

FINDVER: Explainable Claim Verification over Long and Hybrid-content Financial Documents

Anonymous ACL submission

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 contains 4,000 expert-annotated examples across four subsets, each focusing on a type of scenario that frequently arises in real-world financial domains. We assess a broad spectrum of 25 LLMs under long-context and RAG settings. Our results show that even the current best-performing system (i.e., GPT-4o) 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.

1 Introduction

In today’s era of information explosion, 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, 2023a; Koncel-Kedziorski et al., 2024). Extracting and analyzing claim-relevant data from these documents re-

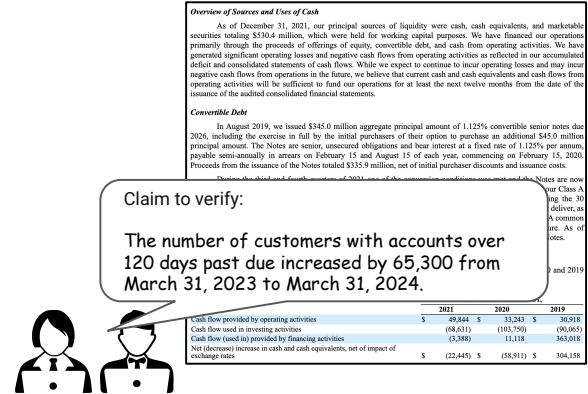


Figure 1: An example from the *numerical reasoning* subset of the FINDVER benchmark. To verify the claim, the LLM is required to first locate claim-relevant data points within long and hybrid-content financial documents, and then apply numerical reasoning over the extracted data points for claim verification.

quires 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, we present **FINDVER**¹, a comprehensive and do-

¹The data and code for this study can be found at <https://anonymous.4open.science/r/FinDVer-B6B0>, and will be released publicly upon publication.

Dataset	Input Context	Annotation / Data Creation	# Label	w. Expla- nation?	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 summarization datasets	2	✗	✗
ContractNLI (Koreeda et al., 2021)	Contract	Expert & Crowd-sourced	2	✗	✗
LLM-AGGREGFACT (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	3	✓	✓

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 four focused subsets, each tailored to mirror distinct real-world financial analysis scenarios.

main 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*”, “*refuted*”, or “*not enough information*”, based solely 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 four key types of scenarios that frequently arise in real-world settings: *information extraction*, *numerical reasoning*, *knowledge-intensive reasoning*, *comparative analysis*. For each scenario, we construct a corresponding 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, 25 models from 13 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 re-

sults indicate that even the existing best-performing LLM (*i.e.*, GPT-4o) still significantly lags behind human experts (59.2% versus 85.0%), 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

Property	FDV-IE	FDV-MATH	FDV-KNOW	FDV-COMPA
Real-world scenarios in financial domains	information extraction	numerical reasoning	knowledge-intensive reasoning	comparative analysis on two companies
# Document	243	251	254	283
Doc Length (<i>i.e.</i> , word count) (Median/Avg/Max)	53K / 54K / 62K	57K / 55K / 63K	51K / 54K / 60K	68K / 65K / 77K
# Tables per document (Median/Avg)	45 / 74.1	49.5 / 52.1	46 / 48.4	61 / 62.7
% Entailed claims	36%	36%	35%	36%
Claim length (Median/Avg)	23 / 24.0	22 / 26.0	35 / 38.7	41 / 42.9
% Refuted claims	36%	35%	35%	35%
Claim length (Median/Avg)	25.5 / 26.4	23 / 25.1	39 / 41.2	44.5 / 45.8
% Not Enough Information claims	28%	29%	30%	29%
Claim length (Median/Avg)	22 / 24.4	26 / 28.9	39.5 / 42	44 / 46.1
# Text evidence per claim (Median/Avg)	2 / 2.3	2 / 1.9	2 / 1.7	3 / 2.6
# Table evidence per claim (Median/Avg)	1 / 0.7	2 / 2.3	2 / 1.9	2 / 2.4
% Claims w. table evidence	85.5%	96.3%	93.1%	96%
Explanation length (Median/Avg)	64 / 68.8	102 / 110.3	98 / 100.1	113 / 115.2
# Reasoning steps (Median/Avg)	3 / 3.2	5 / 5.4	5 / 4.8	7 / 6.7
Benchmark size (# Claims)				
testmini size	250	250	250	250
test size	750	750	750	750

Table 2: Basic statistics of the FINDVER benchmark.

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 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; Zhao et al., 2023a; Koncel-Kedziorski et al., 2024). Despite these advancements, there remains a significant gap in the exploration of claim verification tasks within the financial domain. Verifying the accuracy of claims derived from financial documents is essential, as it directly impacts investment decisions and market perceptions. To address this gap, we propose the first claim verification benchmark specifically designed for real-world scenarios in the financial domain, focusing on the comprehension of financial documents.

3 FINDVER Benchmark

FINDVER provides a robust evaluation benchmark for reasoning-intensive and explainable claim verification over long and hybrid-content financial documents. Table 2 presents the basic statistics of our benchmark. We present an overview of the FIND-

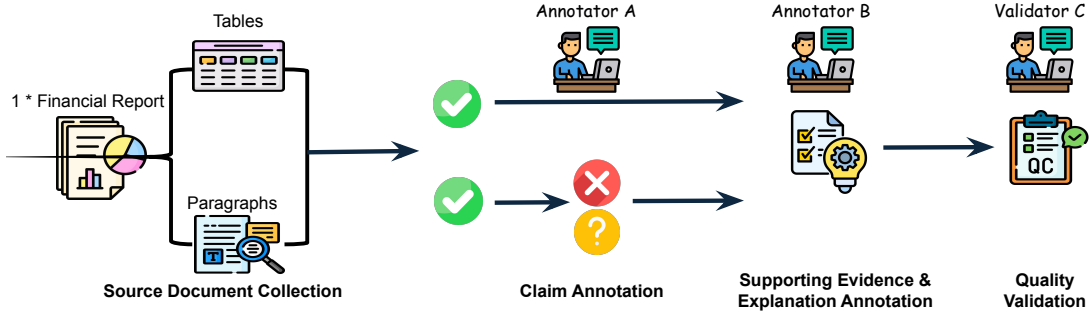


Figure 2: An overview of FINDVER construction pipeline.

