Can Large language model analyze financial statements well?

Xinlin Wang and **Mats Brorsson**
Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
29 Av. John F. Kennedy
L-1855 Kirchberg Luxembourg
xinlin.wang,mats.brorsson@uni.lu

Abstract

Since GPT-3.5's release, large language models (LLMs) have made significant advancements, including in financial analysis. However, their effectiveness in financial calculations and predictions is still uncertain. This study examines LLMs' ability to analyze financial reports, focusing on three questions: their accuracy in calculating financial ratios, the use of these metrics in DuPont analysis and the Z-score model for bankruptcy prediction, and their effectiveness in predicting financial indicators with limited knowledge. We used various methods, including zero-shot and few-shot learning, retrieval-augmented generation (RAG), and fine-tuning, in three advanced LLMs and compared their outputs to ground truth and expert predictions to assess their calculation and predictive abilities.The results highlight both the potential and limitations of LLMs in processing numerical data and performing complex financial analyses.

1 Introduction

Financial reporting analysis plays an important role in a company's analysis of financial health, operational efficiency, and potential risks. Traditionally, this process has relied on skilled financial analysts to manually compute and interpret financial ratios derived from financial statements. Established methods such as DuPont analysis (Soliman, 2008) and the Altman Z-score model (Altman, 1968) have been developed and refined over decades to accurately estimate profitability, financial leverage, and risk of bankruptcy. However, these techniques are time-intensive, costly, and susceptible to human error, limiting their scalability and efficiency, particularly when real-time analysis of large datasets is required.

With the advent of models like GPT-3.5, large language models (LLMs) have shown remarkable potential to automate document analysis across domains (Kalyan, 2023). Advanced LLMs, such as GPT-4 and Llama, excel in natural language understanding, solving complex tasks, and generating contextual insights. Their robust text processing abilities offer an opportunity to transform traditional financial analysis by offering faster and more accessible insights to analysts and decision makers (Zhao et al., 2024).

Despite this promise, significant challenges persist in applying LLMs to quantitative tasks. Studies have noted that while LLMs handle language-based tasks effectively, they often struggle with precise numerical reasoning (Zhao et al., 2023). Recent advances, including fine-tuning on math datasets (Liu et al., 2023) and using hybrid approaches that combine LLMs with symbolic computation tools (Lam and Shareghi, 2024; Yamauchi et al., 2023), have improved numerical reasoning to some extent. However, their applicability to real-world financial contexts remains uncertain (Lee et al., 2024), as financial analysis demands not only linguistic comprehension but also accurate numerical computation from both structured and unstructured data (Li et al., 2023).

Given these challenges, it is crucial to assess whether LLMs can accurately analyze financial data, especially numerical data in financial statements, to support decision-making processes in finance. This study investigates the feasibility of using LLMs to automate three essential tasks in financial statement analysis: (1) calculating financial ratios, (2) utilizing these ratios in established models such as DuPont analysis and Altman's Z-score for bankruptcy prediction, and (3) forecasting critical indicators such as EBITDA and sales. Each task requires precise numerical computation, logical reasoning, and contextual understanding, making them ideal benchmarks for evaluating LLMs in financial statement analysis. By comparing the performance of various approaches (*zero-shot, few-shot, Retrieval-Augmented Generation (RAG), and fine-tuning*) with expert predictions

and ground truth, this research aims to identify both the strengths and limitations of LLMs in financial tasks.

In summary, this study provides a comprehensive evaluation of LLMs in financial statement analysis, providing insights into their strengths, limitations, and areas of improvement. The primary contributions of this study are:

- Systematically evaluating the accuracy of LLMs in computing financial ratios.
- Assessing the reliability of LLM-derived ratios in DuPont and Z-score models for bankruptcy prediction.
- Comparing LLMs with domain experts in forecasting key financial metrics, such as EBITDA and sales.
- Identifying challenges and limitations in applying LLMs to financial analysis, contributing to the broader field of AI in finance.

2 Related work

Financial analysis is a cornerstone of corporate finance, supporting decision-making in areas such as investment, risk management, and corporate governance. Traditional approaches rely on financial metrics derived from balance sheets, income statements, and cash flow statements, with ratios such as *profitability, liquidity, leverage*, and *efficiency* serving as essential indicators (Constantin and Loredana, 2012). These ratios form the basis for advanced analytical frameworks like DuPont analysis and the Altman Z-score model. DuPont analysis decomposes return on equity (ROE) into three components: *profit margin*, *asset turnover*, and *financial leverage*, allowing analysts to identify sources of financial performance (Soliman, 2008). Similarly, the Altman Z-score model predicts bankruptcy risk through a weighted combination of financial ratios (Altman, 1968). However, these methods are labor-intensive, prone to human error, and constrained in their ability to process large datasets or deliver real-time insights.

