral*—capturing a nuanced spectrum of market emotions in Spanish.

For QA task, the **EFPA** [8] dataset, derived from questions used in financial examinations by official examiner associations, present unique challenges. The EFPA dataset broadens this challenge with 228 questions featuring four answer choices each, covering a broader spectrum of financial topics including economic principles, fundamental financial concepts, and intricate calculations associated with financial products. This dataset is evaluated using accuracy (Makridakis, 1993).

For TG task, the **FNS-2023** (Zavitsanos et al., 2023) dataset, consisting of 232 Spanish annual reports from diverse financial companies. It aims to condense voluminous Spanish financial docu-

[8] https://efpa-eu.org/

ments into succinct, informative abstracts, enhancing the accessibility and usability of financial information. Rich in detailed financial data and narratives, these reports pose a distinct challenge: distilling the essence of each document into a summary that captures the key information while preserving the factual integrity and coherence of the original text. This dataset is evaluated using ROUGE-1 (Lin, 2004).

E.5 Greek

We adopt four core financial NLP datasets from Plutus-ben (Peng et al., 2025), the first Greek financial benchmark. These datasets span from IE, TA, QA, and TG tasks.

For IE task, only GRFinNUM (Peng et al., 2025) is included. **GRFinNUM** (Peng et al., 2025) targets fine-grained classification of numerals into *MONETARY, PERCENTAGE, TEMPORAL, QUANTITY, and OTHERS* categories, which captures the nuances of numerical semantics in long-form Greek financial texts. This dataset is evaluated using entity F1 (Derczynski, 2016).

For TA task, **GRMultiFin** (Jørgensen et al., 2023b) focuses on classifying concise Greek financial headlines into one of six predefined thematic categories: *Business & Management, Finance, Government & Controls, Tax & Accounting, Technology, and Industry*, emphasizing the need for lexical disambiguation and contextual inference due to the brevity and ambiguity typical of headlines. This dataset is evaluated using accuracy (Makridakis, 1993).

For QA task, **GRFinQA** (Peng et al., 2025) evaluates models' performance on inferring the correct answer using provided financial text under a multiple-choice format, assessing a model's ability to comprehend and reason over finance-related questions drawn from Greek academic sources. This dataset is evaluated using accuracy (Makridakis, 1993).

For TG task, **GRFNS-2023** (Zavitsanos et al., 2023) tests models' ability of generating fluent, coherent, and factually accurate summaries from complex narrative sections of Greek financial reports, testing a model's capacity for paraphrasing and information compression. This dataset is evaluated using ROUGE-1 (Lin, 2004).

F Audio Benchmark

The Automatic Speech Recognition (ASR) task evaluates AudioLLMs' accuracy in transcribing financial audio clips, directly impacting applications such as financial voice assistants. In this task, each input is represented as (A, Q, R), where A is an audio clip, Q is the prompt instruction, and R is the reference transcript. The transcribed text T is generated as $T = \textit{AudioLLM}(A, Q)$. The Word Error Rate (WER) (Park et al., 2008) is computed as $\text{WER} = \frac{S+D+I}{N}$, where S, D, and I represent the number of substitutions, deletions, and insertions, respectively, and N is the total number of words in the reference ($N = S + D + C$, with C indicating correct words). A lower WER indicates better ASR performance.

The evaluation dataset in FinAudio (Cao et al., 2025b) is developed from two primary data sources: 1) existing open-source financial audio data originally created for non-LLM evaluation purposes. 2) novel datasets introduced in this work. Table 6 summarizes the key statistics.

- Short financial audio clip dataset: the MDRM (Qin and Yang, 2019) dataset derived from earnings conference call recordings. The original audio data was segmented at the sentence level and aligned with corresponding transcripts. This dataset is initially split into training and test subsets. For our paper, we utilize only the test sets for evaluation, which includes 22,208 audio clips totaling 87 hours.