VER 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 $\ell \in \mathcal{L} = \{\text{“entailed”}, \text{“refuted”}, \text{“not enough information”}\}$, based on the document context:

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

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

$$e = \arg \max_e P_{\text{LLM}}(e \mid P, T, c) \quad (2)$$

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.

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 6 in Appendix A 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 four key types of scenarios that frequently arise in real-world settings. Accordingly, we have created four corresponding subsets of FINDVER. Appendix A presents examples for each subset.

- (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.
- (4) **FDV-COMPA** (*comparative analysis*), which involves detailed examination of documents from two different companies, comparing financial metrics, performance indicators, and other critical data.

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 database² of the U.S. Securities and Exchange Commission. We collect a total of 1,537 documents that were first released between January 1 to March 15, 2024, which

²<https://www.sec.gov/edgar/search/>

Annotation Quality	%S \geq 4
Claim	
Fluency	92.5
Meaningfulness	90.0
Alignment with real-world scenarios	94.5
Evidence	
Relevancy	89.5
Completeness	85.0
Reasoning-process Explanation	
Fluency	95.5
Correctness	91.0
Comprehensiveness	90.5
Entailment Label	
Correctness	96.5

Table 3: Human evaluation over 200 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.

is after the cutoff date of most pretraining corpora used for training foundation models. This helps to alleviate issues related to data memorization to some extent. We detail the process of source document processing in the Appendix A. We regard one paragraph or one table within the processed document as a single piece of candidate supporting evidence, which is used for evidence annotation.

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 and Not Enough Information 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*” or “*not enough information*” types is difficult, we

instead perturb the original “*entailed*” claims into claims of “*refuted*” and “*not enough information*” types through expert annotation. Specifically, expert annotators first create an “*entailed*” claim using the same procedure detailed in the “Entailed Claim Annotation” paragraph.

For annotating “*refuted*” claims, the annotators are 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 accordingly.

For annotating “*not enough information*” claims, during our pilot annotation, we found that GPT-4o could include more diverse information compared to human annotators, thereby reducing dataset artifacts. Therefore, we first utilize GPT-4o to generate four candidate claims by adding additional relevant information to the claims without providing any document context. This additional information is intentionally outside the scope of the document context, making it impossible to verify or refute the new claim based on the document context. Expert annotators are then assigned to critically review, rewrite, and finalize one of the candidate claims for inclusion in the final dataset. It is worth noting that the “*not enough information*” candidate claims generated by GPT-4o may actually be entailed or refuted by the document context. Therefore, we require annotators to re-evaluate the relevant evidence and explanations by thoroughly reviewing the entire document.

3.5 Supporting Evidence and Explanation Annotation

After finishing the claim annotation, we pass it to another annotator for supporting evidence and explanation annotation. The annotators are required to first read the claim carefully and annotate all relevant paragraphs and tables within the document as supporting evidence. They are then required to 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 supporting evidence and 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 96.0% 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 mandated 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. Figure 3 in Appendix presents an example with the step-by-step explanation annotation.

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, 532 out of 4,000 initial examples were revised by the validators. We also report the human evaluation scores over 200 sampled examples. As illustrated in Table 3, FINDVER has a high annotation quality.

3.7 Dataset Preparation and Release

We randomly divide the dataset into two subsets: *testmini* and *test*. The *testmini* set is intended for model development and validation. It contains 1,000 examples, with 250 examples from each subset. The *test* set comprises the remaining 3,000 examples and is designed for standard evaluation. To prevent data contamination (Jacovi et al., 2023; Shi et al., 2024), the annotation features related to ground-truth will not be publicly released for the *test* set. Instead, we will maintain an online evaluation platform, allowing researchers to evaluate their models and participate in a public leaderboard.

4 Experiment Setup

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-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), AdaptLLM-Finance (Cheng et al., 2024) and Phi-3 (Abdin et al., 2024). Table 7 in Appendix presents the details of evaluated models (*i.e.*, organizations, release time, max context length, and model version used).

The implementation details, including LLM parameter setting, tabular data serialization, and LLM response processing are discussed in Appendix B.1. 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 4 in Appendix B presents the used prompts.

4.2 Long-Context and RAG Settings

As presented in Table 2, 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 128,000 tokens, which exceeds 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 Human-level Performance

To provide a rough but informative estimate of human-level performance by non-experts and experts on FINDVER, we randomly sampled 2 docu-