Advances in artificial intelligence (AI) and machine learning (ML) offer opportunities to automate financial analysis. While these methods improve efficiency and consistency, they often focus on pure numerical predictions (Zhu et al., 2023; Alessi and Savona, 2021) or textual sentiment analysis (Liu et al., 2021), falling short of replicating traditional frameworks like DuPont and Z-score (Emerson et al., 2019). Large language models (LLMs) represent a transformative technology in this space, demonstrating exceptional abilities in natural language understanding and complex problem-solving (Achiam et al., 2023; Minaee et al., 2024). By mastering complex linguistic patterns, LLMs excel in various domains, including customer support automation, content generation, and coding assistance (Chew et al., 2023).

In financial contexts, however, LLMs face unique challenges. Financial documents often contain jargon, numerical data, and intricate relationships that demand both linguistic and mathematical precision (Harvel et al., 2024). While LLMs like GPT-3.5 and GPT-4 have shown promise in tasks such as sentiment analysis (Liu et al., 2021), their numerical reasoning abilities are limited, particularly in multi-step calculations or exact numerical tasks (Brown, 2020; Zhao et al., 2023). Studies highlight that even state-of-the-art LLMs often miscalculate or misinterpret numerical contexts, leading to inaccurate financial projections (Hendrycks et al., 2020; Zhang et al., 2024). This limitation underscores the critical importance of precise numerical reasoning in financial decision-making, where even minor errors can lead to flawed conclusions.

Efforts to enhance LLMs' numerical reasoning have explored hybrid approaches, such as *Retrieval-Augmented Generation* (RAG), which integrates external databases for improved factual accuracy (Gupta et al., 2024; Ovadia et al., 2023). Fine-tuning on domain-specific datasets (Soudani et al., 2024) and techniques like Chain-of-Thought prompting have also been proposed to improve performance on complex financial tasks (Kim et al., 2024). These methods have demonstrated the potential to bridge gaps between LLM capabilities and traditional financial analysis. For instance, GPT-4 has been shown to outperform human analysts in predicting earnings changes (Kim et al., 2024), while few-shot learning has proven effective for text classification in finance with minimal labeled data (Loukas et al., 2023).

Despite these advances, no consensus exists on the optimal strategies for enhancing LLMs in numerical and domain-specific tasks. This paper seeks to address this gap by systematically benchmarking various methods, including zero-shot, few-shot, RAG, and fine-tuning, to evaluate their efficacy in financial applications. The findings aim to establish a clearer framework for leveraging LLMs

in finance and identify trade-offs between performance and computational efficiency.

3 Problem Formulation

The core objective of this study is to assess the effectiveness of LLMs in analyzing financial statements and making financial projections compared to traditional methods and experts' forecasts. Building on previous research (Section 2), which highlights the potential and limitations of LLMs in financial statements analysis and numerical reasoning, this study aims to identify the most effective models and methodologies for financial analysis tasks.

To achieve this, we address the following research questions:

RQ1: How accurately can LLMs compute financial ratios based on provided financial statement data?

RQ2: How effectively can LLMs predict bankruptcy risks using methodologies such as the Altman Z-score model and DuPont analysis?

RQ3: How capable are LLMs in forecasting critical financial indicators?

RQ4: What is the optimal combination of models and approaches balancing efficiency and effectiveness?

To better study these questions, we prepared a special dataset to simulate a qualified and experienced financial analyst, allowing LLMs to acquire knowledge from this dataset through RAG or fine-tuning.

4 Experimental Design

4.1 Dataset and Data Preprocessing

For this study, data preparation involves selecting both training and validation datasets. Fig. 1 shows the process of constructing the training set and testing set. We have five raw data sources, including a question-answer pair dataset, raw PDF files, and publicly available accessible databases. Combining Compustat and Institutional Brokers' Estimate System (IBES) by company's stock ticker, hybrid Compustat and IBES is constructed. The FinQA and CFA-QA datasets are only involved in the training set, the other three datasets are used in both training set and testing set. The details of these datasets will be introduced in the following.

FinQA Dataset: The FinQA dataset (Chen et al., 2021) includes annotated financial documents and tables derived from S&P 500 earnings reports. We only derive the question-answer pairs from 6251 samples in its training set, each pair comprises question spliced of `post_text`, `pre_text`, `table`, `question` and answer spliced of `answer`, `gold_evidence`.

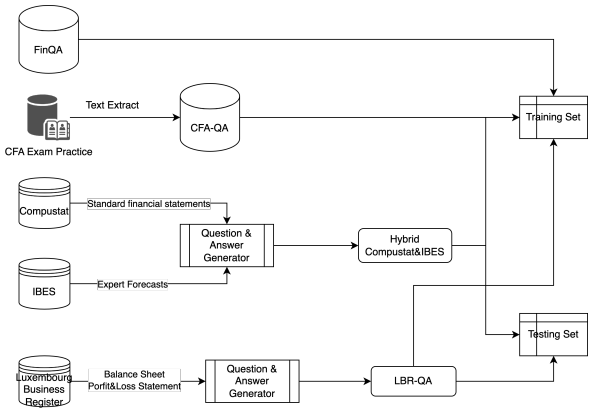

Figure 1: Workflow of constructing datasets for training and testing.