- FinAudioSum: *FinAudioSum* (Cao et al., 2025b) was created based on the ECTSum dataset (Mukherjee et al., 2022), originally designed for earnings call summarization using textual data. ECTSum comprises 2,425 earnings transcripts paired with expert-generated, telegram-style summaries. We obtain corresponding audio recordings for the ECTSum test set from earningcast[9]. Overlapping recordings with Earnings-21 and Earnings-22 (spanning 2019–2022) are removed. The final *FinAudioSum* (Cao et al., 2025b) dataset includes 64 recordings totaling 55 hours.

[9] https://earningscast.com/

Dataset Name	Type	#Samples	# Hours	Task	Metrics
MDRM-test (Qin and Yang, 2019)	Short Clips	22,208	87	short financial clip ASR	WER
FinAudioSum (Cao et al., 2025b)	Long Audio	64	55	long financial audio Summarization	ROUGE-L

Table 6: Statistics of the datasets in the FinAudio benchmark.

G Optical Character Recognition (OCR)

G.1 OCR Statistics

Dataset	# PDF Files	Storage (GB)	Size
EnglishOCR	1,039	1.70	7,961
JapaneseOCR	331	1.57	17,586
SpanishOCR	2,302	0.76	12,819
GreekOCR	100	0.16	6,533

Table 7: Statistics of OCR datasets.

G.2 OCR Additional Results

Model	EnglishOCR			JapaneseOCR			GreekOCR		
	BERT	Num	Edit	BERT	Num	Edit	BERT	Num	Edit
GPT-4o	78.29	60.89	7.08	87.84	69.14	43.25	88.96	90.66	55.94
Llama-4-Scout	78.51	57.49	6.98	89.11	90.37	48.27	88.25	91.31	46.38
Gemma-27B	80.86	79.50	3.56	88.32	81.02	40.66	88.98	77.92	37.73
LLaVA-1.6	71.35	85.65	4.61	74.53	45.61	9.09	73.66	66.74	10.90
Deepseek-VL	54.00	79.82	3.12	76.16	54.59	5.60	78.41	55.96	10.78

Table 8: Additional results on EnglishOCR, JapaneseOCR, and GreekOCR, evaluated using BERTScore (BERT), numeric-consistency checks (Num), and edit distance (Edit; lower is better).

G.3 OCR Instruction Data Conversion

Task Instruction for OCR Datasets

```
Convert this financial statement page
into semantically correct HTML.
Return HTML and nothing else.
```

H PolyFiQA-Easy and PolyFiQA-Expert

H.1 Data Construction

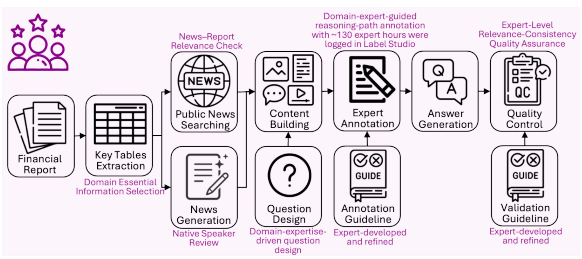


Figure 5: Expert-in-the-loop data construction of *PolyFiQA-Easy* and *PolyFiQA-Expert*.

Expert-in-the-Loop Data Construction. To ensure benchmark fidelity and domain rigor, we adopt an expert-in-the-loop pipeline (Figure 5). Three financial professionals (Appendix H.7) with expertise in economics, business, and accounting oversaw all phases of construction, including news selection, question authoring, guideline development, annotation, and quality control. News articles were meticulously screened for strong alignment with financial reports. Questions were crafted to anchor in real analytical tasks and span two difficulty tiers: easier questions in *PolyFiQA-Easy* and more complex ones in *PolyFiQA-Expert* (Appendix H.4). This structure supports fine-grained model assessment across reasoning levels (Appendix H.5). Rigorous tier-specific annotation guidelines (Appendix H.6) were refined through iterative pilot annotation rounds including tier-specific annotation rules and formatting protocols to promote inter-annotator consistency. In total, over 130 expert hours were logged in Label Studio (Appendix H.8), establishing a high quality, reproducible, auditable, and streamlined workflow aligned with best practices in benchmark creation. The raw datasets were further converted into structured instruction datasets with task-specific prompts thoughtfully crafted by financial professionals (Appendix H.11).