Model	Notes	FDV-IE		FDV-MATH		FDV-KNOW		FDV-COMPA		Average	
		LongC	RAG	LongC	RAG	LongC	RAG	LongC	RAG	LongC	RAG
Baselines											
Random Choice		33.3		33.3		33.3		33.3		33.3	
Frequent Choice		36.4		36.4		36.0		35.6		36.1	
Claim only											
GPT-4o		40.8		36.4		38.0		36.8		38.0	
Llama-3-70B		39.2		38.0		36.8		39.2		38.3	
Human Non-Expert		90.0		82.5		70.0		77.5		80.0	
Human Expert		87.5		87.5		80.0		85.0		85.0	
Open-source LLMs											
AdaptLLM-Finance-7B	Llama-2-7B	–	46.4	–	36.8	–	42.4	–	37.6	–	40.8
Llama-2-7B		–	48.8	–	41.6	–	40.0	–	38.8	–	42.3
InternLM2-math-plus-7b	Math	–	48.8	–	42.4	–	38.4	–	39.6	–	42.3
InternLM2-7B		–	49.2	–	44.0	–	41.6	–	38.4	–	43.3
Llama-2-70B		–	49.6	–	40.0	–	44.4	–	39.6	–	43.4
Phi-3-medium-128k		51.2	54.0	36.4	37.2	43.6	44.4	39.6	40.8	42.7	44.1
Gemma-7B		–	49.6	–	48.4	–	42.0	–	36.8	–	44.2
Phi-3-medium-4k		–	54.0	–	38.4	–	43.6	–	43.2	–	44.8
Meta-Llama-3-8B		–	59.6	–	40.0	–	40.8	–	40.0	–	45.1
GLM-4-9b		50.4	54.4	42.0	41.2	43.2	44.0	41.2	45.6	44.2	46.3
Mistral-7B-v3		–	52.0	–	45.6	–	48.4	–	41.2	–	46.8
Yi-1.5-34B		–	56.8	–	47.6	–	45.2	–	42.0	–	47.9
C4AI Command R+		–	59.6	–	48.4	–	45.6	–	40.4	–	48.5
Qwen2-57B-A14B	MoE	–	59.6	–	50.4	–	46.8	–	38.8	–	48.9
Meta-Llama-3-70B	MoE	–	57.6	–	48.0	–	45.2	–	45.6	–	49.1
Qwen2-72B		57.6	60.4	43.6	49.2	40.8	45.2	42.4	44.8	46.1	49.9
Mixtral-8x22B		–	64.8	–	48.8	–	48.4	–	43.2	–	51.3
Proprietary LLMs											
Gemini-1.5-flash		57.2	60.8	42.4	45.6	42.8	43.6	45.2	44.0	46.9	48.5
Claude-3-haiku		64.8	64.8	45.6	50.0	42.4	47.2	44.8	43.2	49.4	51.3
GPT-3.5-turbo-0125		–	69.6	–	53.2	–	43.2	–	43.2	–	52.3
Gemini-1.5-pro		67.2	68.8	53.2	54.8	42.4	40.8	47.2	46.0	52.5	52.6
Claude-3-sonnet		64.8	66.8	46.4	48.0	48.0	49.2	48.0	48.8	51.8	53.2
Claude-3-opus		72.0	70.8	57.2	56.4	48.0	45.6	50.4	45.6	56.9	54.6
GPT-4-turbo		–	68.4	–	54.8	–	53.2	–	58.0	–	58.6
GPT-4o		67.6	68.8	56.4	55.2	51.6	52.4	61.2	56.8	59.2	58.3

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. Numbers underscored indicate that models under the long-context setting achieve better results than under the RAG setting. We also highlight the best performance in each setting for each subset with red and blue color.

ments \times 5 claims / document = 10 claims from each validation subset, totaling 40 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 Experimental Results

5.1 Main Results

Table 4 displays the primary results for FINDVER. We derive the following key findings:

FINDVER is a high-quality and challenging claim verification benchmark for state-of-the-art LLMs Table 4 reveals a significant accuracy

gap between human experts and LLMs. Notably, GPT-4o, the highest-performing LLM, achieves an accuracy rate of only 59.2%, in stark contrast to the 85% accuracy of financial experts. This discrepancy highlights the complexity and challenges of our benchmark. Furthermore, when only the claim is provided without the context, LLMs’ accuracy drops to nearly random levels. This suggests that the FINDVER benchmark contains minimal artifacts and data contamination, affirming that it effectively evaluates the true capabilities of the models in claim verification.

Comparing RAG and Long-context Settings

For the less competitive LLMs, such as Qwen2-72B, GLM-4-9B, and Phi-3-medium-128k, they exhibit improved performance under the RAG set-

		R @ K	GPT-3.5	Llama-3-70B
$k = 5$	BM25	0.28	37.8 (-7.1)	38.6 (-4.6)
	Contriever	0.34	40.2 (-7.7)	39.8 (-6.0)
	GPT-embed	0.50	48.7 (-3.6)	45.3 (-3.8)
$k = 10$	BM25	0.42	44.9	43.2
	Contriever	0.52	49.9	45.8
	GPT-embed	0.66	52.3	49.1
$k = 15$	BM25	0.53	50.1 (+5.2)	47.6 (+4.4)
	Contriever	0.63	52.5 (+2.6)	48.0 (+2.2)
	GPT-embed	0.69	52.9 (+0.6)	49.3 (+0.2)
Oracle	–	–	73.2 (+20.9)	68.4 (+18.7)

Table 5: Accuracy of entailment classification for GPT-3.5 and Llama-3-70B under various RAG settings.

ting. In contrast, the currently more competitive LLMs, such as GPT-4 and Claude-Opus, 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.

Comparing performance of domain-specific LLMs against their backbones In terms of domain-specific LLMs, AdaptLLM-Finance-7B outperforms its backbone (*i.e.*, Llama-2-7B) on the FDV-KNOW set; InternLM2-Math-Plus-7B outperforms its backbone (*i.e.*, InternLM2-7B) on the FDV-MATH set. However, their overall performance is worse than that of their respective backbones. These observations suggest that while domain-specific training can enhance performance on targeted tasks, achieving consistent excellence across a varied dataset remains a challenge, highlighting the need for advancements in domain adaptation strategies.

5.2 RAG Setting Analysis

This subsection explores the impact of evidence extraction accuracy on the overall performance of LLMs. We assess LLM performance using three different retrieval methods—BM25 (Robertson et al., 1995), Contriever (Izacard et al., 2022), and OpenAI’s text-embedding-3-large—across three retrieval sizes ($k = 5, 10, 15$). As shown in Table 5, providing higher-quality evidence generally leads to improved performance under the RAG setting. This suggests that the effectiveness of LLMs in generating responses can be significantly enhanced by optimizing the accuracy of the underlying evidence retrieval mechanisms.

5.3 Chain-of-Thought Analysis

To better understand the effectiveness of CoT prompting methods for our tasks, we select the best-performing 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. As illustrated in Table 8 in Appendix B, both LLMs’ performance degrades in the w/o CoT setting. This degradation is especially significant in long-context settings, where models without CoT achieve accuracy levels close to random guessing. This demonstrates the effectiveness of CoT methods for our task.

