FinDVer: Explainable Claim Verification over Long and Hybrid-Content Financial Documents

Yilun Zhao **Yitao Long** **Yuru Jiang** **Chengye Wang** **Weiyuan Chen**
Hongjun Liu **Yiming Zhang** **Xiangru Tang** **Chen Zhao** **Arman Cohan**

Yale NLP - FinDVer Team

Abstract

We introduce FinDVer, a comprehensive benchmark specifically designed to evaluate the explainable claim verification capabilities of LLMs in the context of understanding and analyzing long, hybrid-content financial documents. FinDVer is divided into three subsets: information extraction, numerical reasoning, and knowledge-intensive reasoning—each addressing common scenarios encountered in real-world financial contexts. We assess a broad spectrum of LLMs under long-context and RAG settings. Our results show that even the current best-performing system, Claude-3.5-Sonnet, significantly lags behind human experts. Our detailed findings and insights highlight the strengths and limitations of existing LLMs in this new task. We believe FinDVer can serve as a valuable benchmark for evaluating LLM capabilities in claim verification over complex, expert-domain documents.

github.com/yilunzhao/FinDVer

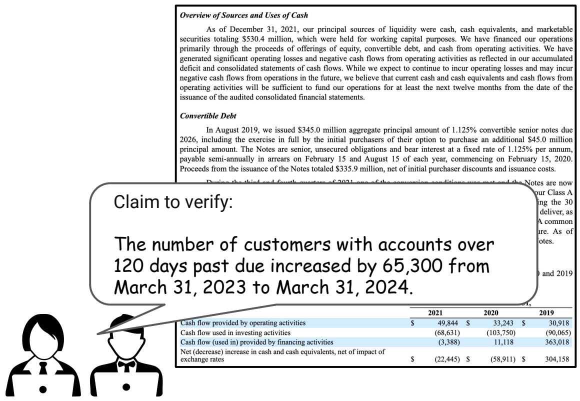

Figure 1: An example from the *numerical reasoning* subset of the FinDVer benchmark. To verify the claim, the LLM is required to first locate claim-relevant data points within long and hybrid-content financial documents, and then apply numerical reasoning over the extracted data points for claim verification.

1 Introduction

In today's information explosion era, the responsibility of verifying the truthfulness of the item is often passed on to the audience.unverified claims about a company's financial performance frequently circulate in online media, potentially misleading investors. Therefore, it is crucial to verify these claims using the companies' original financial documents (*i.e.,* earnings reports and regulatory filings). Recent advancements in Large Language Models (LLMs) have attracted significant attention due to their capabilities in solving a broad range of tasks (Touvron et al., 2023b; Jiang et al., 2023b; OpenAI, 2023a). However, it remains particularly difficult for applying them to document-grounded claim verification in real-world financial domains due to the following two reasons:

First, financial documents are typically long, intricate and dense, and they include both quantitative tables and qualitative text (Chen et al., 2021; Zhu et al., 2021; Zhao et al., 2022, 2023d; Koncel-Kedziorski et al., 2024). Extracting and analyzing claim-relevant data from these documents requires complicated document comprehension abilities and professional knowledge in financial domains. Moreover, the type of reasoning involved encompasses various unique aspects that are less studied, necessitating a dedicated approach to evaluation and application.

Second, in the financial domain, where decisions often involve significant stakes, it is often critical to provide clear and comprehensible rationales for any claim verification decisions (Atanasova et al., 2020, 2023). However, existing *context-grounded* claim verification benchmarks (Chen et al., 2020; Kamoi et al., 2023; Lu et al., 2023; Glockner et al., 2024) primarily focus on the task of entailment classification and do not evaluate the reasoning process. This hinders the practical application and evaluation of LLMs in real-world scenarios.

In response to the aforementioned pressing need,

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 14739–14752
November 12-16, 2024 ©2024 Association for Computational Linguistics

Dataset	Input Context	Annotation / Data Creation	# Label	w. Explanation?	Reasoning-Intensive?
PubHealthTab (Akhtar et al., 2022)	Wikipedia table	Crowd-sourced	4	✗	✗
TABFACT (Chen et al., 2020)	Wikipedia table	Crowd-sourced	2	✗	✓
INFOTABS (Gupta et al., 2020)	Wikipedia table	Crowd-sourced	3	✗	✓
SCITAB (Lu et al., 2023)	Scientific table	Expert & InstructGPT	3	✗	✓
HOVER (Jiang et al., 2020)	Wikipedia articles	Crowd-sourced	2	✗	✗
DOCNLI (Yin et al., 2021)	News article	From summrization datasets	2	✗	✗
ContractNLI (Koreeda et al., 2021)	Contract	Expert & Crowd-sourced	2	✗	✗
LLM-AGGREFACT (Tang et al., 2024)	Doc from various domains	From existing benchmarks	2	✗	✗
WICE (Kamoi et al., 2023)	Wikipedia article	Crowd-sourced	3	✗	✗
AMBIFC (Glockner et al., 2024)	Wikipedia article	Crowd-sourced	3	✗	✗
CLAIMDECOMP (Chen et al., 2022a)	Political article	Expert	6	✗	✗
SCIFACT (Wadden et al., 2020)	Scientific paper abstracts	Expert	2	✗	✓
LIAR++ (Russo et al., 2023)	Political article	From fact-check website	2	✓	✗
FullFact (Russo et al., 2023)	Web page	From fact-check website	2	✓	✗
PUBHEALTH (Akhtar et al., 2022)	Health Web page	From fact check website	4	✓	✗
FINDVER (ours)	**Long** financial doc with tables	Expert	2	✓	✓

