FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models

Anonymous ACL submission

Abstract

We introduce *FinTral*, a suite of state-of-the-art multimodal large language models (LLMs) built upon the Mistral-7b model and tailored for financial analysis. FinTral integrates textual, numerical, tabular, and image data. We enhance **FinTral** with domain-specific pretraining, instruction fine-tuning, and RLAIF training by exploiting a large collection of textual and visual datasets we curate for this work. We also introduce an extensive benchmark featuring nine tasks and 25 datasets for evaluation, including hallucinations in the financial domain. Our FinTral model trained with direct preference optimization employing advanced **T**ools and **R**etrieval methods, dubbed *FinTral-DPO-T&R*, demonstrates an exceptional zero-shot performance. It outperforms ChatGPT-3.5 in all tasks and surpasses GPT-4 in five out of nine tasks, marking a significant advancement in AI-driven financial technology. We also demonstrate that FinTral has the potential to excel in real-time analysis and decision-making in diverse financial contexts.

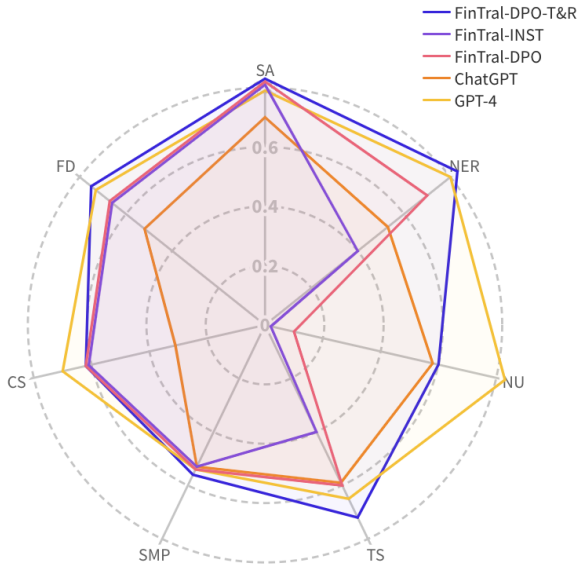


Figure 1: Comparative Performance Analysis on text-based tasks of Key Financial AI Models. We compare three variations of ***FinTral*** with ChatGPT (GPT-3.5) and GPT-4 across seven task clusters: Sentiment Analysis (SA), Named Entity Recognition (NER), Number Understanding (NU), Text Summarization (TS), Stock Movement Prediction (SMP), Credit Scoring (CS), and Firm Disclosure (FD).

1 Introduction

Natural Language Processing (NLP) plays a key role in financial document analysis, interpretation, and utilization. In recent years, a wide range of applications incorporating advances in NLP have emerged. These include sentiment analysis of financial news, event extraction from financial documents, and the generation and summarization of financial reports (Souma et al., 2019; Araci, 2019; Yang et al., 2018). These developments have uncovered the potential for unstructured data for data-driven financial decision-making and the transformation of financial documents into actionable insights and market intelligence. Applying NLP in finance, however, is challenging because financial documents often include dense numerical information and domain-specific jargon requiring advanced numerical processing and reasoning capabilities (Mik, 2017; Liu et al., 2023b). This means that financial NLP models need extensive domain knowledge before they can capture the nuanced implications of accounting and financial measures, economic indicators, and market trends. This is also compounded by the rapid pace of financial markets, where real-time analysis is crucial but challenging to achieve (Gupta, 2023; Yang et al., 2023b).

Similar to other domains, large language models (LLMs) are starting to disrupt financial document understanding (Chapman et al., 2022; La Quatra and Cagliero, 2020) but can also suffer from the same issues as transitional approaches. LLMs are also prone to hallucination, reducing their usability

in financial decision-making (Kang and Liu, 2023). Financial documents can also involve various types of visual content, which require models with multimodal abilities.

To meet these challenges, we introduce a groundbreaking LLM specialising in the financial domain. Our model, dubbed *FinTral*, is designed to overcome hurdles of the financial domain through a multimodal approach that integrates textual, numerical, tabular, and visual data processing for comprehensive document understanding. We train our model off Mistral-7b (Jiang et al., 2023) on a sizeable domain-specific dataset and instruction-tune it for the financial domain using extensive instruction data. We then carefully align it with GPT-4 generated responses leveraging the recently introduced direct policy optimization (DPO) method (Rafailov et al., 2023). In order to evaluate FinTral, we introduce an extensive benchmark of eight different tasks based on 25 different datasets. Our model outperforms all other models of comparable size and, in spite of its much smaller size, performs on par with GPT-4.

To summarize, we offer the following contributions: **(1)** We introduce FinTral a cutting-edge multimodal LLM specialized in financial data, and FinSet, an extensive financial LLM training and evaluation benchmark. FinSet is the largest financial evaluation benchmark and the only one that measures model hallucinations, encompassing nine tasks across 25 datasets. **(2)** FinTrals further instruction-finetuned and carefully aligned using the DPO objective, using AI feedback data, resulting in *FinTralDPO*. **(3)** We have also endowed FinTral with vision capabilities, extending it to *FinTralVL*, which employs the CLIP (Radford et al., 2021) vision encoder. For enhanced performance, we developed a version that utilizes **T**ools and **R**etrieval, *FinTralDPO-T&R*. **(4)** FinTralDPO demonstrates exceptional zero-shot capabilities, outperforming ChatGPT (OpenAI, 2023a) in *all* tasks. Moreover, our best model, FinTralDPO-T&R, surpasses GPT-4 (OpenAI, 2023b) in five of eight text-based tasks.

The rest of this paper is organized as follows: In Section 2, we review related work with a particular emphasis on financial LLMs, their applications and challenges. Section 3 outlines how we built our benchmark dataset: FinSet. We present our approach to model pretraining, instruction tuning, and prompting strategies, and subsequently introduce FinTral models in Section 4. In Section 5, we present our experiments and comprehensively analyse our models. We discuss our results in Section 6 and conclude in Section 7.

2 Related Works

NLP for finance Traditional NLP has been applied to various finance tasks, including named entity recognition, sentiment analysis, event extraction, financial report generation, and text summarization (Salinas Alvarado et al., 2015; Souma et al., 2019; Araci, 2019; Yang et al., 2018; Zheng et al., 2019; Chapman et al., 2022; La Quatra and Cagliero, 2020). However, traditional models face challenges in this domain due to complexity of financial language, scarcity of annotated data, limited inferential capabilities, and the need for real-time analysis. Adaptability of conventional NLP models is also limited, with such models often optimized for single-task functions (Mik, 2017; Mishra et al., 2021; Liu et al., 2023b).

Financial LLMs Advancements in financial models began with FinBERT (Araci, 2019). Recently, models like BloombergGPT (Wu et al., 2023), PIXIU (Xie et al., 2023), Instruct-FinGPT (Zhang et al., 2023a), and GPT-FinRE (Rajpoot and Parikh, 2023) are notable contributions. Other innovations include introduction of multimodal capabilities (FinVis-GPT (Wang et al., 2023b)), enhancement of investment strategies (GPT-InvestAR (Gupta, 2023), InvestLM (Yang et al., 2023b)), and efforts to address challenges such as economic sentiment analysis and hallucination in information extraction (Zhang et al., 2023b; Sarmah et al., 2023). FinLMEval (Guo et al., 2023) and DISC-FinLLM (Chen et al., 2023) focus on evaluation and model performance in monetary scenarios. Other work, such as Chu et al. (2023), emphasizes sophisticated data preprocessing for better handling of financial tasks. Appendix A provides a further discussion of the NLP and LLMs literature in finance.

3 FinSet

We develop comprehensive and diverse datasets to build FinTral. We first describe our raw datasets rich in domain-specific tokens, setting a solid foundation for model training, then our instruction fine-tuning and AI-driven feedback datasets. Subsequently, we present a multi-modal financial dataset to facilitate a nuanced approach to data interpretation. Finally, we introduce an extensive set of

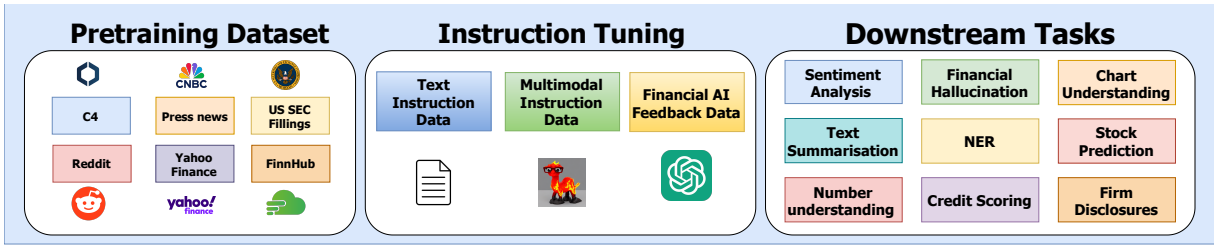


Figure 2: FinSET, a Financial Training and Evaluation Benchmark.

evaluation benchmark datasets tailored to test the model's performance across diverse financial tasks.