5.4 Error Analysis of Reasoning Process

The GPT-4o model achieves a top accuracy of 59.2% 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 four subsets where the GPT-4o model fails to perform correctly. Our analysis has identified two primary error categories pervasive across all subsets: *label confusion* and *reasoning errors*. We also pinpoint four specific types of errors, each predominating in a particular subset: *extraction errors*, *numerical reasoning errors*, *domain knowledge deficiencies*, and *comparison errors*. For each type of error, illustrative examples and corresponding error analysis are provided in Appendix C. We believe that such error analysis could provide insights into potential areas for improving LLM-based system performance on our task.

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 25 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, among recently released finance-specific LLMs (Wu et al., 2023; Yang et al., 2023; Xie et al., 2023), we only evaluate AdaptLLM-Finance-7B, as it is the only work with a checkpoint available at HuggingFace and compatible with the vllm framework. 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 Norrick, 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 Santacrose, 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, Xiyu Zhou, and

642	William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification . In <i>International Conference on Learning Representations</i> .	699
643		700
644		701
645	Zhiyu Chen, Wenhui 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 <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	702
646		703
647		
648		704
649		705
650		706
651		707
652		708
653		709
654		710
655	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 <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	711
656		712
657		713
658		714
659		
660		715
661		716
662		717
663	Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension . In <i>The Twelfth International Conference on Learning Representations</i> .	718
664		719
665		720
666		721
667		722
668	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 <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 320–335, Dublin, Ireland. Association for Computational Linguistics.	723
669		724
670		725
671		726
672		727
673		728
674		729
675	Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context .	
676		
677	Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence . <i>Transactions of the Association for Computational Linguistics</i> , 11:1–18.	730
678		731
679		732
680		733
681	Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikrumar. 2020. INFOTABS: Inference on tables as semi-structured data . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 2309–2324, Online. Association for Computational Linguistics.	734
682		735
683		
684		736
685		737
686		738
687		739
688	Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning . <i>Transactions on Machine Learning Research</i> .	740
689		741
690		742
691		743
692		744
693	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 <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 5075–5084, Singapore. Association for Computational Linguistics.	745
694		746
695		747
696		748
697		749
698		750
		751
		752
		753
		754
		755

756	for Computational Linguistics: EMNLP 2022, pages	OpenAI. 2023b. GPT-4V(ision) system card .	812
757	1995–2010, Abu Dhabi, United Arab Emirates. Asso-	OpenAI. 2024. Hello gpt-4o .	813
758	ciation for Computational Linguistics.		
759	Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang,	Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu.	814
760	Ruochen Xu, and Chenguang Zhu. 2023. G-eval:	2024. Fin-fact: A benchmark dataset for multimodal	815
761	NLG evaluation using gpt-4 with better human align-	financial fact checking and explanation generation .	816
762	ment . In <i>Proceedings of the 2023 Conference on</i>		
763	<i>Empirical Methods in Natural Language Processing</i> ,	Stephen E Robertson, Steve Walker, Susan Jones,	817
764	pages 2511–2522, Singapore. Association for Com-	Micheline M Hancock-Beaulieu, Mike Gattford, et al.	818
765	putational Linguistics.	1995. <i>Okapi at trec-3. Nist Special Publication Sp</i> ,	819
766		109:109.	820
767	Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun-	Daniel Russo, Serra Sinem Tekiroğlu, and Marco	821
768	yuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-	Guerini. 2023. Benchmarking the generation of fact	822
769	Wei Chang, Michel Galley, and Jianfeng Gao. 2024.	checking explanations . <i>Transactions of the Associa-</i>	823
770	Mathvista: Evaluating mathematical reasoning of	tion for Computational Linguistics , 11:1250–1264.	824
771	foundation models in visual contexts . In <i>Inter-</i>		
772	<i>national Conference on Learning Representations</i>	Julio Cesar Salinas Alvarado, Karin Verspoor, and Tim-	825
773	(<i>ICLR</i>).	othy Baldwin. 2015. Domain adaption of named	826
774	Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov,	entity recognition to support credit risk assessment .	827
775	and Min-Yen Kan. 2023. SCITAB: A challenging	In <i>Proceedings of the Australasian Language Tech-</i>	828
776	benchmark for compositional reasoning and claim	<i>nology Association Workshop 2015</i> , pages 84–90,	829
777	verification on scientific tables . In <i>Proceedings of the</i>	Parramatta, Australia.	830
778	<i>2023 Conference on Empirical Methods in Natural</i>		
779	<i>Language Processing</i> , pages 7787–7813, Singapore.	Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and	831
780	Association for Computational Linguistics.	Sudheer Chava. 2023. Finer: Financial named entity	832
781		recognition dataset and weak-supervision model .	833
782	Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang	Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo	834
783	Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024.	Huang, Daogao Liu, Terra Blevins, Danqi Chen, and	835
784	Ex-fever: A dataset for multi-hop explainable fact	Luke Zettlemoyer. 2024. Detecting pretraining data	836
785	verification .	from large language models . In <i>The Twelfth Interna-</i>	837
786		<i>tional Conference on Learning Representations</i> .	838
787	Macedo Maia, Siegfried Handschuh, André Freitas,	Yejun Soun, Jaemin Yoo, Minyong Cho, Jihyeong Jeon,	839
788	Brian Davis, Ross McDermott, Manel Zarrouk, and	and U Kang. 2022. Accurate stock movement predic-	840
789	Alexandra Balahur. 2018. Www’18 open challenge:	tion with self-supervised learning from sparse noisy	841
790	Financial opinion mining and question answering . In	tweets . In <i>2022 IEEE International Conference on</i>	842
791	<i>Companion Proceedings of the The Web Conference</i>	<i>Big Data (Big Data)</i> , pages 1691–1700.	843
792	2018, WWW ’18, page 1941–1942, Republic and		
793	Canton of Geneva, CHE. International World Wide	Liyan Tang, Philippe Laban, and Greg Durrett. 2024.	844
794	Web Conferences Steering Committee.	Minicheck: Efficient fact-checking of llms on ground-	845
795		ing documents .	846
796	Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korho-	Gemma Team, Thomas Mesnard, Cassidy Hardin,	847
797	nen, and Jyrki Wallenius. 2013. Good debt or bad	Robert Dadashi, Surya Bhupatiraju, Shreya Pathak,	848
798	debt: Detecting semantic orientations in economic	Laurent Sifre, Morgane Rivi�re, Mihir Sanjay	849
799	texts .	Kale, Juliette Love, Pouya Tafti, L�onard Hussenot,	850
800	AI Meta. 2024. Introducing meta llama 3: The most	Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam	851
801	capable openly available llm to date . <i>Meta AI</i> .	Roberts, Aditya Barua, Alex Botev, Alex Castro-	852
802		Ros, Ambrose Slone, Am�lie H�liou, Andrea Tac-	853
803	Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee,	chetti, Anna Bulanova, Antonia Paterson, Beth	854
804	Soumya Sharma, Manjunath Hegde, Afreen Shaikh,	Tsai, Bobak Shahriari, Charline Le Lan, Christo-	855
805	Shivani Shrivastava, Koustuv Dasgupta, Niloy Gan-	pher A. Choquette-Choo, Cl�ment Crepy, Daniel Cer,	856
806	guly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECT-	Daphne Ippolito, David Reid, Elena Buchatskaya,	857
807	Sum: A new benchmark dataset for bullet point sum-	Eric Ni, Eric Noland, Geng Yan, George Tucker,	858
808	marization of long earnings call transcripts . In <i>Pro-</i>	George-Christian Muraru, Grigory Rozhdestvenskiy,	859
809	<i>ceedings of the 2022 Conference on Empirical Meth-</i>	Henryk Michalewski, Ian Tenney, Ivan Grishchenko,	860
810	<i>ods in Natural Language Processing</i> , pages 10893–	Jacob Austin, James Keeling, Jane Labanowski,	861
811	10906, Abu Dhabi, United Arab Emirates. Associa-	Jean-Baptiste L�spiau, Jeff Stanway, Jenny Bren-	862
	tion for Computational Linguistics.	nan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin	863
	OpenAI. 2022. Chatgpt: Optimizing language models	Mao-Jones, Katherine Lee, Kathy Yu, Katie Milli-	864
	for dialogue .	can, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon,	865
	OpenAI. 2023a. Gpt-4 technical report . <i>ArXiv</i> ,	Machel Reid, Maciej Miku�a, Mateo Wirth, Michael	866
	abs/2303.08774 .	Sharman, Nikolai Chinaev, Nithum Thain, Olivier	867