Table 1: Comparison between FINDVER and existing *context-grounded* claim verification datasets. FINDVER is distinguished by four unique characteristics: (1) **Expert Annotation**: It is annotated by financial experts to ensure high data quality; (2) **Complex Document Comprehension**: It requires interpreting a mix of textual and tabular data within a long-context financial document; (3) **Examination on Reasoning-Process Explanation**: It enhances claim verifications with detailed explanations about the reasoning process, increasing the benchmark's practical value; and (4) **Diverse Reasoning for Real-world Scenarios**: It incorporates various reasoning challenges, such as extracting complicated information, performing numerical calculations, applying external professional knowledge, and conducting comparative analyses. Accordingly, we divide the benchmark into three focused subsets, each tailored to mirror distinct real-world financial analysis scenarios.

we present **FINDVER**, a comprehensive and domain expert-annotated explainable claim verification benchmark that first explores in the context of financial documents. The LLMs are tasked with generating explanations of their reasoning to verify claims labeled as *"entailed"* or *"refuted"*, based on the information in the provided document, which contains both textual and tabular data. To identify the common reasoning-intensive scenarios in claim verification based on financial documents, we engage with domain experts and conducted a preliminary study. This helped us determine three key types of scenarios that frequently arise in real-world settings: *information extraction*, *numerical reasoning*, and *knowledge-intensive reasoning*. For each scenario, we construct an evaluation set. Each example in our dataset is annotated with detailed supporting evidence and step-by-step reasoning-process explanations.

We evaluate a wide spectrum of open- and closed-source LLMs, specifically, 19 models from 10 organizations. The documents in our benchmark are exceedingly long; therefore, we employ two widely adopted real-world application settings—*retrieval augmented generation* (RAG) and *long-context*—in this study. The experimental results indicate that even the existing best-performing LLM (*i.e.,* Claude-3.5-Sonnet) still significantly lags behind human experts (77.2% versus 93.3%), demonstrating the challenges of our proposed benchmark. Our contributions are summarized below:

- We introduce FINDVER, the first comprehensive context-grounded claim verification benchmark for financial domains, presenting new challenges for state-of-the-art LLMs.
- We conduct an extensive evaluation that encompasses a wide range of LLMs, including those specialized in finance and math. We also evaluate both long-context and RAG settings to comprehensively assess the capabilities and limitations of existing LLMs in our task.
- Our experimental results reveal a noticeable performance gap compared to human experts. This highlights the limitations of current LLMs in complex real-world applications and the need for continued advancements.

2 Related Work

Claim Verification Benchmark Claim verification is a well-established research area with two main settings. The first is the open-domain setting, which involves using an external retriever to find the most relevant information from a large corpus to verify claims (Vlachos and Riedel, 2014; Thorne et al., 2018; Aly et al., 2021; Wadden et al., 2022; Rangapur et al., 2024; Ma et al., 2024). The second setting is context-grounded claim verification, which requires models to verify claims based on the provided document context (Chen et al., 2020; Kamoi et al., 2023; Lu et al., 2023; Glockner et al., 2024). This work focuses on the second setting, as it allows us to eliminate variability and dependency on the retriever's performance, thereby focusing on the evaluation of LLM performance on on accurately verifying claims within a given context. However, as illustrated in Table 1, existing context-grounded claim verification benchmarks have four notable limitations: they typically 1) focus on general domains, overlooking the specific challenges and intricacies present in specialized fields, 2) focus solely on entailment classification and do not evaluate the reasoning processes of models, 3) do not involve claims that require intensive reasoning and complicated document comprehension. These limitations hinder their effectiveness for evaluating LLMs in real-world practice.