3.1 Pretraining Dataset

We introduce FinSet, a 20 billion token, high quality dataset we build for financial LLM training. FinSet is acquired based on a collection of large text corpora (2.9 billion documents, making 135 billion tokens; see Table 1) from which we extract finance-specific data that we then clean using a careful filtering pipeline. The datasets are described in Appendix B. Our cleaning pipeline is detailed in Appendix C. Our document cutoff date is August 1, 2023, which affords recent information to our models.

Dataset	Documents	Tokens	Deduplicated Tokens
C4	2.8B	124.0B	11.75B
News	51.5M	8.7B	5.65B
SEC	4.3M	3.1B	2.55B
Social Media	717.7K	8.2M	7.87M
Press	12.0K	3.1M	1.55M
Total	2.9B	135.9B	**20.0B**

Table 1: Details of our pretraining resources.

3.2 Financial Instruction Data

We assemble an extensive collection of instruction tuning datasets to enhance capabilities of our models. The datasets originate from various sources, notably including interactions with GPT-3.5 and GPT-4 for a diverse host of tasks. Again, we apply a deduplication and filtering pipeline (detailed in Appendix C) to exclude non-financial instructions, thereby focusing solely on financial reasoning. Table 2 shows our various data sources, along with the resultant (final) dataset.

3.3 Financial AI Feedback Data

Human feedback is valuable for aligning LLMs. Traditionally, this feedback is derived from human preferences as to the quality of LLM responses. In this work, we employ AI feedback through a refined version of the finance reasoning instruction dataset described in Section 3.2.

Dataset	Source	Instructions
FLUPE	ChanceFocus/FLUPE	123.0k
finance-alpaca	Gbharti/Finance-alpaca	68.91k
finest-finred	FinGPT/Hingpt-finred	32.67k
Math Instruct	TIGER-Lab/MathInstruct	26.2k
fin-llama-dataset	bavest/fin-llama-dataset	16.9k
llama-2-finance	AdiOO7/llama-2-finance	4.84k
Total instructions	-	272.6k
Total after deduplication	-	**226.3k**

Table 2: Instruction tuning datasets.

Along with the output generated by GPT-4 (OpenAI, 2023c), we generate responses using the FinMA-7B (Xie et al., 2023) and LLaMa-7B-chat (Touvron et al., 2023) models to each prompt. For a given prompt, the GPT-4 output is selected as the 'chosen' response while we select randomly one from FinMA-7B and LLaMa outputs as the 'rejected' response. Our AI feedback data includes a total of 43k samples, and we show an example of this data in Figure D.5.

3.4 Visual Financial Instruction Dataset

For aligning the vision language components in FinTral, we use LAION, CC, and SBU datasets from the Llava pretraining data (Liu et al., 2023a). We also use the ChartQA training set (Masry et al., 2022) for the same purpose. In addition, we follow the same approach by Wang et al. (2023b) to further expand our visual pretraining dataset. While Wang et al. (2023b) use Chinese data, we use the Fortune-500 companies stock price data, allowing us to create our own English dataset, dubbed *FinVis-PT*. We then use LLava Instruct data to improve the instruction understanding of our multimodal LLMs, creating our instruction tuning dataset *FinVis-IT*. While the FinVis-PT dataset includes stock market charts and asks simple questions about them, FinVis-IT is multi-turn and includes more complex charts and instructions. Our visual instruction datasets are described in Table 3.

Multimodal Training	Dataset	Source	Instructions
Alignment	LAION/CC/SBU	Liu et al. (2023a)	558k
	FinVis-PT	**Our Paper**	185k
	ChartQA	Masry et al. (2022)	20.9k
Multiturn	FinVis-IT	**Our Paper**	427k
	LLava 1.5	Liu et al. (2023a)	665k
Total			1.1M

Table 3: Visual financial instruction datasets. We generated FinVis using the same method from Wang et al. (2023b).

3.5 Downstream Evaluation Datasets

A diverse array of downstream datasets is crucial for effective LLM performance benchmarking. In this work, we develop an extensive benchmark using existing and new datasets to evaluate our models. Our benchmark covers the following tasks: (1) chart understanding (CU), (2) sentiment analysis (SA), (3) named entity recognition (NER), (4) number understanding (NU), (5) text summarization (TS), (6) stock movement prediction (SMP), (7) credit scoring (CS), (8) firm disclosure (FD), and (9) hallucination analysis (HI). Table 4 summarizes all the datasets used in our evaluation, each along with the corresponding evaluation metric employed. We also provide more details about the datasets in Appendix D.

4 Fintral

We use *Mistral-7B-v0.1* (Jiang et al., 2023) as our base model for further development, due to its strong performance and employment of a BPE tokenizer that segments numbers into single digits, which is suitable for numerical tasks.

Domain-Specific Pretraining We further pretrain Mistral-7B-v0.1 on our 20 billion token FinSet financial data described in Section 3. We perform pretraining with flash attention 2 (Dao, 2023). We employ a sequence length of up to 8k tokens, thus accommodating long financial documents. We use LoRA (Hu et al., 2021) for pretraining and train the model for one epoch with a learning rate of $2.5e^{-5}$. Pretraining takes 80 hours on four 40GB A100 GPUs.

Prompting for Financial LLMs We employ a prompting method suited for a financial LLM with multimodal capabilities. The model is assigned a memetic proxy (Reynolds and McDonell, 2021) as a financial expert signifying key expected behaviors, encouraged to think step by step, and that consider diverse inputs which may be texts, tables, or images. This is followed by a strategic retrieval of pertinent information, ensuring the model's focus aligns with the query's requirements. The model then engages with a task-based question, demanding an application of the model's financial expertise and analytical thinking. This structured approach is pivotal in eliciting focused answers from the model, especially in complex financial scenarios. The application of constraints further refines the model's output, leading to enhanced accuracy and context-appropriate responses. A visual representation of FinTral's prompting method is depicted in Figure 3.

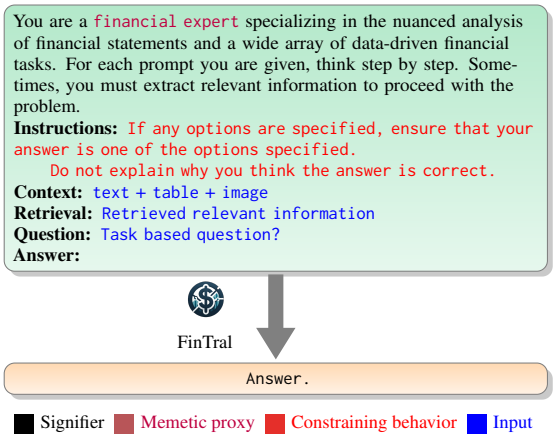


Figure 3: FinTral prompting method

Instruction Tuning We use our instruction tuning dataset described in section 3.2 to perform instruction finetuning on our pretrained model.[1] We adopt QLoRA to perform instruction finetuning using all the linear layers as target modules as this gives us a performance that is close to full fine-tuning (Dettmers et al., 2023).

Alignment with AI Feedback Large language models may fail to respond well to natural prompts even after instruction fine-tuning. To address this challenge, we use direct preference optimization (DPO) (Rafailov et al., 2023) which allows us to preferentially tune the model without the usage of a reward model. Tunstall et al. (2023) introduces a method to use LoRA to train LLMs using DPO objective. This is known as distilled direct preference optimization (dDPO). [2] We describe how we generate the binarized preference data for this process in Section 3.3.

Multimodal Instruction Tuning Once we teach

[1] We standardize all the datasets to have the same format of prompting, as explained earlier.

[2] We use the scripts provided by Tunstall et al. (2023) to train our dDPO model.