868	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 .	
884	InternLM Team. 2024. Internlm2 technical report .	
885	James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification . In <i>Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)</i> , pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.	
894	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal 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 .	
917	Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal 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 .	929 930 931 932 933 934 935 936 937 938 939
	Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction . In <i>Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science</i> , pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.	940 941 942 943 944 945
	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 <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 7534–7550, Online. Association for Computational Linguistics.	946 947 948 949 950 951 952
	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 <i>Findings of the Association for Computational Linguistics: EMNLP 2022</i> , pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	953 954 955 956 957 958 959
	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 <i>Advances in Neural Information Processing Systems</i> .	960 961 962 963 964
	Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid deep sequential modeling for social text-driven stock prediction . In <i>Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18</i> , page 1627–1630, New York, NY, USA. Association for Computing Machinery.	965 966 967 968 969 970 971
	Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance . <i>ArXiv</i> , abs/2303.17564.	972 973 974 975 976
	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 .	977 978 979 980
	Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices . In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.	981 982 983 984 985 986

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, 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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao 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*

Example

Claim: Pfizer has shown superior revenue growth compared to Moderna due to the widespread adoption of its COVID-19 vaccine, BNT162b2, and strong sales across its diverse portfolio of treatments. Moderna’s revenue growth, primarily driven by its COVID-19 vaccine, mRNA-1273, has been slightly slower than Pfizer’s.

Relevant Evidence Index:
[13, 14, 108, 110]

Reasoning Process Explanation:

Part 1: Extract relevant information:

1. Paragraph 13 states that Pfizer reported a substantial 25% year-over-year increase in revenue for its latest fiscal quarter, reaching \$20 billion.
2. Paragraph 14 states that the growth was primarily driven by the global demand for its COVID-19 vaccine, BNT162b2, which accounted for \$9 billion in revenue alone due to widespread vaccine distribution agreements and ongoing booster campaigns.
3. Paragraph 108 states that Moderna disclosed a 15% growth in revenue during the same fiscal period, totaling \$5 billion.
4. Paragraph 110 states that the growth was predominantly fueled by sales of its COVID-19 vaccine, mRNA-1273, which generated \$3.5 billion in revenue.

Part 2: Reason over extracted information:

1. To calculate the revenue growth of Pfizer, we use the equation $20 - 20 / (1 + 25\%) = 4$ billion.
2. To calculate the revenue growth of Pfizer, we use the equation $5 - 5 / (1 + 15\%) = 0.65$ billion.
3. It is clear that 4 billion > 0.65 billion.

Part 3: Return final entailment label:
entailed

Figure 3: An example from the FDV-COMPA subset. We provide detailed annotation for the features of claim, supporting evidence, and reasoning-process explanation in a standardized step-by-step format.

on Natural Language Processing (Volume 1: Long Papers), pages 3277–3287, Online. Association for Computational Linguistics.

A FINDVER Benchmark

A.1 Source Document Preparation

After collecting the raw HTML-format documents, we utilize the SEC API³, a commercial platform API for extracting financial document content, to process the collected documents, obtaining documents with both textual and tabular data.

For the subsets of FDV-IE, FDV-MATH, and FDV-KNOW, we utilize a single document as

³<https://sec-api.io/>

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

Table 6: 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.

the source document for each claim. The FDV-COMPA subset focuses on verifying claims by analyzing and comparing documents from two different companies. To achieve this, we concatenate two documents from companies within the same industry for the same fiscal year or quarter into a single document. We detail this process in Appendix A.1.