Financial Evaluation Benchmark NLP techniques have been applied to various financial tasks, such as named entity recognition (Salinas Alvarado et al., 2015; Shah et al., 2023), sentiment analysis (Malo et al., 2013; Maia et al., 2018), stock movement prediction (Soun et al., 2022; Xu and Cohen, 2018; Wu et al., 2018), and summarization (Zhou et al., 2021; Mukherjee et al., 2022; Liu et al., 2022). More recently, there has been an increasing focus on tasks involving financial documents (*e.g.,* annual reports and regulatory filings), which are crucial for providing insights into a company's performance and strategies. Several QA benchmarks have been proposed to evaluate models' performance in answering questions based on financial documents, with a particular focus on numerical reasoning (Chen et al., 2021; Zhu et al., 2021; Zhao et al., 2022; Chen et al., 2022b; Koncel-Kedziorski et al., 2024; Zhao et al., 2024b,a). Despite these advancements, there remains a significant gap in the exploration of claim verification tasks within the financial domain. While the recent FIN-FACT benchmark (Rangapur et al., 2024) addresses explainable multimodal financial fact-checking, it primarily focuses on open-domain scenarios. Verifying claims derived from financial documents is crucial, as inaccuracies can significantly influence investment decisions and market perceptions. To bridge this gap, we introduce FINDVER, the first context-grounded claim verification benchmark, specifically designed for real-world financial document comprehension.

3 FINDVER Benchmark

FINDVER provides a robust evaluation benchmark for reasoning-intensive and explainable claim verification over long and hybrid-content financial documents. We present an overview of the FINDVER construction pipeline in Figure 2; and detail the task formulation, data construction, and quality validation process in the following subsections.

3.1 Task Formulation

We formally define the task of FINDVER within the context of LLMs as follows: Consider a *single* financial document d, containing textual data P and tabular data T, associated with a claim c that needs verification. The expert-annotated data we collect supports the following two tasks:

Entailment Classification The model is required to determine the entailment label $\ell \in \mathcal{L} = \{\textit{“entailed”}, \textit{“refuted”}\}$, based on the document context:

$$\ell = \arg\max_{\ell \in \mathcal{L}} P_{\mathbf{LLM}}(\ell \mid P, T, c) \tag{1}$$

Reasoning-process Explanation Generation The model is required to generate a natural language explanation e:

$$e = \arg\max_{e} P_{\mathbf{LLM}}(e \mid P, T, c) \tag{2}$$

which articulates the reasoning process behind the validity of the provided claim c, based solely on the provided textual content P and tabular content T within the financial document.

Notably, some claim verification systems, particularly those developed prior to the era of LLMs and for previous datasets that did not require explanation generation (Chen et al., 2020; Yin et al., 2021; Koreeda et al., 2021), might not explicitly perform explanation generation. Instead, they directly output the final label. For such systems, FINDVER can also be used for evaluation by focusing on the entailment classification task.

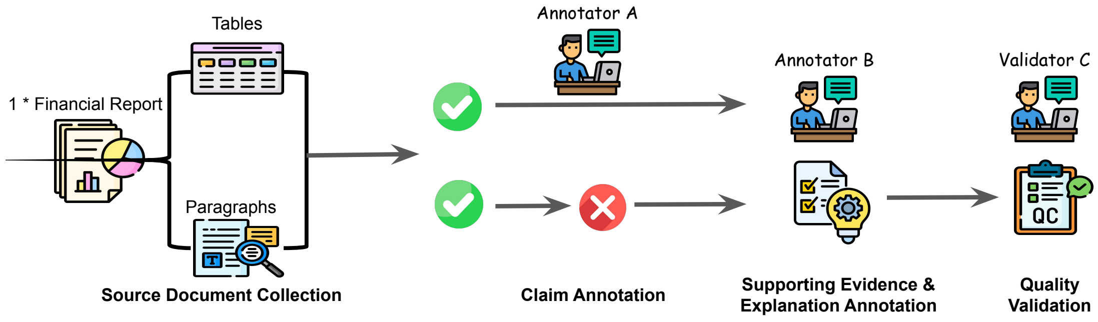

Figure 2: An overview of FINDVER construction pipeline.

3.2 FINDVER Subset Design

FINDVER is designed to mirror the real-world challenges encountered in the financial domain. Therefore, we ensure that the included annotators are financial experts with professional experience in comprehending and processing financial documents. Table 7 in Appendix presents the detailed annotator biographies for FINDVER annotation.

To identify the common reasoning-intensive scenarios in claim verification based on financial documents, we engaged with domain experts and conducted a preliminary study. This helped us determine three key types of scenarios that frequently arise in real-world settings. Accordingly, we have created three corresponding subsets of FINDVER. (1) **FDV-IE** (*information extraction*), which involves extracting information from both *textual* and *tabular* content within a *long-context* document.

(2) **FDV-MATH** (*numerical reasoning*), which necessitates performing *calculations* or *statistical analysis* based on data within the document.

(3) **FDV-KNOW** (*knowledge-intensive reasoning*), which requires integrating *external domain-specific knowledge* or *regulations* for claim verification.