Data	Task	Instruction	Data Types	Modalities	Source	Metrics
ChartQA	chart understanding	2,500	general charts	text, images	Masry et al. (2022)	Accuracy
FinVQAv1		500	stock market charts		**Our paper**	
FinVQAv2		525	complex financial charts		**Our paper**	
Australian	credit scoring	690	credit records	table	Quinlan	Accuracy
German		1,000			Hofmann (1994)	
CS	firm disclosure	240	SEC filings	text	Cao et al. (2023)	Accuracy
FSR		3,931			Cao et al. (2020)	
ITR		1,196			**Our paper**	
FinTerms-MCQ	hallucination analysis	1,129	financial terms,Wikipedia	text	**Our Paper**	Accuracy
FinanceBench		150	financial documents	text,tables	Islam et al. (2023)	Human Evaluation
FinTerms-Gen		150	financial terms,Wikipedia	text	**Our Paper**	
ConvFinQA	numerical understanding	3,892	earnings reports	text, table	Chen et al. (2022)	Exact Match
FinQA		8,281			Chen et al. (2021)	
Finer-Ord	named entity recognition	1,080	news articles	text	Shah et al. (2023b)	Entity-F1
FiNER		13,660	financial agreements		Salinas Alvarado et al. (2015)	
ACL18	stock movement prediction	27,053	tweets, historical prices	text, time series	Xu and Cohen (2018)	Accuracy
BigData22		7,164			Soun et al. (2022)	
CIKM18		4,967			Wu et al. (2018)	
FiQA-SA	sentiment analysis	11,730	news headlines, tweets	text	Maia et al. (2018)	Accuracy
FOMC		496	FOMC hawkish-dovish		Shah et al. (2023a)	
FPB		48,450	news		Malo et al. (2013)	
Headline		11,412	news headlines		Sinha and Khandait (2020)	
ECTSUM	text summarization	495	earning call transcript	text	Mukherjee et al. (2022)	Rouge-score
EDTSUM		2,000	news articles		Zhou et al. (2021)	
Risk Eval		3,000	SEC articles		Loukas et al. (2021)	

Table 4: The details of the downstream data. FinTerms-Gen is extracted from Investopedia (2024) and FinTerms-MCQ is generated using code from Ghosh et al. (2022)

our model to handle various financial queries, we also empower it with visual understanding. This is done using the architecture suggested by Liu et al. (2023a). Specifically, we add an `<image>` token to our prompt and replace the `<image>` token with its image embedding after tokenization. We use a CLIP model (Radford et al., 2021) as our vision encoder and a 2-layer MLP visual abstractor, allowing us to convert image inputs into text embeddings fed to the LLM.

Tool Usage In addressing the inherent challenges faced by LLMs in dealing with quantitative tasks, we integrate tools (Schick et al., 2023) to our model. These tools enable the LLM to offload mathematically intensive tasks to a more suitable computational environment. For instance, functions such as `Add()`, `Subtract()`, and `Multiply()` are used by model to generate outputs in a structured format interpretable as Python function calls, thereby enhancing model accuracy in financial applications.

Retrieval Augmented Generation (RAG) As shown in Zhang et al. (2023b) for financial sentiment analysis, using retrieval augmented generation (RAG) can significantly boost performance. To better facilitate our tool usage and, in some cases, text extraction from complex data, we deploy a RAG system employing the BGE (Xiao et al., 2023) models, which are SoTA for document retrieval. This is useful for LLMs since users commonly ask out-of-domain questions. We use 30,000 financial documents derived from multiple sources covering January 1, 2022 to September 30, 2023. We use the chain of retrieval, as shown in Figures D.3 and its example is provided in Figure D.4.

5 Experiments

We conducted multiple experiments to illustrate the efficacy of the methods described in section 4. We evaluated our model on the downstream tasks described in section 3.5. Symbols in the following tables indicate the types of models: ♣, ♠, ♢, ♡, ★m and, ■ represent the pre-trained model, the fine-tuned model, the instruction fine-tuned model, the RL-Tuned Models, tools, and retrieval, respectively. We then performed a hallucination index accuracy check to assess how well our model mitigates one of the biggest challenges for LLMs.

We introduce three versions of our model. Firstly, FinTral-INST is our instruction-fine-tuned model obtained by fine-tuning our pre-trained model. Note that we do not assess the performance of the pre-trained model as it serves as an intermediate step to the instruction fine-tuning model. Secondly, We introduce FinTral-DPO, which has been further trained based on FinTral-INST utilizing reinforcement learning using AI feedback with the dDPO objective. Then, we introduce our FinTral-DPO-T&R, which combines the FinTral-DPO with tools and retrieval.

Model	Type	SA	NER	NU	TS	SMP	CS	FD	Average
FinMA-7B-trade	♠	0.20	0.00	0.00	0.08	0.46	0.39	0.00	0.16
Llama-2-7b-hf	♣	0.26	0.00	0.00	0.00	0.48	0.50	0.09	0.19
Mistral-7B-v0.1	♣	0.25	0.00	0.00	0.05	0.49	0.52	0.09	0.20
Vicuna-7B	♢	0.54	0.01	0.00	0.20	0.46	0.39	0.00	0.23
Mistral-7B-Instruct-v0.1	♢	0.49	0.00	0.00	0.30	0.49	0.48	0.29	0.29
Llama-2-13b-chat-hf	♡	0.58	0.02	0.00	0.30	0.50	0.52	0.31	0.32
FinMA-7B	♠	0.72	0.38	0.16	0.29	0.46	0.29	0.00	0.33
Llama-2-7b-chat-hf	♡	0.54	0.07	0.00	0.31	0.52	0.56	0.32	0.33
FinMA-7B-full	♠	0.78	0.35	0.12	0.35	0.51	0.29	0.30	0.38
FinTral-INST	♢	0.81	0.40	0.02	0.40	0.53	0.61	0.66	0.49
ChatGPT (gpt-3.5-turbo)	♡	0.70	0.53	0.58	0.59	0.53	0.31	0.52	0.53
FinTral-DPO	♡	**0.82**	0.70	0.15	0.60	**0.54**	0.62	0.67	0.59
GPT-4 (gpt-4-0613)	♡	0.79	**0.80**	**0.63**	**0.65**	**0.54**	**0.70**	**0.73**	**0.69**

Table 5: Comparative analysis of LLMs on diverse tasks. Models in bold are introduced in this paper. This analysis includes **SA:** Sentiment Analysis, **NER:** Named Entity Recognition, **NU:** Number Understanding, **TS:** Text Summarization, **SMP:** Stock Movement Prediction, **CS:** Credit Scoring, and **FD:** Firm Disclosure.

We also compare performance of our models to nine other baselines LLMs. These are LLama-2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), three versions of FinMA (Xie et al., 2023), Vicuna (Chiang et al., 2023), ChatGPT (OpenAI, 2023a), GPT-4 (OpenAI, 2023c).

5.1 Instruction Tuning and Model Alignment

As seen from Table 5, our instruction fine-tuned model FinTral-INST outperforms all pretrained and fine-tuned open-source models with an average score of 0.49. One of the causes of concern here is the tasks that require a specific format as the output, like the numerical understanding and NER tasks. We see that in some instances, the model struggles to follow instructions and often deviates from what the task asks for.

Furthermore, models that have undergone reinforcement learning with AI feedback (RLAIF), like FinTral-DPO, ChatGPT, and GPT-4, show even more marked improvements. Adding RLAIF dramatically boosts performance to the average score of 0.59, resulting in FinTral-DPO outperforming ChatGPT.

GPT-4, in particular, stands out with the highest average score, indicating its robust performance across a diverse set of tasks. Its high NER, NU, and FD scores suggest exceptional capabilities in understanding complex text, identifying specific entities, and interpreting numerical data.

5.2 Retrieval and Tools Usage

As detailed in section 4, the use of retrieval and tools plays a pivotal role in enhancing the capabilities of our model, FinTral-DPO-T&R, similar to their impact on GPT-4. Integrating these features into these models allows the models to access a broader range of information and apply more specialized processing techniques, leading to significant improvements in performance across various tasks. In the case of FinTral-DPO-T&R, combining the FinTral-DPO model with retrieval and tool capabilities has proven particularly effective. The FinTral-DPO model's ability to follow instruction prompts accurately enables seamless integration with external tools and retrieval data. The performance of GPT-4-Turbo, with its latest update incorporating tools and retrieval, is also noteworthy.

In 5 downstream tasks, FinTral-DPO-T&R outperformed GPT-4, while GPT-4 surpassed FinTral-DPO-T&R in two downstream tasks. Since GPT-4 has done exceptionally well in those two tasks, its average performance is slightly better than FinTral-DPO-T&R (0.72 vs. 0.70, as shown in table 6). The edge that FinTral-DPO-T&R and GPT-4 have over other models is a testament to the potential of combining sophisticated AI models with additional data and tool integration for more refined and accurate outputs.

5.3 Multimodal Evaluation

To evaluate our financial multimodal model, we use ChartQA and our FinVis datasets. We compare various state-of-the-art multimodal large language models (MLLMs) such as GPT-4V (OpenAI, 2023b), Gemini-Pro (Team et al., 2023), Qwen-VL-Plus (Bai et al., 2023), LLaVa-NEXT (Liu et al., 2024), and our FinTral-VL model which

Model	Type	SA	NER	NU	TS	SMP	CS	FD	Average
Mistral-7B-Instruct-v0.1	◇	0.49	0.00	0.00	0.30	0.49	0.48	0.29	0.29
Llama-2-7b-chat-hf	♡ + ★ + ■	0.54	0.07	0.00	0.31	0.52	0.56	0.32	0.33
FinTral-INST	◇	0.81	0.40	0.02	0.40	0.53	0.61	0.66	0.49
ChatGPT (gpt-3.5-turbo-1106)	♡	0.70	0.53	0.58	0.59	0.53	0.31	0.52	0.53
FinTral-DPO	♡	0.82	0.70	0.15	0.60	0.54	0.62	0.67	0.59
FinTral-DPO-T&R	♡ + ★ + ■	**0.83**	**0.83**	0.60	**0.72**	**0.56**	0.62	**0.75**	0.70
GPT-4-Turbo (gpt-4-1106-preview)	♡ + ★ + ■	0.79	0.80	**0.83**	0.65	0.54	**0.70**	0.73	**0.72**

Table 6: Comparative analysis of LLMs using external tools on diverse tasks.

comprises of CLIP and FinTral-DPO. As Table 7 shows, GPT-4V performs best, with scores of 0.79 in ChartQA and 0.89 in FinVis, averaging 0.84. Gemini-Pro follows closely, with a consistent performance across both datasets, scoring an average of 0.78. Other models like Qwen-VL-Plus, FinTral-VL, and LLaVa-NEXT show varying degrees of efficacy: Qwen-VL-Plus performing notably better in ChartQA (0.78) than in FinVQA (0.64), while FinTral-VL and LLaVa-NEXT trail behind, indicating areas for potential improvement in their visual data interpretation capabilities. FinTral-VL performs well on the FinVQA dataset, making it highly suited for multimodal financial usage. Figure D.6 shows examples of models' outputs on questions from the FinVQA dataset.