Quality Validation. Evaluating the quality of free-text generation datasets remains a persistent challenge in both QA and TG. To ensure annotation reliability, we adopt a structured scoring framework inspired by prior summarization benchmarks (Takeshita et al., 2024; Gliwa et al., 2019), evaluating each instance along two key dimensions: **Relevance** (scored 1–4) captures whether the response includes the key information required to answer the question. **Consistency** (scored 1–3) measures factual accuracy, especially numerical values. Each question is initially annotated by one expert and independently scored by two additional reviewers using detailed, pilot-refined validation guidelines (Appendix H.9). Only responses with cumulative scores above 5 are retained in the final dataset. To further validate scoring reliability, we report inter-

annotator agreement as normalized difference percentage across dimensions. *PolyFiQA-Easy* and *PolyFiQA-Expert* achieved average inter-annotator agreements of 89.38% and 91.21%, demonstrating the benchmark's high quality and scoring consistency.

H.2 News Data Sources

We collected news data from the following resources:

- **English:** Articles were primarily obtained from prior work (Yu et al., 2024); additional samples were generated by the authors.
- **Chinese:** Articles were collected from public websites including https://guba.eastmoney.com/, https://caifuhao.eastmoney.com, https://www.tobaccochina.com, https://www.jiemian.com, and https://xueqiu.com.
- **Japanese:** Articles were primarily collected from https://news.mynavi.jp/, https://www3.nhk.or.jp, https://jp.investing.com, and https://sustainablejapan.jp; additional samples were generated by the authors.
- **Spanish:** Articles were primarily collected from https://www.diariolibre.com, https://www.abc.es, https://es.marketscreener.com, https://www.infobae.com/, and https://es.investing.com; additional samples were generated by the authors.
- **Greek:** Since no publicly available Greek news sources were identified, all articles were generated by the authors.

H.3 News Generation

To support high-quality, multilingual financial news creation in our benchmark, we designed a semi-automated pipeline that integrates human curation with LLM assistance. The pipeline consists of three main stages:

Step 1: Extracting key financial highlights. For each target company and reference date, we collect relevant financial news and the corresponding SEC filings (10-Q or 10-K). From these sources, we distill 2–3 bullet points summarizing key developments, with emphasis on material indicators such as revenue performance, profit margins, and liquidity or cash flow metrics.

Step 2: AI-assisted drafting. We prompt GPT-4o to generate a concise financial news article based on the extracted bullet points. The prompt enforces a neutral and professional journalistic tone, while discouraging unsupported statements:

> `You are a journalist, specializing in finance. I am giving you some basic points, based on which I want you to generate a financial news article. This article should not be very large. Use neutral tone and professional language. Do not generate fallacious statements. You can change the order of the bullet points.`

Step 3: Human post-editing and quality assurance. The AI-generated draft is subsequently reviewed by human annotators to improve readability, ensure adherence to journalistic standards, and remove any hallucinated or unverifiable content. This step guarantees that the final article is copyright-compliant and fully authored by the benchmark creators.

H.4 Question Design

The questions are categorized into two difficulty levels: **Easy** and **Expert**.

PolyFiQA-Easy:

- What trends can be observed in the company's revenue amount over the past few years? Please quote the piece(s) of news aligned with the finding(s), if any.
- How does the company's balance sheet reflect its financial health in terms of total amount of current assets and the ratio of total liability divided by total equity? Please provide only the final result(s); do not include any calculation steps. Please quote the piece(s) of news aligned with the finding(s), if any.
- Are there any significant changes or irregularities in the amount of operating, investing and financing cash flow? Please quote the piece(s) of news aligned with the finding(s), if any.