3.3 Source Document Collection

Similar to Zhao et al. (2023a), we use the quarterly (Form 10-Q) and annual reports (Form 10-K) of companies as the source documents, which are publicly available in the open-source database[1] of the U.S. Securities and Exchange Commission. We collect a total of 523 documents that were first released between January 1 to April 30, 2024, which is after the cutoff date of most pretraining corpora for training foundation models. This helps to alleviate issues related to data memorization to some extent. After collecting the raw HTML-format documents, we utilize the SEC API[2], a commercial platform API for extracting financial document content, to process the collected documents, obtaining documents with both textual and tabular data.

3.4 Claim Annotation

Entailed Claim Annotation To address the potential bias concerning the position of evidence within the documents, we initiate the process by randomly sampling multiple document contexts from the given document. Annotators are then tasked with creating *"entailed"* claims based on the textual and tabular data within these contexts. The annotators are instructed to simulate real-world document comprehension scenarios, ensuring the annotated claims are representative of practical financial analysis and align with the scenarios defined by the corresponding subsets. Annotators are then tasked with carefully locating all evidence (*i.e.,* indices of relevant paragraphs and tables) within the entire document that support the claims, which are used for the subsequent data validation.

Refuted Claim Annotation Following established practices in the field (Wadden et al., 2020; Chen et al., 2020; Lu et al., 2023), and since directly obtaining *"refuted"* types is difficult, we instead perturb the original *"entailed"* claims into *"refuted"* claim through expert annotation. Specifically, expert annotators first create an *"entailed"* claim using the same procedure detailed in the "Entailed Claim Annotation" paragraph. The annotators are then instructed to perturb the *"entailed"* claim to introduce factual errors that are directly contradicted by the annotated evidence, and rewrite the annotated reasoning-process explanation.

[1] https://www.sec.gov/edgar/search/

[2] https://sec-api.io/

Annotation Quality	%S ≥ 4
Claim	
Fluency	92
Meaningfulness	90
Alignment with real-world scenarios	94
Evidence	
Relevancy	89
Completeness	85
Reasoning-process Explanation	
Fluency	95
Correctness	92
Comprehensiveness	90
Entailment Label	
Correctness	94

Table 2: Human evaluation over 100 samples from the FINDVER *testmini* set. Two internal evaluators were asked to rate the samples on a scale of 1 to 5 individually. We report percent of samples that have an average score ≥ 4 to indicate the annotation quality of FINDVER.

3.5 Explanation Annotation

After finishing the claim annotation, we pass it to another annotator for explanation annotation. The annotators are required to first read the claim carefully and annotate a detailed explanation of the reasoning process. Such reasoning-process explanations allow for a granular and informative evaluation of model outputs, helping future work identify reasoning errors and provide more accurate error feedback. We compare the entailment label annotated in this step with those in the claim annotation step. A third annotator is introduced if the two annotation versions are different. In practice, we achieve an inter-annotator agreement of 90.3% for entailment label annotation.

During our pilot annotation phase, we observed variability in the format of reasoning-process explanation annotated by different annotators, which made the dataset less standardized. To ensure consistency and clarity in our benchmark, we developed a predefined template for annotators to follow. Specifically, annotators are required to commence with the **extraction of relevant information** phase, where they need to list all claim-relevant information in a numbered list. Subsequently, they are required to annotate the **reasoning over the extracted information** segment in a step-by-step manner. For each step, they should elucidate the associated reasoning. Finally, they annotate the **entailment label** feature.

3.6 Data Quality Validation

To ensure the high quality of our annotated data, for every annotated example, a qualified annotator is assigned to validate several key aspects: (1) the claim and reasoning-process explanation should be grammatically correct and free of spelling errors; (2) the claim should be closely related to financial domains and meaningful in real-world scenarios; (3) the annotated evidence should be relevant to the claim and complete enough to verify it; (4) the entailment label of the claim should be supported by the annotated evidence; and (5) the reasoning-process explanation should correctly interpret the extracted evidence and apply appropriate reasoning steps to correctly verify the claim. The validators are asked to revise examples that do not meet these standards. In practice, 347 out of 2,100 initial examples were revised by the validators. We also report the human evaluation scores over 100 sampled examples. As illustrated in Table 2, FINDVER has a high annotation quality.

3.7 Dataset Preparation and Release

Table 3 provides an overview of the key statistics for our benchmark. The dataset is randomly split into two subsets: *testmini* and *test*. The *testmini* set is intended for model development and validation. It contains 600 examples, with 200 examples from each subset. The *test* set comprises the remaining 1,500 examples and is designed for standard evaluation. To prevent data contamination (Jacovi et al., 2023; Shi et al., 2024; Deng et al., 2024), the ground-truth-related annotation features for the *test* set will not be publicly released. Instead, we provide an online evaluation platform where researchers can assess their models and participate in a public leaderboard.