Method	LLM	ChartQA	FinVQA	CU
Closed-source API				
Gemini-Pro	-	0.74	0.82	0.78
QwenVL-Plus	-	0.78	0.64	0.71
GPT-4V	-	**0.79**	**0.89**	**0.84**
Open-source MLLMs				
LLaVA	Vicuna-7B	0.12	0.25	0.19
InstructBLIP	Vicuna-7B	0.34	0.23	0.29
LLaVA-1.5	Vicuna-13B	0.44	0.32	0.38
Qwen-VL-Chat	Qwen-7B	0.53	0.34	0.44
LLaVa-NEXT	Yi-34B	**0.65**	0.58	0.62
FinTral-VL (ours)	FinTral-DPO	0.63	**0.75**	**0.69**

Table 7: Comparison with available MLLMs on Chart Understanding datasets.

5.4 Financial Hallucination Evaluation

Since financial hallucinations can be complex to measure, we have used three different methods and datasets to quantify hallucinations. We first assess how much models hallucinate in selecting definitions of financial terms. We then conduct human evaluations of the appropriateness of responses from top LLM models based on our first task. Finally, we evaluated them on the Finance Bench (Islam et al., 2023) dataset, a complex numerical question-answering dataset requiring mathematical tools and retrieval.

FinTerms-MCQ In FinTerms-MCQ dataset, we convert definitions of financial terms from Investopedia (2024) to a multiple choice format using the right definition and three other closely related definitions. We then ask the models to select the right definition. We derive a hallucinations index (HI), defined as the proportion of correctly generated definitions by each model (higher is better), based on the models' performance in this MCQ task. As seen in Table 8, the models' performances on the HI vary significantly. GPT-4 and ChatGPT lead the pack with exceptionally high scores of 98% and 95%, respectively. All three of our models perform better than the other open-source LLMs. In particular, FinTral-DPO-T&R show a strong performance with an HI of 97%.

Model	Type	HI
FinMA-7B-trade	♠	0.28
Vicuna-7B	◇	0.55
Llama-2-7b	♣	0.64
FinMA-7B	♠	0.64
Mistral-7B	♣	0.67
Llama-2-7b-chat	♡	0.70
Llama-2-13b-chat	♡	0.75
Mistral-7B-Instruct	◇	0.76
FinMA-7B-full	♠	0.80
FinTral-INST	◇	0.82
FinTral-DPO	♡	0.88
ChatGPT	♡	0.95
FinTral-DPO-T&R	♡ + ■	0.97
GPT-4-Turbo	♡ + ■	**0.98**

Table 8: Comparison of various models based on Hallucinations Index (HI). This index represents the proportion of correctly generated definitions by each model (higher is better).

FinTerms-Gen In Table D.1, we show an example of how popular LLMs, like ChatGPT, hallu-

cinate in the financial domain. We generate answers to questions related to the financial terms in the FinTerms-Gen dataset (n=150, see Table 4) using the three models with best performance on FinTerms-MCQ (i.e, GPT-4, ChatGPT, and FinTral-DPO+T&R). We then ask two humans, each with at least four years of background in finance, to label the responses with one of the four correctness tags shown in Figure 4. The two annotators agree with a Cohen's kappa (K) of 0.85. As Figure 4[3] shows our FinTral-DPO-T&R produces more correct and satisfying responses (category A in Figure 4) than ChatGPT but falls short of GPT-4.

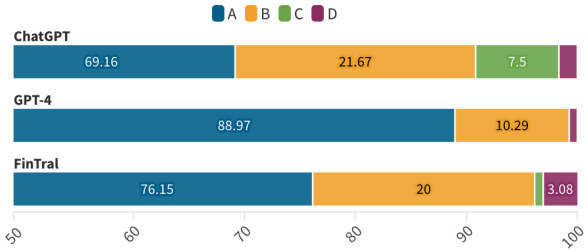


Figure 4: Human Evaluation on FinTerms Dataset. ***FinTral:*** is our FinTral-DPO-T&R. Each bar is segmented into four colors representing the quality of responses: **A**: correct and satisfying response **B**: acceptable response with minor imperfection, **C**: responds to the instruction but has significant errors, **D**: irrelevant or invalid response.

Finance Bench Finance Bench (Islam et al., 2023) is a proprietary dataset designed to assess the capabilities of LLMs in the context of open-book financial question answering (QA). While the full version includes 10,231 questions related to publicly traded companies, each accompanied by evidence strings and relevant answers, we evaluate our models using FinanceBench's open-source sample of 150 questions as provided in Islam et al. (2023) using the same methodology adopted by the authors. As presented in Figure 5, the FinTral-DPO-T&R performs very well on this dataset, outperforming the other models, GPT-4 (OpenAI, 2023c), Claude (Models, 2023), and Llama-70B (Touvron et al., 2023), evaluated in Islam et al. (2023). Using retrieval and tools in FinTral-DPO-T&R proves its efficiency and puts the model ahead of all the other models.

[3]We use only Q&A pairs where both annotators agree (n=128 pairs) for this analysis.

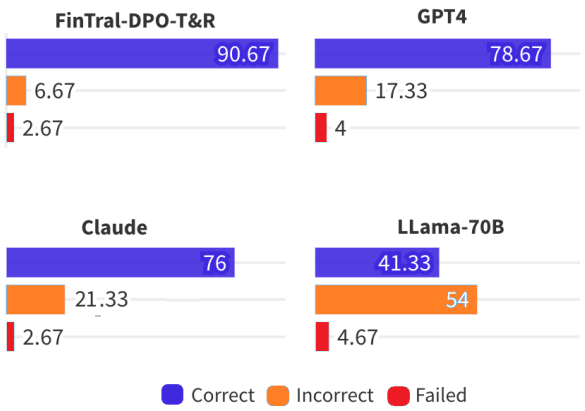


Figure 5: Performance comparison of various models on the FinanceBench dataset. Each model's percentage of correct, incorrect, and failed responses is shown. FinTral-DPO-T&R and GPT4 outperform other models, with LLama-70B having the highest failure rate.

6 Discussion

Advancements in financial LLMs FinTral leverages extensive datasets and diverse training methods, including instruction fine-tuning and RLAIF, to enhance its analysis of complex financial data across multiple modalities. The integration of advanced tools further augment its financial capabilities.

Reducing model hallucinations FinTral combats financial hallucinations by pretraining with up-to-date, clean financial data and employing RLAIF and retrieval methods, enhancing model accuracy and reliability.

Human-AI collaboration in financial decision-making Enhancing FinTral's real-time adaptability to financial markets through dynamic data retrieval and live data analysis can significantly boost its predictive accuracy and assist in informed decision-making. Figure E.1 shows how this model can be used in real world.

7 Conclusion

We presented FinTral an advanced multimodal financial language model with remarkable capabilities. Key advancements include integrating textual, numerical, and visual data, a training pipeline with various finetuning capabilities, and employment of tools and retrieval mechanisms. The model effectively addresses challenges like financial hallucination, evidenced by high performance in various financial tasks compared to baseline models. The achievements of FinTral hold a great potential for financial models of a moderate size (e.g., 7B).

8 Limitations

While FinTral represents a significant advancement in the realm of financial large language models (LLMs), it is important to acknowledge inherent limitations:

1. **Domain-Specific Adaptability:** Tailored for the financial domain, FinTral may not perform as effectively outside its trained scope, potentially limiting its generalizability.

2. **Handling of Real-Time Data:** While designed for real-time analysis, the model's predictive accuracy depends on the timeliness and accuracy of incoming data, which may be affected by rapidly changing market conditions.

3. **Maintenance and Updating:** Continuous updating and maintenance are required to keep the model relevant and effective in evolving financial markets and regulations.

Acknowledging these limitations is crucial for the responsible deployment and continued development of FinTral and similar financial LLMs.

9 Ethics Statement

Energy Efficiency. Our FinTral models, similar to many large language models (LLMs), required significant training time and computational resources, and thus are not particularly energy efficient. We acknowledge this as a critical issue and advocate for ongoing research into developing more energy-efficient models.