- What is the company's R&D ratio (R&D divided by revenue)? Please provide only the final result(s); do not include any calculation steps. Please quote the piece(s) of news aligned with the finding(s), if any.

PolyFiQA-Expert:

- Please list the top three focuses on revenue from news. Please quote all the relevant financial information from the financial statements supporting the finding(s), if any.
- How is the company allocating capital (e.g., investments, share repurchases, dividends) based on its performance and market outlook from the news? Please quote all the relevant financial information from the financial statements supporting the finding(s), if any.
- What is the company's strategy on maintaining profit margins from the news? Please quote all the relevant financial information from the financial statements supporting the finding(s), if any.
- What are the company's capital expenditures and their strategic significance from the news? Please quote all the relevant financial information from the financial statements supporting the finding(s), if any.

H.5 Reasoning Trajectory

PolyFiQA-Easy:

- **Q:** What trends can be observed in the company's revenue amount over the past few years? Please quote the piece(s) of news aligned with the finding(s), if any.

  Steps:

  1. Extract revenue amounts over the past few years.

- **Q:** How does the company's balance sheet reflect its financial health in terms of total amount of current assets and the ratio of total liability divided by total equity? Please provide only the final result(s); do not include any calculation steps. Please quote the piece(s) of news aligned with the finding(s), if any.

  Steps:

  1. Extract current assets, total liability, and total equity amounts.
  2. Calculate the ratio (total liability divided by total equity).

- **Q:** Are there any significant changes or irregularities in the amount of operating, investing and financing cash flow? Please quote the piece(s) of news aligned with the finding(s), if any.

  Steps:

  1. Extract all three cash flow components (operating, investing and financing cash flow amounts).

- **Q:** What is the company's R&D ratio (R&D divided by revenue)? Please provide only the final result(s); do not include any calculation steps. Please quote the piece(s) of news aligned with the finding(s), if any.

  Steps:

  1. Extract R&D and revenue amounts.
  2. Calculate R&D ratio (R&D amount divided by revenue amount).

PolyFiQA-Expert:

- **Q:** Please list the top three focuses on revenue from news. Please quote all the relevant financial information from the financial statements supporting the finding(s), if any.

  Steps:

  1. Extract revenue for all focuses.
  2. Rank the revenue amounts of different focuses from high to low.
  3. Select the top three.

- **Q:** How is the company allocating capital (e.g., investments, share repurchases, dividends) based on its performance and market outlook from the news? Please quote all the relevant financial information from the financial statements supporting the finding(s), if any.

  Steps:

  1. Extract capital allocation items (investments, dividends, etc).
  2. Extract market outlook information from news.
  3. Analyze capital amounts and market outlook information together.

- **Q:** What is the company's strategy on maintaining profit margins from the news? Please quote all the relevant financial information from the financial statements supporting the finding(s), if any.

  Steps:

  1. Extract net profit and revenue.
  2. Calculate profit margin (net profit divided by revenue amount).
  3. Extract strategy information.
  4. Analyze profit margin ratio and strategy information.

- **Q:** What are the company's capital expenditures and their strategic significance from the news? Please quote all the relevant financial information from the financial statements supporting the finding(s), if any.

  Steps:

  1. Extract capital expenditures.
  2. Extract strategic context.
  3. Analyze capital expenditures amounts and strategy information together.

H.6 Annotation Guideline

H.6.1 Task Description

The task is a question answering (QA) and text generation (TG) task based on both financial statements and financial news.

H.6.2 Annotation Rules

General rules

- Limit your response to 100 words or fewer.
- Write `None` if the question cannot be answered or if no relevant evidence is found.

Specific rules

(1) PolyFiQA-Easy:

- **Answer format:**
  - **Answer:** {Answer using the financial statement.}
  - **News Evidence:** {Verify your answer using quote(s) from financial news. Write `None` if no news evidence is available.}

(2) PolyFiQA-Expert:

- **Answer format:**
  - **Answer:** {Answer by summarizing the financial news.}
  - **Financial Statements Evidence:** {Verify your answer using quote(s) from the financial statements (include original amounts if relevant). Write `None` if no statement evidence is available.}

H.7 Annotator Demography

The construction of our MULTIFINBEN relies on the deep domain expertise of a team of highly qualified annotators with strong backgrounds in finance, economics, and data science. Their interdisciplinary training and multilingual fluency ensure accurate and contextually grounded annotations across both financial statements and cross-lingual news articles.