4 Experiment Setup

We next present the experimental setup, covering the evaluated LLMs, long-context and RAG setups, implementation details, and the measurement of human-level performance.

4.1 Experimented LLMs

We examine the performance of LLMs across two distinct categories on FINDVER: (1) **Proprietary LLMs**, including GPT-4* (OpenAI, 2023a,b, 2024), Gemini-1.5-* (Gemini, 2024), and Claude-3 (Anthropic, 2024); and (2) **Open-source LLMs**, including Gemma (Team et al., 2024), Llama-

Property	FDV-IE	FDV-MATH	FDV-KNOW
Real-world scenarios in financial domains	information extraction	numerical reasoning	knowledge-intensive reasoning
# Document	221	225	217
Doc Length (*i.e.,* word count) `(Median/Avg/Max)`	42K / 41K / 71K	43K / 41K / 71K	43K / 41K / 71K
# Tables per document `(Median/Avg)`	62 / 78.9	62 / 79.1	62 / 79.0
Claim length `(Median/Avg)`	47 / 47.2	24 / 25.1	36 / 37.1
# Text evidence per claim `(Median/Avg)`	2 / 1.8	1 / 1.3	3 / 2.6
# Table evidence per claim `(Median/Avg)`	1 / 1.0	1 / 0.9	1 / 1.2
% Claims *w.* table evidence	66.3%	71.1%	70.8%
Explanation length `(Median/Avg)`	70 / 73.1	74 / 76.2	96 / 100.7
Benchmark size (# Claims)			
testmini size	200	200	200
test size	500	500	500

Table 3: Basic statistics of the FINDVER benchmark.

Adopted Chain-of-Thought Prompt

[System Input]
As a financial expert, your task is to assess the truthfulness of the given claim by determining whether it is entailed or refuted based on the provided financial document. Follow these steps:
1. Carefully read the given context and the claim.
2. Analyze the document, focusing on the relevant financial data or facts that related to the claim.
3. Document each step of your reasoning process to ensure your assessment is clear and thorough.
4. Conclude your analysis with a final determination. In your last sentence, clearly state your conclusion in the following format: "Therefore, the claim is {entailment_label}." Replace {entailment_label} with either 'entailed' (if the claim is supported by the document) or 'refuted' (if the claim contradicts or partially contradicts the document).

[User Input]
Financial Report:
{Financial Report}

Claim to verify:
{Claim}

Follow the instructions and think step by step to verify the claim.

Figure 3: The Chain-of-Thought prompt used.

2&3 (Touvron et al., 2023a; Meta, 2024), Yi-1.5 (AI et al., 2024), Qwen-2 (qwe, 2024), Mistral & Mixtral (Jiang et al., 2023a, 2024), InternLM2 (Team, 2024), C4AI (Aryabumi et al., 2024), GLM (Du et al., 2022), and Phi-3 (Abdin et al., 2024). Table 8 in Appendix presents the details of evaluated models (*i.e.,* organizations, release time, max context length, and model version).

The experiments for open-sourced LLMs were conducted using the `vLLM` framework (Kwon et al., 2023). For all the experiments, we set temperature as 1.0 and maximum output length as 512. We adopt the *Chain-of-Thought (CoT)* prompting methods (Wei et al., 2022) for the FINDVER benchmark. Specifically, the model is instructed to first output a detailed reasoning process for verifying claims, and then provide the entailment label of the claim based on the generated reasoning process. Figure 3 presents the used prompts.

4.2 Long-Context and RAG Settings

As presented in Table 3, the documents within our benchmark are notably lengthy. To effectively handle this, we have implemented two real-world application settings that are widely recognized for their utility in dealing with extensive texts. For **Long-context Setting**, we input the entire financial document into the model. We limit our evaluation to those models that have a context window larger than 100,000 tokens, which exeeds the maximum length of the included financial document. For **RAG Setting**, we leverage the current best-performing embedding models (*i.e.,* OpenAI's `text-embedding-3-large`) to retrieve the top-10 paragraphs or tables that are most relevant to the claims. These elements are then concatenated in their original order as found in the document before being fed into the model.

4.3 Implementation Details

Input Tabular Data Serialization Building on previous research that assessed LLMs on tasks involving tabular data (Chen, 2023; Zhao et al., 2023b,c), we introduce our methodology for processing tables within documents. Our approach