Data. Our pretraining datasets are collected from public domains, encompassing a wide range of financial topics and sources. While these datasets provide comprehensive coverage for financial language modeling, we must be aware of the potential biases and limitations inherent in publicly available data, ensuring our model remains as objective and unbiased as possible.

Data Copyright. We emphasize that all datasets used, including those from SEC filings, news sources, and social media, are collected from publicly available sources. We confirm that our data collection process respects the copyrights of these sources and does not infringe upon any proprietary data.

Model Release. We plan to release our models responsibly. Given the sensitive nature of financial data and the potential for misuse, we will implement strict guidelines and conditions for the use of FinTral, particularly in real-world applications. This includes clear guidelines on ethical usage and the avoidance of deployment in contexts that could lead to unethical practices such as market manipulation or privacy violations.

Privacy. FinTral is developed using publicly available data, which mitigates concerns regarding personal information leakage. However, given the sensitive nature of financial data, we have taken extra precautions to ensure that no identifiable personal or corporate financial information is retrievable from our trained models.

Human Annotation. The human annotators involved in this project are professionals with expertise in finance and natural language processing. No sensitive or personally identifiable data was used in the annotation process, adhering to ethical guidelines and data privacy standards. The human annotators are co authors on this paper.

Bias Analysis. We recognize that any language model can inadvertently perpetuate biases present in its training data. In FinTral's case, potential biases might be related to financial markets, regions, or corporate entities. We conducted thorough analysis to identify and mitigate such biases, ensuring that our model's outputs are as fair and unbiased as possible. However, users should remain aware of these potential biases, especially when applying the model to real-world scenarios.

Applications. While FinTral offers advanced capabilities for financial analysis, like any powerful tool, it can be misused. It's crucial to emphasize responsible usage, particularly in sensitive financial contexts. Users should avoid deploying FinTral for speculative trading, market manipulation, or any activity that could contravene financial regulations or ethical standards. Conversely, FinTral has the potential for beneficial applications such as financial education, research, and improving the accessibility of financial information.

AI usage. It's pertinent to acknowledge the role of AI tools such as ChatGPT in our project. Specifically, ChatGPT was utilized minimally and primarily for grammar corrections in our documents. This use was strictly confined to enhancing linguistic accuracy and improving the readability of our written materials. It's important to clarify that the core research, analysis, and development were conducted independently by our team.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Rui Cao, Nazli Ozum Kafaee, Arslan Aziz, and Hasan Cavusoglu. 2023. Market reaction to cyber strategy disclosure: Word embedding derived approach. In *Hawaii International Conference on Systems Science (HICSS 2023)*, pages 6078–6087.

Rui Cao, Gene Moo Lee, and Hasan Cavusoglu. 2020. Corporate social network analysis: A deep learning approach. *Workshop on Information Technologies and Systems (WITS 2020)*.

Clayton Leroy Chapman, Lars Hillebrand, Marc Robin Stenzel, Tobias Deußer, David Biesner, Christian Bauckhage, and Rafet Sifa. 2022. Towards generating financial reports from tabular data using transformers. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 221–232. Springer.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning.

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. 2021. 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, 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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Zhixuan Chu, Huaiyu Guo, Xinyuan Zhou, Yijia Wang, Fei Yu, Hong Chen, Wanqing Xu, Xin Lu, Qing Cui, Longfei Li, Jun Zhou, and Sheng Li. 2023. Data-centric financial large language models.

Tri Dao. 2023. FlashAttention-2: Faster attention with better parallelism and work partitioning.

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

Sohom Ghosh, Shovon Sengupta, Sudip Kumar Naskar, and Sunny Kumar Singh. 2022. Finrad: Financial readability assessment dataset - 13,000+ definitions of financial terms for measuring readability. In *Proceedings of the The 4th Financial Narrative Processing Workshop @LREC2022*, pages 1–9, Marseille, France. European Language Resources Association.

Yue Guo, Zian Xu, and Yi Yang. 2023. Is chatgpt a financial expert? evaluating language models on financial natural language processing.

Udit Gupta. 2023. Gpt-investar: Enhancing stock investment strategies through annual report analysis with large language models.

Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

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

Investopedia. 2024. Financial terms dictionary.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination.

Moreno La Quatra and Luca Cagliero. 2020. End-to-end training for financial report summarization. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 118–123.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023b. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 13–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. pages 1941–1942.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2013. Good debt or bad debt: Detecting semantic orientations in economic texts.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Eliza Mik. 2017. Smart contracts: terminology, technical limitations and real world complexity. *Law, innovation and technology*, 9(2):269–300.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Claude Models. 2023. Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

NIST. 2018. Framework for improving critical infrastructure cybersecurity. https://nvlpubs.nist.gov/nistpubs/cswp/nist.cswp.04162018.pdf.

OpenAI. 2023a. Chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023b. Gpt-4 technical report.

OpenAI. 2023c. Gpt-4 technical report.

Ross Quinlan. Statlog (Australian Credit Approval). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C59012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Pawan Kumar Rajpoot and Ankur Parikh. 2023. Gptfinre: In-context learning for financial relation extraction using large language models.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.

Bhaskarjit Sarmah, Tianjie Zhu, Dhagash Mehta, and Stefano Pasquali. 2023. Towards reducing hallucination in extracting information from financial reports using large language models.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023a. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, Toronto, Canada. Association for Computational Linguistics.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023b. Finer: Financial named entity recognition dataset and weak-supervision model.

Ankur Sinha and Tanmay Khandait. 2020. Impact of news on the commodity market: Dataset and results.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Wataru Souma, Irena Vodenska, and Hideaki Aoyama. 2019. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2(1):33–46.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay

Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah,

Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

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

Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023b. Finvis-gpt: A multimodal large language model for financial chart analysis.

Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1627–1630.

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*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

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.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*.

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

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *FinLLM Symposium at IJCAI 2023*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Ali Babar, and Xiao-Yang Liu. 2023b. Enhancing financial sentiment analysis via retrieval augmented large language models.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. *arXiv preprint arXiv:1904.07535*.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124, Online. Association for Computational Linguistics.

Appendices

A Detailed Related Works

Financial NLP Models and Their Challenges There have been successful applications of traditional Natural Language Processing (NLP) techniques a range of finance related problems. These include named entity recognition (Salinas Alvarado et al., 2015) sentiment analysis of financial news (Souma et al., 2019; Araci, 2019), event extraction (Yang et al., 2018; Zheng et al., 2019), generating financial reports (Chapman et al., 2022), and text summarization in a financial context (La Quatra and Cagliero, 2020).

However, deploying NLP models for domain-specific tasks in the financial sector faces several distinct challenges. Firstly, the complex and jargon-rich nature of financial language poses a significant barrier in achieving desirable performance from the models, often leading to a gap in understanding the domain-specific documents (Mik, 2017). Secondly, the scarcity of annotated datasets, combined with the high costs associated with data annotation, in finance, hinders the advancement of these models. Thirdly, existing NLP models often fall short in inferential capabilities, particularly in critical tasks such as risk assessment and making informed decision-making in investment contexts (Liu et al., 2023b). Additionally, the dynamic nature of financial markets requires models to be capable of real-time analysis, a feature that many current models do not possess. Numerical information processing, a common element in financial documents filled with figures and symbols, also poses a significant challenge for understanding financial documents. The challenge is further exacerbated by the fact that many graphs and figures are in image formats in these documents. Lastly, the wide-spread adaptability of many NLP models remains limited, as they are typically optimized for a particular single-task function and lack the ability to generalize across multiple tasks (Mishra et al., 2021). In light of these challenges, it is imperative for ongoing research efforts to develop more advanced, versatile, and robust NLP models tailored to the dynamic and complex requirements for financial document undertanding.