One annotator, currently working as a senior analyst at a major U.S. financial institution, holds a master's degree in Business Analytics from a leading Ivy League university and a bachelor's degree in Statistics and Economics. Their background includes research on LLMs, financial data analysis, and economics, enabling precise annotation of complex financial information. Fluent in Chinese and experienced in multilingual reasoning, they bring a high level of rigor to aligning financial data with multilingual news content.

Another annotator earned their master's degree in Financial Mathematics from a prominent U.S. institution. They are currently pursuing a second master's degree in Computer Science with a focus on machine learning at a major public research university in the United States. With over seven years of professional experience in strategic finance and consulting within the FinTech industry, they contribute practical expertise in corporate finance, along with a strong foundation in computational methods.

The team is further strengthened by an annotator currently serving as a Research Associate at a major international financial research institute in Tokyo. They hold a Ph.D. in Financial Economics from a leading Australian university, with research spanning mergers and acquisitions, machine learning, and firm behavior. Their background includes prior academic positions in both

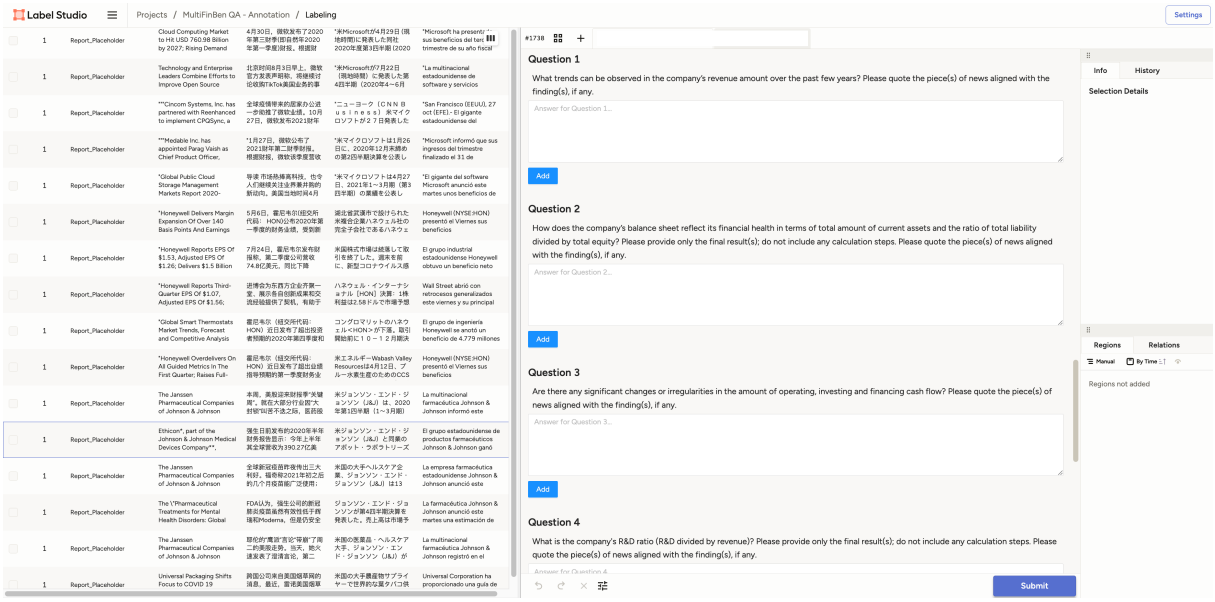


Figure 6: The Label Studio interface of the PolyFiQA-Easy and PolyFiQA-Expert annotation process.