Model	Notes	FDV-IE		FDV-MATH		FDV-KNOW		Average	
		LongC	RAG	LongC	RAG	LongC	RAG	LongC	RAG
Random Choice		50.0		50.0		50.0		50.0	
Human Non-Expert		90.0		85.0		85.0		86.7	
Human Expert		95.0		90.0		95.0		93.3	
Open-source LLMs									
InternLM2-Math-7b	Math	–	58.0	–	53.5	–	54.5	–	55.3
InternLM2-7B		–	59.5	–	54.5	–	56.0	–	56.7
Gemma-7B		–	59.5	–	55.5	–	55.0	–	56.7
GLM-4-9b		58.5	61.0	54.5	54.5	55.0	56.5	56.0	57.3
Llama-2-7B		–	60.0	–	56.5	–	56.5	–	57.7
Mistral-7B-v3		–	59.5	–	56.5	–	57.0	–	57.7
Phi-3-medium-4k		–	61.5	–	54.0	–	58.0	–	57.8
Llama-2-70B		–	61.5	–	54.5	–	58.0	–	58.0
Phi-3-medium-128k		58.0	61.5	54.0	55.5	56.5	57.5	56.2	58.2
Meta-Llama-3-8B		–	62.5	–	55.0	–	59.5	–	59.0
Yi-1.5-34B		–	62.5	–	58.0	–	58.0	–	59.5
Meta-Llama-3-70B	MoE	–	65.5	–	61.5	–	61.5	–	62.8
C4AI Command R+		–	67.5	–	60.0	–	64.5	–	64.0
Qwen2-72B		67.0	68.0	<u>62.5</u>	61.5	60.5	65.0	63.3	64.8
Mixtral-8x22B		–	70.0	–	62.0	–	67.0	–	66.3
Proprietary LLMs									
Gemini-1.5-Flash		<u>71.0</u>	70.5	<u>62.5</u>	60.5	65.0	65.5	<u>66.2</u>	65.5
GPT-3.5-turbo		–	79.0	–	64.0	–	70.5	–	71.2
GPT-4o		<u>80.0</u>	78.5	<u>70.5</u>	68.0	<u>76.5</u>	74.5	<u>75.7</u>	73.7
Claude-3.5-Sonnet		<u>**83.5**</u>	80.5	<u>**71.0**</u>	69.0	<u>**77.0**</u>	75.5	<u>**77.2**</u>	75.0

Table 4: Accuracy of entailment classification on the *testmini* set of FINDVER. We report results for LLMs with *CoT* prompting under the *long-context* (LongC) and *RAG* settings. <u>Numbers</u> underscored indicate that models under the long-context setting achieve better results than under the RAG setting.

involves distinguishing headers and cells in different columns using a vertical bar (|) and separating rows with new lines. This format allows us to input flattened table data directly into LLMs. In our initial experiments, we found that LLMs such as GPT-* and Llama-3 can effectively interpret this table representation. However, we suggest that future studies should investigate more sophisticated methods for encoding tabular data to enhance comprehension by LLMs.

Model Response Processing Following previous work (Lu et al., 2024), we adopt LLM for processing model response. Specifically, we utilize GPT-4o-mini to extract labels from the LLM output, which can be either *"entailed"*, *"refuted"* or *"none"*. The *"none"* label typically indicates that the LLM output contains nonsensical symbols or unintelligible text rather than meaningful content. In cases where the output is labeled as *"none"*, we assign the final label by making a random guess.

4.4 Human-level Performance

To provide a rough but informative estimate of human-level performance by non-experts and experts on FINDVER, we randomly sampled 5 documents $\times$ 4 claims / document = 20 claims from each validation subset, totaling 60 claims. We enroll two experts (*i.e.,* professionals with CFA license) and two non-experts (*i.e.,* undergraduate students majored in computer science) to individually verify the claims by providing the NL explanations. Table 4 presents the human-level performance.

5 Experiment Results

5.1 Main Findings

Table 4 and Table 5 display the primary results for FINDVER. We reveals a significant accuracy gap between human experts and LLMs. Notably,

Model	IE	MATH	KNOW	Avg
Human Non-Expert	90.0	85.0	85.0	86.7
Human Expert	95.0	90.0	95.0	93.3
Open-source LLMs				
InternLM2-Math-7B	57.0	53.8	54.6	55.1
InternLM2-7B	59.6	53.4	55.4	56.1
Gemma-7B	59.8	54.4	55.0	56.4
GLM-4-9B	61.4	54.2	57.4	57.7
Llama-2-7B	61.0	57.2	57.2	58.5
Mistral-7B-v3	59.8	56.0	56.4	57.4
Phi-3-medium-4k	61.8	54.0	57.2	57.7
Llama-2-70B	60.6	53.8	58.0	57.5
Phi-3-medium-128k	61.2	55.4	58.2	58.3
Meta-Llama-3-8B	63.4	55.0	60.2	59.5
Yi-1.5-34B	62.8	57.2	56.8	58.9
Meta-Llama-3-70B	65.0	61.2	60.4	62.2
C4AI Command R+	67.4	59.0	65.4	63.9
Qwen2-72B	69.0	62.2	65.2	65.5
Mixtral-8x22B	71.0	62.8	68.2	67.3
Proprietary LLMs				
Gemini-1.5-Flash	71.2	61.0	65.8	66.0
GPT-3.5-turbo	79.6	65.2	70.8	71.9
GPT-4o	78.6	**69.4**	73.8	73.9
Claude-3.5-Sonnet	**81.0**	68.2	**74.0**	**74.4**

Table 5: Accuracy of entailment classification on the FINDVER *test* set. We report results for LLMs with *CoT* prompting under the *RAG* setting. Due to computation constraint, we did not evaluate the long-context setting.