Financial Large Language Models Finance has witnessed significant advancements in large language models, starting with the introduction of FinBERT (Araci, 2019). This early contribution sets a precedence for using pre-trained language models in financial sentiment analysis, demonstrating marked improvements in performance metrics. In 2023, a series of groundbreaking models have further propelled the field. BloombergGPT (Wu et al., 2023) emerged as a 50-billion parameter model trained on an extensive financial data corpus. Its training on a diverse dataset enabled it to excel in financial tasks while maintaining robust performance in general LLM benchmarks. PIXIU (Xie et al., 2023) followed, presenting a comprehensive framework with a financial LLM fine-tuned with instruction data. PIXIU was a crucial development in advancing the open-source development of financial AI, combining a novel instruction dataset and an evaluation benchmark for financial LLMs. The same year saw the introduction of Instruct-FinGPT (Zhang et al., 2023a), which utilized instruction tuning to enhance financial sentiment analysis. This model particularly excelled in scenarios requiring deep numerical understanding and contextual comprehension. Another significant advancement was GPT-FinRE (Rajpoot and Parikh, 2023), focusing on financial relation extraction using in-context learning. This model demonstrated high effectiveness and accuracy by employing two distinct retrieval strategies. Adding to the multimodal capabilities in financial LLMs, FinVis-GPT (Wang et al., 2023b) was proposed, designed explicitly for financial chart analysis. This model leveraged the power of LLMs along with instruction tuning and multimodal capabilities, showcasing superior performance in related tasks. GPT-InvestAR (Gupta, 2023) aimed to enhance stock investment strategies by analyzing annual reports using LLMs. This approach yielded promising results in outperforming traditional market returns, highlighting the potential for LLMs in investment strategies. InvestLM (Yang et al., 2023b) showed strong capabilities in understanding economic text and providing practical investment advice. With retrieval-augmented LLMs (Zhang et al., 2023b) addressed the challenges of applying LLMs directly to economic sentiment analysis, achieving considerable performance gains. FinGPT (Wang et al., 2023a) focused on creating a benchmark for Instruction Tuning of LLMs in financial datasets, emphasizing the integration challenges and potential solutions for GPT-based models specialized in the financial domain. Sarmah et al. (2023) reduced hallucination in information extraction from earning call transcripts and achieved improved the accuracy by combin-

ing retrieval-augmented generation techniques with metadata. FinLMEval (Guo et al., 2023) assessed the performance of LLMs in financial natural language processing tasks, offering foundational evaluations for ongoing efforts to enhance LLMs in the financial domain. DISC-FinLLM (Chen et al., 2023) introduced a Chinese financial LLM based on a Multiple Experts Fine-tuning Framework, showing improved performance in various monetary scenarios compared to baseline models. Lastly, the work on data-centric financial LLMs (Chu et al., 2023) presented a novel approach to better handle financial tasks with LLMs, emphasizing data preprocessing and pre-understanding, resulting in substantial performance improvements on economic analysis and interpretation tasks. These contributions collectively illustrate the rapid growth in utilization of LLMs and their tremendous potential in various financial applications, showcasing their capacities in revolutionizing financial analysis, forecasting, and decision-making processes.

B Pretraining Data Details

Common Crawl Data The Common Crawl dataset, specifically the C4 snapshot from 2019 to 2021, comprising over 10 billion files, was an initial broad data source. Text classification via the ELECTRA Finance domain-specific language model ensured that the dataset maintained a strong relevance to financial content. Rigorous domain filtering and data pruning were employed, isolating financial-specific texts and discarding irrelevant content. The final dataset consisted of 800 million documents, including 300 million English-only and 500 million multilingual files, providing a comprehensive base for financial analysis.

News Scraping Our approach extended to news scraping, particularly focusing on the period from July 2022 to July 2023. With 300 million data lines, this dataset allowed for in-depth analysis of market trends and financial narratives. The dataset encapsulated a global view of financial markets by integrating sources like Yahoo, Seeking Alpha, Eastmoney, and Yicai. This multi-source strategy ensured a robust, cross-referenced, and credible dataset. We used scrapers implemented in (Yang et al., 2023a) to build out News datasets.

SEC Filings An exhaustive scrape of the EDGAR SEC database from 1993 to 2023 provided detailed records of accurate business, financial and accounting information from official filings. This dataset, exclusively in English, added substantial depth, allowing for analysis of historical market regulatory impacts and corporate financial maneuvers.

Company Websites and Social Media Further data were obtained from the top 5000 company websites and their social media presence on platforms like Facebook, Instagram, and Reddit. This dataset provided direct corporate communications and captured broader market sentiments and public perceptions, notably through an extensive scrape of the r/WallStreetBets Reddit community.

C Financial data cleaning and deduplication pipeline

We started of gathering various text corpora shown in table 1, resulting in a dataset consisting of 2.9B documents. The data that we collected is not only unclean but also suffers from large-scale duplication. As shown by (Soboleva et al., 2023), using clean and deduplicated data is computationally efficient for model training. The data cleaning and deduplication pipeline for financial data begins with URL filtering, in which the raw data is initially processed. This crucial step ensures the inclusion of only pertinent URLs, enhancing the dataset's quality by excluding irrelevant or unsuitable sources. Once the URLs are streamlined, the Text extraction phase commences, whereby contents of documents from the selected URLs are meticulously extracted, filtering out images while maintaining the large dataset scale. Following this, the language identification phase excluded non-English documents by categorizing them based on the language of their tokens. Subsequently, the pipeline further refines the data through Document-wise domain-based filtering, narrowing down to 100 billion tokens pertinent to the financial domain by excluding 55B-token non-financial documents. Recognizing the importance of data privacy and relevance, the pipeline incorporates removing sensitive information, which is done using a classifier built using FinBERT (Araci, 2019). Line-wise corrections enhance accuracy and filter out 5B tokens of sensitive information. An extensive Fuzzy deduplication process reduces the data to 38 billion tokens. This is followed by an Exact deduplication method, which trims another 13 billion tokens. Finally, the text cleaning process identifies and excludes 5B improper tokens, including all sensitive information. Ultimately, the pipeline crafts a streamlined financial dataset, culminating in a

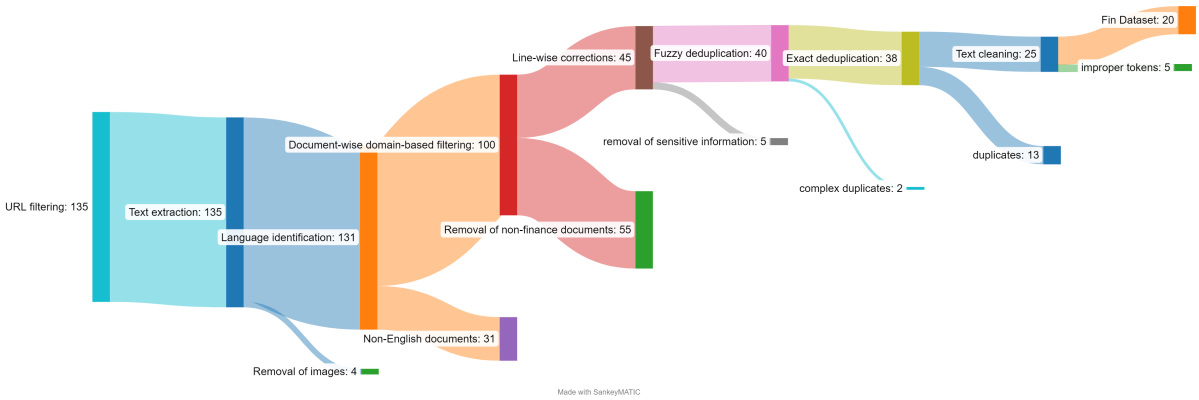


Figure C.1: Explanation of our data deduplication pipeline

concise 20B-token financial dataset. The pipeline is illustrated in figure C.1.

D Downstream Dataset details

Notable datasets include FPB and FiQA-SA, both utilized for sentiment analysis, with the former comprising 48,450 news texts (Malo et al., 2013) and the latter encompassing 11,730 news headlines and tweets (Maia et al., 2018). The FOMC dataset, consisting of 496 FOMC transcripts, serves the hawkish-dovish classification task (Shah et al., 2023a), whereas the Headline dataset, with 11,412 news headlines, aids in news headline classification (Sinha and Khandait, 2020). Named entity recognition is the focus of the NER and Finer-Ord datasets (Salinas Alvarado et al., 2015; Shah et al., 2023b). We brought in ECTSUM and EDTSUM (Mukherjee et al., 2022; Zhou et al., 2021) for text summarisation. For text classification, we included two credit scoring datasets from the German and Australia (Hofmann, 1994; Quinlan). We employed FinQA introduced by the current paper and ConvFinQA (Chen et al., 2021, 2022) for numerical understanding task. We used three existing datasets for stock movement prediction, namely BigData22, ACL18, and CIKM18 (Soun et al., 2022; Xu and Cohen, 2018; Wu et al., 2018).

Firm Disclosure Datasets This study employed three datasets that serve as a microcosm of firm regulatory disclosures. Each consists of labelled text segments from comprehensive reports annually filed with the Security and Exchange Commission (SEC) by public companies to inform investors regarding their financial health and business risks. The 'Firm Social Relationships' (FSR) dataset provides insight into the intricate network of corporate interactions, categorized into several key relational dimensions: ownership, alliances, competition, and board interlock relationships (Cao et al., 2020). They identified 3931 sentences stating another firm in a focal firm's disclosure. Domain experts classified the relationship between the focus firm and the firm into one or none of these relationships. The 'Cyber Strategies' (CS) dataset contains disclosure sentences describing the firm's cybersecurity strategies (Cao et al., 2023). Experts labelled 240 cybersecurity-related sentences from firms' disclosures into one of five strategies delineated by the National Institute of Standards and Technology: Identification, Protection, Detection, Response, and Recovery (NIST, 2018). The 'IT Risk Disclosure' (ITR) dataset is created for this study using the Risk Factors section of the firm's annual disclosure. Domain experts categorized 1,196 sentences related to Information Technology into one or none of the 11 IT risk categories. These datasets curated by domain experts are pivotal to our zero-shot evaluation framework, which tests the models' utility against genuine instructional data—thus bridging the gap between theoretical model performance and practical utility in real-world scenarios.