Australia and China, multiple peer-reviewed publications in finance journals, and presentations at major international conferences. They also have experience as an editorial assistant, ad hoc referee, and research assistant, contributing a strong academic foundation to the annotation process.

Together, the team's combined financial acumen, analytical precision, and multilingual fluency enable the construction of a high-quality, cross-lingual financial QA dataset. Their expertise ensures the benchmark is linguistically, technically, and financially robust, setting a strong foundation for advancing multilingual LLM research in financial domain.

H.8 Annotation Process

The annotation workflow implemented in Label Studio is illustrated in Figure 6.

H.9 Validation Guideline

To ensure consistent annotation quality in PolyFiQA-Easy and PolyFiQA-Expert, we define the following validation guidelines:

(1) **Relevance**: Assesses whether the response includes the key information needed to answer the question.

- 4 – The response includes all key points or if no answer could be generated from the content.
- 3 – The response partially includes key points without irrelevant information.
- 2 – The response partially includes key points with minor irrelevant information.
- 1 – The response includes no relevant information.

(2) **Consistency (Factual accuracy)**: Evaluates the correctness of numerical values.

- 3 – All numbers are correct or if no numbers are needed.
- 2 – Some numbers are incorrect.
- 1 – All numbers are incorrect.

H.10 Annotation Agreement Metric

To account for the annotation agreement in our consistency rating task, we report **Gwet's AC1** (Gwet, 2014), a chance-corrected agreement coefficient that is more robust to prevalence and marginal distribution issues than Cohen's Kappa. The AC1 statistic is defined as:

$$\text{AC1} = \frac{P_o - P_e}{1 - P_e}, \tag{3}$$

where P_o is the observed agreement, and P_e is the chance agreement estimated by:

$$P_e = \sum_{i=1}^{k} p_i(1 - p_i), \tag{4}$$

with p_i being the proportion of annotations assigned to class i. This metric was selected due to the class imbalance present in our dataset.

H.11 Instruction Data Conversion

Task Instruction for PolyFiQA-Easy

```
You are tasked with answering the
user's question using the provided
context, which includes financial
statements (Income Statements,
Balance Sheets, and Cash Flow
Statements) and financial news
articles in multiple languages
(English, Chinese, Japanese,
Spanish, and Greek).
Please provide a detailed and well-
supported answer based on the
information available. Limit your
response to 100 words or fewer. If
you cannot answer the question or
if no relevant evidence is found,
write “None”.

Answer Format:
Answer: {Answer using the financial
statement.}
News Evidence: {Verify your answer
using quote(s) from the financial
news. Write “None” if no news
evidence is available.}

Context:
Financial Statements:
{financial_statements}

English News:
{english_news_content}

Chinese News:
{chinese_news_content}

Japanese News:
{japanese_news_content}

Spanish News:
{spanish_news_content}

Greek News:
{greek_news_content}

Question:
{user_question}

Answer:
```

Task Instruction for PolyFiQA-Expert

```
You are tasked with answering the
user's question using the provided
context, which includes financial
statements (Income Statements,
Balance Sheets, and Cash Flow
Statements) and financial news
articles in multiple languages
(English, Chinese, Japanese,
Spanish, and Greek).
Please provide a detailed and well-
supported answer based on the
information available. Limit your
response to 100 words or fewer. If
you cannot answer the question or
if no relevant evidence is found,
write “None”.

Answer Format:
Answer: {Answer by summarizing the
financial news.}
Financial Statements Evidence:
{Verify your answer using quote(s)
from the financial statements
(include original amounts if
relevant). Write “None” if no
statement evidence is available.}

Context:
Financial Statements:
{financial_statements}

English News:
{english_news_content}

Chinese News:
{chinese_news_content}

Japanese News:
{japanese_news_content}

Spanish News:
{spanish_news_content}

Greek News:
{greek_news_content}

Question:
{user_question}

Answer:
```