Adopted Chain-of-Thought Prompt

[System Input]
As a financial expert, your task is to assess the truthfulness of the given statement by determining whether it is entailed or refuted based on the provided financial document. You should directly output the entailment label ('entailed' or 'refuted') without any intermediate steps.

[User Input]
Financial Report:
{Financial Report}

Claim to verify:
{Claim}

Directly output the entailment label ('entailed' or 'refuted') of the claim.

Figure 4: The *Direct Output* prompt used.

Model	**w/o CoT**		**w/ CoT**	
	LongC	RAG	LongC	RAG
Qwen2-72B	57.7 (-5.6)	59.5 (-5.3)	63.3	64.8
GPT-4o	70.1 (-7.1)	69.5 (-5.5)	77.2	75.0

Table 6: Accuracy of entailment for GPT-4o and Qwen2-72B with and without CoT Prompting methods on the FINDVER *testmini* set.

Claude-3.5-Sonnet, the highest-performing LLM, achieves an accuracy rate of only 77.2%, in stark contrast to the 93.3% accuracy of financial experts. This discrepancy highlights the complexity and challenges of our benchmark.

For the less competitive LLMs, such as Qwen2-72B, GLM-4-9B, and Phi-3-medium-128k, they exhibit improved performance under the RAG setting. In contrast, the currently more competitive LLMs, such as GPT-4o and Claude-3.5-Sonnet, generally perform better under the long-context setting compared to the RAG setting. This indicates the potential of developing long-context techniques to manage tasks involving extensive documents in specialized domains.

5.2 Chain-of-Thought Analysis

To better understand the effectiveness of CoT prompting methods for our tasks, we select the commonly-used proprietary and open-source LLMs, GPT-4o and Qwen2-72B, for our experiments. In the w/o CoT setting, we instruct the LLMs to directly output the entailment label of the claim using the provided document context (Figure 4). As illustrated in Table 6, both LLMs' performance degrades in the w/o CoT setting. These results highlight the importance of CoT reasoning in enhancing performance for our task.

5.3 Error Analysis of Reasoning Process

The Claude-3.5-Sonnet model achieves a top accuracy of 77.2% under the long context setting. To better understand the model's limitations, we perform a detailed error analysis with human evaluators. We randomly select 25 instances from each of the three subsets where the Claude-3.5-Sonnet model fails to perform correctly. Our analysis has identified four primary categories of errors: (1) *Extraction error*: The model fails to correctly retrieve relevant information from the context, resulting in inaccurate verification. (2) *Numerical reasoning error*: The model encounters difficulties with correct mathematical reasoning. (3) *Domain knowledge deficiency*: The model lacks sufficient knowledge in finance-related areas, which hampers its ability to reason accurately. (4) *Computation error*: While the model's reasoning is correct, it makes computational mistakes during intermediate or final steps, resulting in incorrect verification.

6 Conclusion

This paper presents FINDVER, a comprehensive benchmark designed to evaluate LLMs in claim verification over long and hybrid-content financial documents. Through extensive experiments involving 19 LLMs under long-context and RAG settings, we have demonstrated that even the top-performing models exhibit a significant performance gap compared to financial experts. Our detailed findings and insights reveal the strengths and limitations of current LLMs in this new task. We believe that FINDVER provides a valuable benchmark for future research on LLMs' ability to handle complex claim verification tasks within the expert domain.

Limitations

In this work, we propose FINDVER and conduct comprehensive analysis of different LLMs' capabilities on our task. However, there are still some limitations: First, our evaluation does not include recently released finance-specific LLMs (Wu et al., 2023; Yang et al., 2023; Xie et al., 2023, 2024), as these models are not yet compatible with the vLLM framework used for inference. Due to computational resource constraints, we do not tune LLMs on a large-scale finance-domain data ourselves. However, we believe that training on finance data can help improve LLMs' capabilities in FINDVER. Moreover, we only conduct human error analysis on the generated reasoning process of models. We believe future work could explore the development of LLM-based automated evaluation systems (Liu et al., 2023; Zheng et al., 2023) for automatically detecting reasoning errors within the generated explanation.

References

2024. Qwen2 technical report.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Anthropic. 2024. Introducing the next generation of claude.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 7352–7364, Online. Association for Computational Linguistics.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022a. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

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. 2022b. 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.

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024. Unveiling the spectrum of data contamination in language model: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16078–16092, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023b. Mistral 7b. *ArXiv*, abs/2310.06825.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. Bizbench: A quantitative reasoning benchmark for business and finance.