Financial Chart Understanding Dataset

The FinVQA dataset addresses tasks involving questions about trends and details depicted in plots and graphs embedded in images. This dataset includes a variety of financial charts, such as line, bar, and candle charts , all meticulously annotated by humans and accompanied by multi-turn conversations associated with each image. We developed two versions of the FinVQA dataset. The first, illustrated in Fig. D.1, focuses on simple questions related to stock market charts and requires the model to interpret these charts while identifying trends

FinVQA: Simple Intuitive Questions about the chart

Figure D.1: Illustration of the FinBEAT VL Model's Proficiency in Chart Understanding Tasks. The panel demonstrates the model's ability to interpret stock price movements and identify peak values from a candlestick chart.

FinVQAv2: Complex instructions for chart understanding

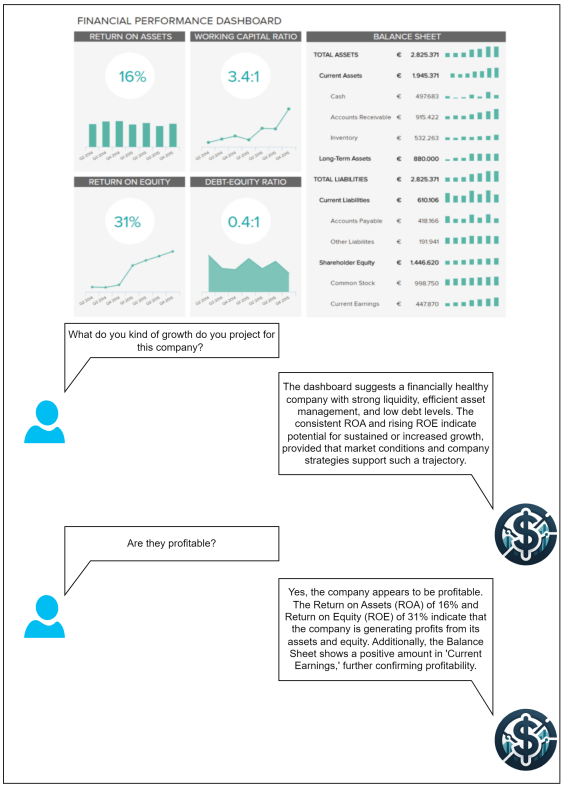


Figure D.2: Illustration of the FinBEAT VL Model's Proficiency in Chart Understanding Tasks. The panel showcases the model's adeptness at analyzing financial dashboards, projecting growth, and evaluating profitability based on key financial ratios and balance sheet data.

and performing analysis. These are randomly chosen 100 images that have been human annotated with different types of questions.

FinVQAv2, presented in Fig. D.2, is a dataset that captures more advanced and challenging situations. We gathered 100 images containing financial visualizations from various sources and collected relevant question-and-answer pairs from the experts to build our dataset. It encompasses a diverse array of financial graphics and various relevant questions posed in relation to each image. For instance, we present a financial dashboard containing various financial metrics in numerical and graphical formats, and we ask the model to perform complex calculations using the data extracted from the image.

Hallucinations Evaluation In the financial context, Large Language Models (LLMs) like GPT-4 are prone to hallucinations, giving incorrect answers or misinterpreting basic facts, as shown by (Kang and Liu, 2023). We generated two datasets: FinTerms-MCQ and FinTerms-Gen.

To build FinTerms-MCQ, we generated a dataset containing 1129 financial terms and their definitions, using the method described by (Ghosh et al., 2022). This dataset assesses the foundational financial knowledge of various LLMs and investigates if retrieval-based methods can reduce the incidence of hallucinations. We built this dataset in a multiple-choice format with the question and four options; all four are closely related, and only one is correct.

FinTerms-Gen is built as a generation task where we collected terms from Investopedia (2024), and then we asked our models to answer the definitions. Examples from this dataset are presented in Table D.1.

E Applications

Figure E.1 shows how this model can be used in real life.

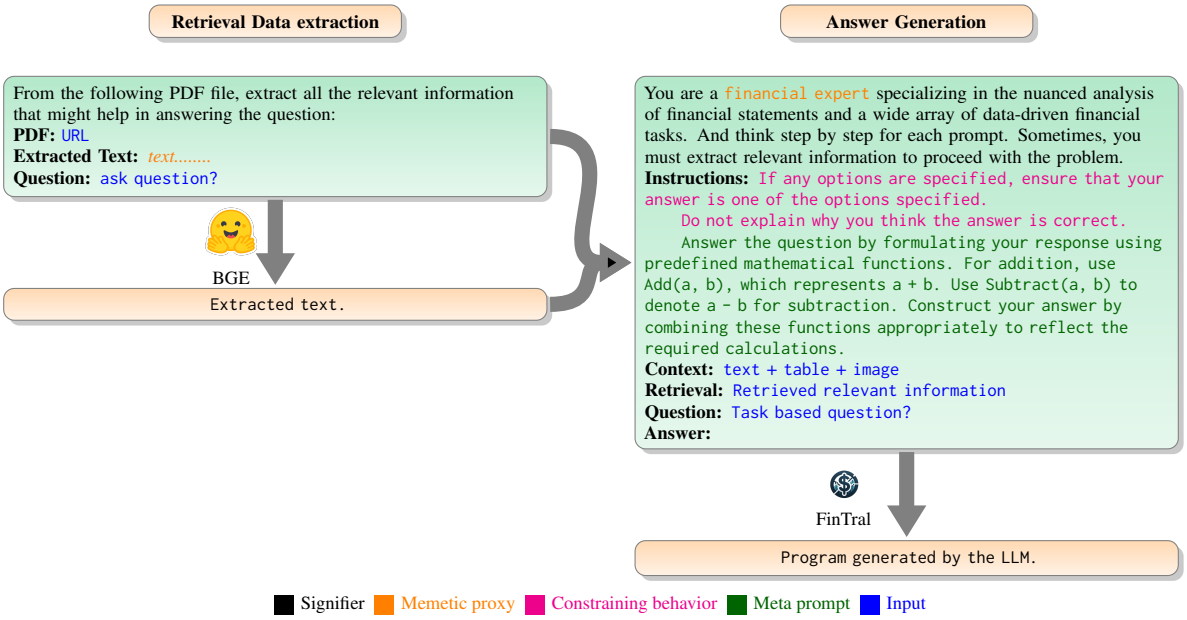


Figure D.3: Prompting Method for FinTral-RL-T&R

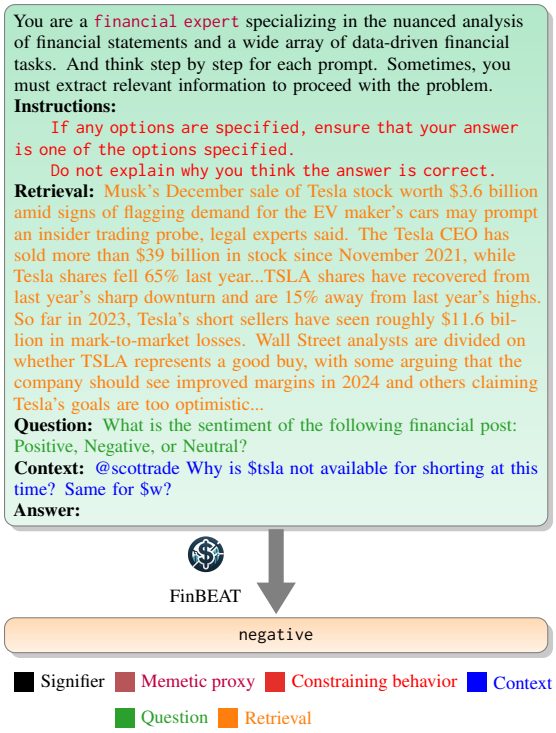


Figure D.4: FinBEAT Prompting method Example

```
{
  "prompt": "How can you correlate a company stock's performance
            with overall market performance?",
  "chosen":
     "To correlate a company's stock performance with overall market
      performance, start by selecting a relevant market index like the
      S&P 500, Dow Jones, or NASDAQ, which matches the company's
      industry and geographical location. Collect historical data for
      the company's stock prices and the chosen market index over the
      same period. Convert this data into returns using the formula
      (Current Price - Previous Price) / Previous Price to standardize
      it for comparison. Calculate the Pearson correlation coefficient
      between the stock's and market index returns to measure how
      closely they move together. Also, determine the stock's beta
      coefficient to assess its volatility relative to the market. For
      visual insight, plot a scatter plot of the stock's returns
      against the market index. Conduct a linear regression analysis
      with the market index as the independent variable and the stock
      as the dependent variable to further explore the market's
      influence on the stock. Lastly, consider external factors like
      economic conditions and industry-specific news, as they can
      significantly impact stock performance.",
  "rejected":
      "There are several ways to correlate a company's stock
      performance with overall market performance:\n\n1. Market
      capitalization: The market capitalization, or market cap, of a
      company is calculated by multiplying the total number of shares
      outstanding by the current stock price. The market cap of a
      company can be compared to the overall market cap of the stock
      market to determine the company's market exposure.\n2. Industry
      average stock performance: Comparing a company's stock
      performance to the average stock performance of its industry can
      provide insight into how the company performs relative to its
      peers. Sector performance: Comparing a company's stock
      performance to the performance of stocks in the same sector can
      provide insight into how the company performs relative to other
      companies in the same industry. Economic indicators: Economic
      indicators such as GDP, inflation, and interest rates can impact
      the stock market's and individual companies' overall performance.
      By analyzing the correlation between a company's stock
      performance and these economic indicators, investors can gain
      insight into how broader economic trends may impact the company's
      performance.",
}
```


Figure D.5: Example from the Finance AI Feedback Dataset

Hallucinations index

Examples of No Hallucinations

Prompt: What is the definition of 'Low/No Documentation Loan' in the financial domain?
Wikipedia: A No-Doc or Low-doc loan (abbr: No/Low Documentation Loan) refers to loans that do not require borrowers to provide income documentation to lenders or require little documentation.