For the FDV-COMPA subset, we identified suitable company pairs for analysis as follows: 1) Extracted the "Item 1. Business" text from each company's latest annual report, as it describes the core business. 2) Used OpenAI's *text-embedding-3-large* model to represent each company's business description as embeddings and calculated cosine similarity between all pairs to identify highly related companies. 3) Had human annotators verify that the identified pairs were indeed related companies in the real world. 4) Selected only report pairs where the combined length did not exceed 80,000 words to accommodate the context window constraints of long-context LLMs.

B Experiment Setup

B.1 Implementation Details

LLM Experiment 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, Top P as 1.0, and maximum output length as 512.

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 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 GPT-3.5-turbo for processing LLM response. Specifically, we utilize GPT-3.5-turbo to extract labels from the LLM output, which can be either "entailed," "refuted," "not enough information," 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 resort to a random guess as the final label assignment.

C Experimental Results

Model	Organization	Release Time	Max Length	Source
GPT-4o & GPT-4-turbo (OpenAI, 2023a)	OpenAI	2023-03	128k	https://platform.openai.com/
GPT-3.5 (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 (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-2-70b-chat-hf
Llama-3 (Meta, 2024)	Meta	2024-04	8k	meta-llama/Meta-Llama-3-70B-Instruct
Yi-1.5 (AI et al., 2024)	01-ai	2024-05	32k	01-ai/Yi-1.5-34B-Chat
Qwen-2 (qwe, 2024)	Qwen	2024-06	128k	Qwen/Qwen2-72B-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-20b
C4AI (Aryabumi et al., 2024)	CohereForAI	2024-05	8k	CohereForAI/aya-23-35B
GLM (Du et al., 2022)	THUDM	2024-06	128k	THUDM/glm-4-9b-chat
AdaptLLM-Finance (Cheng et al., 2024)	AdaptLLM	2024-04	4k	AdaptLLM/finance-chat
Phi-3 (Abdin et al., 2024)	microsoft	2024-04	128k	microsoft/Phi-3-medium-128k-instruct

Table 7: 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.

Model	w/o CoT		w/ CoT	
	LongC	RAG	LongC	RAG
Qwen2-72B	37.0 (-9.1)	45.0 (-4.9)	46.1	49.9
GPT-4o	42.1 (-17.1)	54.1 (-4.2)	59.2	58.3

Table 8: Accuracy of entailment for GPT-4o and Qwen2-72B with and without CoT Prompting methods.

Adopted Chain-of-Thought Prompt

[System Input]

Given the provided context and a claim, your task is to verify the claim. The verification can result in three outcomes: entailed, refuted, or not enough information. Follow this structured approach:

Reasoning-process Explanation:

Part 1: Extract relevant information:

Part 2: Reason over extracted information:

Part 3: Return final entailment label:

Here is one example:

Claim: Apple's revenue increased in FY2024Q1 compared to the previous year.

Reasoning Process Explanation:

Part 1: Extract relevant information:

1. According to paragraph 13, the revenue of Apple is \$119.6 billion in FY2024Q1.
2. According to table 72, the revenue of Apple in the previous year's Q1 was \$117.15 billion.

Part 2: Reason over extracted information:

1. The revenue increase from the previous year's Q1 to FY2024Q1 is \$119.6 billion - \$117.15 billion = \$2.45 billion.

Part 3: Return final entailment label:

Therefore, the label for this claim is entailed.

[User Input]

{Financial Report}

{Claim}

Please follow the steps to verify this claim.

Figure 4: The Chain-of-Thought prompt used in this study. To verify the given claim, LLMs must first conduct a step-by-step reasoning process to identify the claim-relevant evidence and reason over this evidence. The LLMs then return the entailment label of the claim.

Common Error: Label Confusion

Claim: The percentage change in the net cash used in operating activities between 2022 and 2023 of Biogen Inc. is -29.73%.

Ground Truth Explanation:

Part 1: Extract relevant information:

1. According to paragraph 285, the net cash utilized in operating activities for 2023 amounts to \$4.1 million.

Part 2: Reason over extracted information:

1. To calculate the the change in the net cash used in operating activities between 2022 and 2023, we need both the values, but we can only find the data of 2023.

Part 3: Return final entailment label:

Not enough information

GPT-4o Generated Explanation:

Part 1: Extract relevant information:

1. From paragraph 285, the net cash used in operating activities for 2023 is \$4.1 million.
2. From paragraph 291, **because there is no net cash from operating activities, we use the net cash provided by financing activities in 2022, which is \$4.6 million, as a substitute.**

Part 2: Reason over extracted information:

1. The formula to calculate the percentage change is $(4.1 - 4.6) / 4.6 = -10.87\%$.
2. The percentage in the claim is -29.73%.

Part 3: Return final entailment label:

Therefore, the label for this claim is refuted.

Error Explanation: It confuses "refuted" with "not enough information". The assumption (mark in red) is made due to the absence of the desired value. Therefore, it should be labeled as "not enough information" rather than "refuted" because we lack sufficient evidence to disprove the claim.

Figure 5: An Example of Common Error: Label Confusion.

Common Error: Reasoning Error

Claim: VAL has a larger fleet of offshore drilling rigs than OII.

Ground Truth Explanation:

Part 1: Extract relevant information:

1. According to paragraph 201 of VAL, the number of rigs is 53.

Part 2: Reason over extracted information:

1. VAL's financial report explicitly states the number of rigs owned, which is 53, whereas OII's financial report does not mention the number of rigs owned or operated. It only describes the services and solutions provided by the OPG segment.
2. Due to the lack of information regarding the number of rigs owned or operated by OII, we cannot compare the two companies directly in terms of fleet size based on the provided reports.

Part 3: Return final entailment label:

Not enough information

GPT-4o Generated Explanation:

Part 1: Extract relevant information:

1. From VAL's financial report, specifically paragraph 201, it states: 'We currently own 53 rigs...' inclusive of various types of drilling rigs and a joint venture interest.
2. From OII's financial report, specifically paragraph 216, it indicates that OII's 'OPG segment provides a broad portfolio of integrated subsea project capabilities and solutions...' However, it does not provide a specific number of rigs owned or operated.