Yuta Koreeda, Manning, and Christopher. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Long text and multi-table summarization: Dataset and method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1995–2010, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. Ex-fever: A dataset for multi-hop explainable fact verification.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECT-Sum: 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.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023a. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2023b. GPT-4V(ision) system card.

OpenAI. 2024. Hello gpt-4o.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2024. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

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.

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

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

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.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam

Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

InternLM Team. 2024. Internlm2 technical report.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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. 2023a. Llama 2: Open foundation and fine-tuned chat models.

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

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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*, CIKM '18, page 1627–1630, New York, NY, USA. Association for Computing Machinery.

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*, abs/2303.17564.

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.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, Shunian Chen, Yifei Zhang, Lihang Shen, Daniel Kim, Zhiwei Liu, Zheheng Luo, Yangyang Yu, Yupeng Cao, Zhiyang Deng, Zhiyuan Yao, Haohang Li, Duanyu Feng, Yongfu Dai, VijayaSai Somasundaram, Peng Lu, Yilun Zhao, Yitao Long, Guojun Xiong, Kaleb Smith, Honghai Yu, Yanzhao Lai, Min Peng, Jianyun Nie, Jordan W. Suchow, Xiao-Yang Liu, Benyou Wang, Alejandro Lopez-Lira, Jimin Huang, and Sophia Ananiadou. 2024. Open-finllms: Open multimodal large language models for financial applications.

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, Melbourne, Australia. Association for Computational Linguistics.

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

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024a. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024b. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Hongjun Liu, Linyong Nan, Lyuhao Chen, Ryo Kamoi, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2023a. Docmath-eval: Evaluating numerical reasoning capabilities of llms in understanding long documents with tabular data.

Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Ruizhe Chen, Xiangru Tang, Yumo Xu, Dragomir Radev, and Arman Cohan. 2023b. QTSumm: Query-focused summarization over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1157–1172, Singapore. Association for Computational Linguistics.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023c. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. 2023d. RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6081, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

ID	Finance Industry Experience	English Proficiency	Annotation Sets	Evaluator?
1	1 working and 1 internship at US	Native speaker	FDV-KNOW	✓
2	>= 2 internship at US	> 15 years	FDV-MATH	✓
3	1 working at Singapore and 2 internship at US	Native speaker	FDV-KNOW	✓
4	2 working and >= 1 internship at US	Native speaker	FDV-KNOW	✗
5	1 internship at US, 2 internship at China	10 years	FDV-IE	✗
6	1 internships at HK, China	15 years	FDV-IE, FDV-MATH	✓
7	1 internships at China	10 years	FDV-IE, FDV-MATH	✗

Table 7: Details of annotators involved in dataset construction. FINDVER is annotated by financial professionals with extensive experience in comprehending financial documents, ensuring it accurately reflects the real-world challenges in the financial domain.

Model	Organization	Release Time	Max Length	Source
GPT-4o (OpenAI, 2023a)	OpenAI	2023-03	128k	`https://platform.openai.com/`
GPT-3.5-turbo (OpenAI, 2022)	OpenAI	2022-11	16k	`https://platform.openai.com/`
Gemini-1.5-* (Gemini, 2024)	Google	2024-02	128k	`https://ai.google.dev/`
Claude-3.5 (Anthropic, 2024)	Anthropic	2024-03	200k	`https://www.anthropic.com/api`
Gemma (Team et al., 2024)	Google	2024-02	8k	google/gemma-7b-it
Llama-2 (Touvron et al., 2023a)	Meta	2023-02	4k	meta-llama/Llama-2-*-chat-hf
Llama-3 (Meta, 2024)	Meta	2024-04	8k	meta-llama/Meta-Llama-3-*-Instruct
Yi-1.5 (AI et al., 2024)	01-ai	2024-05	32k	01-ai/Yi-1.5-*-Chat
Qwen-2 (qwe, 2024)	Qwen	2024-06	128k	Qwen/Qwen2-*-Instruct
Mistral (Jiang et al., 2024, 2023a)	Mistral AI	2024-05	32k	mistralai/Mistral-7B-Instruct-v0.3
Mixtral (Jiang et al., 2024, 2023a)	Mistral AI	2024-04	64k	mistralai/Mixtral-8x22B-v0.1
InternLM2 (Team, 2024)	internlm	2024-01	200k	internlm/internlm2-chat-*
GLM (Du et al., 2022)	THUDM	2024-06	128k	THUDM/glm-4-9b-chat
Phi-3 (Abdin et al., 2024)	microsoft	2024-04	128k	microsoft/Phi-3-*-instruct

Table 8: Details of the organization, release time, maximum context length, and model source (*i.e.,* url for proprietary models and Huggingface model name for open-source models) for the LLMs evaluated in FINDVER.