FinTral-RL-T&R: A low/no documentation loan allows a potential borrower to apply for a mortgage while providing little or no information regarding their employment, income, or assets. Regulation of these loans has evolved significantly since 2008, but they remain an option for some borrowers in nontraditional financial situations.
Annotator Rating: A

ChatGPT: A low/no documentation loan is a type of mortgage or loan where the borrower is not required to provide extensive financial documentation, such as income and asset verification, making it easier to qualify but often resulting in higher interest rates.
Annotator Rating: A

GPT-4: A Low/No Documentation Loan is a type of loan that requires minimal or no verification of the borrower's income, assets, or employment, often used by individuals who may not have traditional income sources or documentation.
Annotator Rating: A

Examples of LLMs Hallucinations

Prompt: What is the definition of 'Open Market' in the financial domain?
Wikipedia: An open market is an economic system without trade barriers to free market activities. In an open market, buyers and sellers can do business freely without common barriers, such as unfair licensing agreements, arbitrary taxes, unionization, subsidies, and other regulations affecting regular market operations.

FinTral-RL-T&R: An open market is an economic system with little to no barriers to free-market activity. An open market is characterized by the absence of tariffs, taxes, licensing requirements, subsidies, unionization, and any other regulations or practices that interfere with free-market activity. Open markets may have competitive barriers to entry, but never any regulatory barriers to entry.
Annotator Rating: A

ChatGPT: In finance, the Open Market refers to a marketplace where securities are traded directly between buyers and sellers without the intervention of a third party, such as an exchange.
Annotator Rating: C

GPT-4: An "Open Market" in the financial domain refers to an economic system with few barriers to free market activity. It typically involves unrestricted buying and selling of goods and services, with prices determined by supply and demand, and minimal government intervention in terms of tariffs, quotas, or regulations. This concept is fundamental in the operation of capital markets, where securities are freely traded.
Annotator Rating: A

Table D.1: Examples of Hallucinations in LLMs

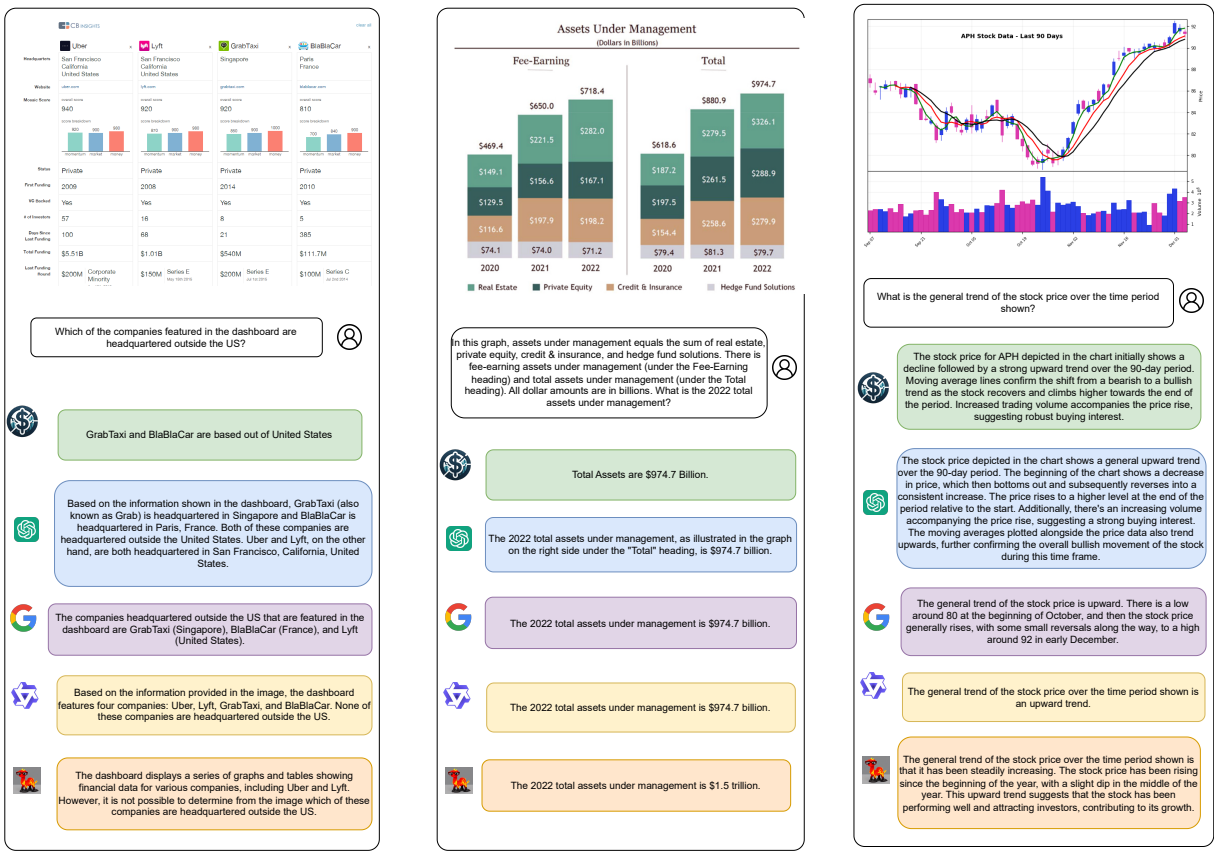
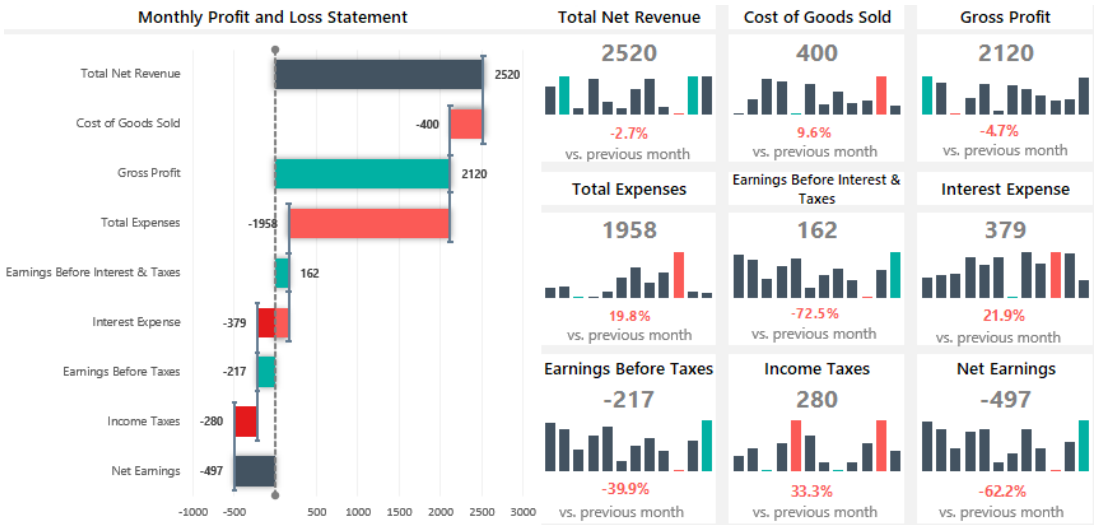


Figure D.6: Example of different VL models on different FinVQA tasks

Applications of FinTral



Calculate the EBITA Margin here

Retrieval:

To calculate the EBITA (Earnings Before Interest, Taxes, and Amortization) margin, you would use the formula:

EBITA Margin = EBITA / Total Net Revenue × 100

However, in the provided Profit and Loss Statement, only EBIT (Earnings Before Interest and Taxes) is given. Amortization is not listed, so we'll assume EBITA is equal to EBIT in this case.

From the image, the EBIT (Earnings Before Interest & Taxes) is given as 162, and the Total Net Revenue is 2520.

Hence the EBIT = multiply(divide(162,2520),100)

Hence the EBIT is 6.43%

Figure E.1: Applications of the FinTral Model