Part 2: Reason over extracted information:

1. The information from VAL clearly quantifies its fleet size, showing it owns a significant number of rigs.
2. OII's report focuses more on service offerings and operational capabilities without specifying the number of rigs, **indicating that they likely operate fewer or none of their own.**

Part 3: Return final entailment label:

Therefore, the label for this claim is entailed.

Error Explanation: In Part 2, the GPT-4o's reasoning about OII's rigs lacks solid evidence, as it makes an assumption based on the absence of specific details in OII's report and incorrectly links the absence of specific figures to fewer rigs (mark in red). Without concrete information directly confirming or denying the number of rigs owned by OII, the reasoning is speculative rather than based on firm evidence.

Figure 6: An Example of Common Error: Reasoning Error.

Common Error: Reasoning Error

Claim: ALGM has a stronger automotive segment presence in terms of percentage of revenue compared to MPWR.

Ground Truth Explanation:

Part 1: Extract relevant information:

1. ALGM: From table 60, Automotive net sales: \$577,515 out of \$808,786 (Nine-Month Period Ended)
2. MPWR: From table 453, Automotive net sales: \$394,665 out of \$1,821,072 (Nine-Month Period Ended)

Part 2: Reason over extracted information:

1. For ALGM, the automotive segment accounts for 71.4% of its total net sales for the nine-month period.
2. For MPWR, the automotive segment accounts for 21.7% of its total net sales for the same nine-month period.
3. Comparing these percentages, it is clear that ALGM has a much stronger presence in the automotive segment in terms of the percentage of revenue.

Part 3: Return final entailment label:
entailed

GPT-4o Generated Explanation:

Part 1: Extract relevant information:

1. ALGM: From table 60, Automotive net sales: $(577,515 / 808,786) * 100 = 71.4\%$ (Nine-Month Period Ended)
2. MPWR: From table 453, Automotive net sales: $(394,665 / 1,821,072) * 100 = 21.7\%$ (Nine-Month Period Ended)

Part 2: Reason over extracted information:

ALGM has around 71.4% (Nine-Month Period Ended) of revenue from automotive

MPWR only has 21.7% of its revenue from automotive for the nine-month period ended.

Thus, the proportion of ALGM's revenue from the automotive segment is significantly higher than MPWR's.

Part 3: Return final entailment label:

Therefore, the label for this claim is refuted.

Error Explanation: The error lies in Part 3. Despite the correct reasoning in Part 2, Part 3 incorrectly returns the label 'refuted'. The claim should be marked as "entailed" based on the provided data and reasoning. Therefore, the mistake is the contradiction between the correct reasoning in Part 2 and the incorrect conclusion in Part 3.

Figure 7: An Example of Common Error: Reasoning Error.

Specific Error: Extraction Error

Claim: NuStar Energy's Net Cash Used in Investing Activities decreased in the first quarter of 2024 compared to the first quarter of 2023.

Ground Truth Explanation:

Part 1: Extract relevant information:

1. Paragraph 224 states that there was a net cash outflow of \$40.1 million from investing activities, whereas for the same period, there was a net cash inflow of \$88.7 million from investing activities.

Part 2: Reason over extracted information:

1. The data clearly shows an increase in net cash usage year-over-year. In 2023, there was a cash inflow (net cash provided), and in 2024 there was a cash outflow (net cash used). This directly contradicts the claim made.

Part 3: Return final entailment label:

Refuted

GPT-4o Generated Explanation:

Part 1: Extract relevant information:

1. Paragraph 224 presents that 'For the three months ended March 31, 2024, we recorded net cash provided by investing activities of \$40.1 million, compared to net cash used in investing activities of \$88.7 million for the three months ended March 31, 2023'.

Part 2: Reason over extracted information:

1. The data indicates a notable improvement in the company's investing activities over the year. In 2023, the company experienced a significant cash outflow (net cash used of \$88.7 million).

2. In contrast, for the same period in 2024, the company experienced a cash inflow (net cash provided of \$40.1 million), indicating that the investing activities generated more cash than was spent.

3. This year-over-year comparison clearly demonstrates a decrease in net cash usage, transitioning from a cash outflow to a cash inflow. This positive shift supports the claim that there has been an improvement in net cash usage.

Part 3: Return final entailment label:

Therefore, the label for this claim is supported.

Error Explanation: The final conclusion is incorrect due to an error in the key information extraction (mark in red). The original extraction incorrectly stated that net cash was provided in 2024 and net cash was used in 2023. This erroneous information led to the flawed reasoning and conclusion. The correct information indicates a shift from net cash provided in 2023 to net cash used in 2024, which actually refutes the claim.

Figure 8: An Example of Specific Error: Extraction Error.

Specific Error: Numerical Reasoning Error

Claim: The interest income generated from the proceeds in the Trust Account was \$4,066,250 as of December 31, 2023, based on an annual average interest rate of 2%.

Ground Truth Explanation:

Part 1: Extract relevant information:

1. Paragraph 7 states that the funds deposited into the Trust Account totaled \$116,150,000, earning an average annual interest rate of 2%.

Part 2: Reason over extracted information:

1. The formula for computing interest income is: $\text{Interest} = \text{Principal} \times \text{Rate} \times \text{Time}$.

2. Given a principal of \$116,150,000, an annual interest rate of 2% (0.02), and a time period from March 2022 to December 2023 of approximately 21 months or 1.75 years, the interest calculation is: $\text{Interest} = \$116,150,000 \times 0.02 \times 1.75 = \$4,066,250$.

3. Therefore, the interest income amounts to \$4,066,250.

Part 3: Return final entailment label:

Entailed

GPT-4o Generated Explanation:

Part 1: Extract relevant information:

1. According to paragraph 7: The net proceeds placed into the Trust Account were \$116,150,000 and an annual average interest rate of 2%.