H.12 Additional Results

Model	PolyFiQA-Easy		PolyFiQA-Expert	
	BERT	Num*	BERT	Num*
GPT-4o	79.72	99.42	79.60	100
o3-mini	62.92	100	64.24	100
Deepseek-V3	84.65	74.06	86.10	83.85
Llama-4-Scout	82.99	87.21	84.86	21.81
Llama-3.1-70B	83.18	76.27	83.59	96.14
Gemma-4B	81.53	92.08	83.24	98.51
Gemma-27B	82.46	85.14	69.08	99.16
Qwen2.5-32B	81.89	85.25	83.11	99.76
Qwen2.5-Omni	82.69	61.38	83.06	93.83
FinMA-7B	79.58	59.36	81.08	97.58
XuanYuan	77.56	100	73.65	100
R1-Qwen32B-JA	78.63	100	77.31	100
FinMA-ES	20.62	82.56	0.88	100
Plutus-8B	79.50	71.63	80.42	89.47

Table 9: Additional results on PolyFiQA datasets, evaluated using BERTScore (BERT) and numeric-consistency checks (Num). * indicates that a score of 100 corresponds to outputs containing no numerical values.

I Error Analysis

I.1 *PolyFiQA-Easy* and *PolyFiQA-Expert*

Expert reviewers analyzed model outputs from GPT-4o and LLaMA3.1-70B-Instruct, identifying **information extraction failures** as the most common error source. Models often cited **incorrect financial fields** (e.g., using total assets instead of cash flow, or retained earnings instead of net profit margin), reflecting confusion over semantically similar numeric entries. In some cases, models returned "none" when they failed to locate relevant information.

These errors indicate persistent challenges in **retrieving and interpreting structured financial data**, especially in multilingual and cross-document settings. The significance of this task lies in its real-world complexity: financial professionals routinely integrate structured filings and multilingual news to form investment insights. *PolyFiQA-Easy* and *PolyFiQA-Expert*, as the first benchmark to capture this **multilingual, cross-document reasoning challenge**, provides a high-fidelity testbed for evaluating and advancing LLMs toward realistic financial QA capabilities.

I.2 *EnglishOCR*, *JapaneseOCR*, *SpanishOCR*, and *GreekOCR*

Expert review of GPT-4o's outputs identified five major error types:

1. **Omission of tables and charts**, especially in complex or multi-column layouts, leading to loss of key financial content;
2. **Hallucinations**, where models generated plausible but ungrounded content;
3. **Numeric misinterpretation**, including rounding errors, dropped digits, and misreading of formatting conventions;
4. **Edge-based omissions**, where headers, footers, or marginal content was ignored, possibly due to incomplete OCR bounding or attention bias;
5. **Skipping bracketed or vertical text**, which often contained essential qualifiers or context.

These diverse and persistent error types underscore the real-world difficulty and importance of our OCR benchmark. Accurate, layout-aware extraction of financial content is essential for downstream tasks such as auditing, reporting, and compliance. With four novel OCR datasets (*EnglishOCR*, *JapaneseOCR*, *SpanishOCR*, and *GreekOCR*), our benchmark supports **fine-grained diagnosis** of model limitations and advances the frontier of multimodal financial NLP.

J Model Openness Framework (MOF)

To systematically assess the openness and completeness of the models listed above, we adopt the Model Openness Framework (MOF) (White et al., 2024b), a three-tier ranked classification system designed for machine learning models. The MOF defines 17 components spanning the model development lifecycle and categorizes models into three hierarchical classes: "Class III - Open Model", "Class II - Open Tooling", and "Class I - Open Science", with each level subsuming the requirements of the preceding one. Class III represents the minimal threshold of openness. Although many models are advertised as "open", few satisfy even the Class III criteria. Most release only partial artifacts, such as model architecture, weights, model cards, or technical reports, without permissive open-source