Part 2: Reason over extracted information:

1. The interest income can be calculated using: $\text{Interest} = \text{Principal} \times \text{Rate} \times \text{Time}$.

2. The principal is \$116,150,000, the rate is 2% per year (0.02) and the time period from March 2022 to December 2023 is approximately 21 months, or 1.75 years. $\text{Interest} = \$116,150,000 \times 0.02 \times 1.75 = \$406,625$.

3. Hence, the interest income should be \$406,625, not \$4,066,250.

Part 3: Return final entailment label:

Therefore, the label for this claim is refuted.

Error Explanation: The final conclusion is incorrect due to an error in the mathematical calculations. In Part 2, the calculation of the interest income was done incorrectly (mark in red). The correct calculation should be $\$116,150,000 \times 0.02 \times 1.75 = \$4,066,250$. Due to this miscalculation, the reasoning was flawed, leading to the incorrect conclusion that the claim is refuted. The correct conclusion should have been based on the accurate calculation, which would have supported the claim.

Figure 9: An Example of Specific Error: Numerical Reasoning Error.

Specific Error: Domain Knowledge Error

Claim: FOXO Technologies Inc.'s total assets decreased greatly from December 31, 2022, to September 30, 2023, due to substantial reductions in cash and cash equivalents, prepaid expenses, and the absence of reinsurance recoverables.

Ground Truth Explanation:

Part 1: Extract relevant information:

1. From Table 8, the total assets reported were \$35,408 on December 31, 2022, and \$3,131 on September 30, 2023. 2. Notably, cash and cash equivalents showed a drastic reduction from \$5,515 to \$42, prepaid expenses decreased from \$2,686 to \$1,306, and reinsurance recoverables dropped from \$18,573 to \$0 within the same period.
2. "Total assets": "The sum of all assets owned by a company",
"Cash and cash equivalents": "Liquid currency and other short-term investments that can be quickly converted into cash",
"Prepaid expenses": "The amount paid for goods or services in advance of receiving them",
"Reinsurance recoverables": "The portion of an insurance company's losses that can be recovered from reinsurance companies".

Part 2: Reason over extracted information:

1. The significant reduction in cash and cash equivalents suggests diminished liquidity or heavy cash usage during the period.
2. The decrease in prepaid expenses reflects a lower future financial obligation. The absence of reinsurance recoverables in 2023 could indicate a change in insurance arrangements or a swift closure of pending claims.
3. All these factors contribute substantially to the reduction in total assets.

Part 3: Return final entailment label:

Entailed

GPT-4o Generated Explanation:

Part 1: Extract relevant information:

1. From Table 8, the total assets reported were \$35,408 on December 31, 2022, and \$3,131 on September 30, 2023. 2. Notably, cash and cash equivalents showed a drastic reduction from \$5,515 to \$42, prepaid expenses decreased from \$2,686 to \$1,306, and reinsurance recoverables dropped from \$18,573 to \$0 within the same period.

Part 2: Reason over extracted information:

1. The reduction in cash and cash equivalents from \$5,515 to \$42 means an increase in cash reserves, **suggesting that the company has become more liquid and is in a better position to meet short-term obligations.**
2. The decrease in prepaid expenses from \$2,686 to \$1,306 **implies that the company has fewer expenses paid in advance, seen as a reduction in future liabilities.**
3. The absence of reinsurance recoverables, dropping from \$18,573 to \$0, **represents as a reduction in liabilities.**

Part 3: Return final entailment label:

Therefore, the label for this claim is refuted.

Error Explanation: The conclusion is incorrect due to a misunderstanding of financial concepts. The reduction in cash and cash equivalents from \$5,515 to \$42 does not mean an increase in cash reserves; it actually indicates a significant decrease, reducing the company's liquidity. The decrease in prepaid expenses from \$2,686 to \$1,306 should not be seen as a reduction in future liabilities but rather as a decrease in assets that were prepaid for future expenses. The absence of reinsurance recoverables, dropping from \$18,573 to \$0, does not represent a reduction in liabilities but rather a loss of potential assets that could be recovered. These misinterpretations led to the incorrect conclusion that the company's financial position has improved when, in fact, it has deteriorated significantly, supporting the original claim

Figure 10: An Example of Specific Error: Domain Knowledge Error.

Specific Error: Comparison Error

Claim: According to the financial reports, Apple's revenue grew by 28.62% in fiscal year 2021 compared to fiscal year 2020. Meanwhile, Microsoft's revenue increased from \$143.02B to \$168.09B, representing a growth rate of 17.48%. Therefore, Apple had a higher revenue growth rate than Microsoft in the fiscal year 2021.

Ground Truth Explanation:

Part 1: Extract relevant information:

1. According to paragraph 63, Apple's revenue grew by 28.62% in fiscal year 2021 compared to fiscal year 2020.
2. From paragraph 181, Microsoft's revenue increased from \$143.02B in fiscal year 2021 to \$168.09B in fiscal year 2022, representing a growth rate of 17.48%.

Part 2: Reason over extracted information:

1. The time periods for the revenue growth rates of Apple and Microsoft does not match. Apple's growth rate is for fiscal year 2021 compared to fiscal year 2020. Microsoft's growth rate is for fiscal year 2022 compared to fiscal year 2021.
2. The comparison should within the same time period.

Part 3: Return final entailment label:
refuted

GPT-4o Generated Explanation:

Part 1: Extract relevant information:

1. Paragraph 63 states that Apple's revenue increased by 28.62% from fiscal year 2020 to fiscal year 2021.
2. According to paragraph 181, Microsoft experienced a revenue growth of 17.48%.

Part 2: Reason over extracted information:

1. Since 28.62% is greater than 17.48%, the growth rate of Apple is larger than that of Microsoft.

Part 3: Return final entailment label:
Therefore, the label for this claim is entailed.

Error Explanation: The comparison between the two companies is problematic because meaningful comparisons require using a consistent metric and, preferably, evaluating both companies during the same time frame (mark in red).

Figure 11: An Example of Specific Error: Comparison Error.