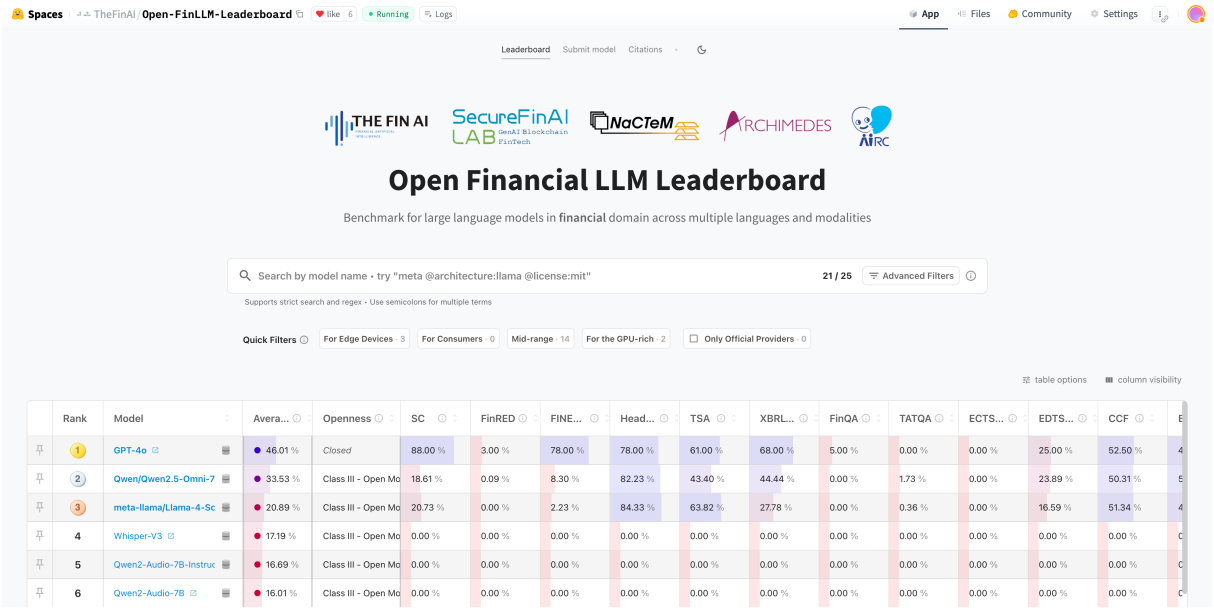


Figure 7: The MULTIFINBEN leaderboard.

licenses. Consequently, downstream use, modification, and redistribution are often legally constrained, posing potential legal and compliance risks. Among the models in our benchmark, most fall under Class III, with only llava-v1.6-vicuna-13b qualifying for Class II (Table 4, Appendix C).

K Open Multilingual and Multimodal Financial LLM Leaderboard

To promote transparency, openness, and community engagement, we have developed an interactive leaderboard for MULTIFINBEN (Figure 7). The leaderboard displays model performance across all benchmark tasks and includes Model Openness Framework (MOF) (White et al., 2024b) tags, which provide structured metadata about each model's openness, accessibility, and reproducibility. Users can filter, compare, and explore submissions based on performance metrics and openness levels. The leaderboard is continuously updated and publicly accessible to encourage standardized and responsible evaluation practices in financial AI. [10]

L Author Contribution

The author contributions are summarized below:

- **Science Leadership**: Kaleb E. Smith, Meikang Qiu, Arman Cohan, Xiao-Yang Liu, Jimin Huang, Guojun Xiong, Alejandro Lopez-Lira, Xi Chen, Junichi Tsujii, Jian-Yun Nie, Qianqian Xie, Sophia Ananiadou

- **Contributors**: Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, Shengyuan Lin, Keyi Wang, Shanshan Yang, Yilun Zhao, Zhiwei Liu, Peng Lu, Jerry Huang, Suyuchen Wang, Triantafillos Papadopoulos, Polydoros Giannouris, Efstathia Soufleri, Nuo Chen, Zhiyang Deng, Heming Fu, Yijia Zhao, Mingquan Lin

[10] https://huggingface.co/spaces/TheFinAI/Open-FinLLM-Leaderboard