CFA-QA Dataset: Derived from Level I CFA exam materials[1], this dataset includes 208 question-answer pairs. A study proved that with few-shot learning, ChatGPT can pass the accounting certification exams (Eulerich et al., 2023), which means LLMs could have the ability to act like a certified expert. As the Level I CFA exam covers various topics in financial statement analysis, this dataset is particularly valuable for LLMs with RAG and fine-tuning to align with expert-level financial analysis standards.

Compustat: Compustat provides standardized financial statements and market data for North American companies, supporting robust bankruptcy risk evaluation. For this study, we focus on the fiscal years 2014 to 2019, extracting 50 accounting subjects and excluding pandemic-related anomalies.

Institutional Brokers' Estimate System (IBES): IBES includes expert analyst forecasts for EBITDA and sales, serving as benchmarks for evaluating LLM prediction accuracy. Joint with the samples selected from Compustat, we have 4957 companies with 21496 fiscal years in total. We randomly chose 1000 samples for training set and 1000 samples for testing set considering the experimental time of LLMs inference.

[1] https://www.cfainstitute.org/

Luxembourg Business Register: LBR[2] offers balance sheets and profit-and-loss statements from Luxembourg-based companies. Unlike other datasets, these documents feature diverse formats and accounting standards, testing the adaptability of LLMs to unstructured financial data. To standardize, only companies with both balance sheets and profit-and-loss statements for the same fiscal year were included. A total of 15908 samples were processed, with 1000 randomly selected for training and 1000 for testing. In summary, the number of samples included in the training and testing sets and their sources can be seen in Table 1.

Table 1: Summary of datasets

	Dataset	# samples
Training set	FinQA	6251
	CFA-QA	208
	Hybrid Compustat& IBES	1000
	LBR-QA	1000
Testing set	Hybrid Compustat& IBES	1000
	LBR-QA	1000

4.2 Methodology

To understand which models and methods are most effective for analysing financial statements, we chose three state-of-the-art open-source LLMs: Llama 3.2 3B[3], Llama 3.1 8B[4], Mistral 7B[5]. Compared to closed-source models like GPT-4. These we can have complete control over the model's architecture, parameters, and training data without dependence on third-party platforms, which permits us to make flexible adjustments and optimizations. The capability of researching open-source models could offer enterprises or research institutions the solutions rather than relying solely on commercial models.

Llama 3 models, particularly the latest version, exhibit competitive capabilities compared to leading models like GPT-4, especially in multilingual support and complex reasoning tasks (Dubey et al., 2024). Llama 3.2, being the latest version, incorporates higher parameter optimization and knowledge updates, and holds the potential to perform outstandingly in understanding complex language tasks and mathematical reasoning. While Llama 3.1, as the previous version, can be used for comparison to assist in analyzing whether version iterations bring about significant improvements. Mistral focuses on efficient parameter utilization, excelling in minimizing hallucinations and achieving performance approaching while using fewer parameters (Jiang et al., 2023). It is suitable for contrast experiments that are sensitive to resource efficiency, especially for analyzing the actual performance of the model under limited computing power. We use the same setting for LLMs in this paper considering the needs of comparison: max_new_tokens is set to 2048 to ensure a complete answer, temperature is set to 0 or 1e-5 to have a consistency answer set, load_in_4bit is true to smoothly deploy LLMs.

To optimize the performance of these LLMs, this study employed three primary strategies: prompt engineering, retrieval-augmented generation (RAG), and fine-tuning. Prompt engineering involved zero-shot and few-shot learning. In zero-shot learning, no previous examples were provided, allowing the evaluation of the model's baseline capabilities. Few-shot learning was conducted by presenting the model with a limited number of question-answer pairs, testing its ability to generalize from minimal context in financial tasks. For RAG, a vector database was incorporated to retrieve domain-specific financial knowledge, which the models used to enhance accuracy in question answering and financial ratio computations. Fine-tuning was performed using supervised training on domain-specific question-answer pairs, allowing the models to align more closely with the requirements of financial statement analysis.

Fig 2 illustrates the overall experimental design, where the training set is exclusively used for RAG and fine-tuning, while the testing set evaluates all combinations of models and optimization techniques. This study designs three categories of questions according to the RQs. Question 1 focused on computing financial ratios, Z-score values, and bankruptcy risks using the Altman Z-score model. Question 2 involved calculating financial ratios, return on equity (ROE), and bankruptcy risks by DuPont analysis. Question 3 is to ask for the predicted EBIDTA and sales based on provided financial statements and its own knowledge. Combining the financial statements from hybrid Compustat/IBES and LBR, we can have the full text of questions. For the answers, we populate the manually calculated financial ratios, Z-score value and ROE value into the fixed-format text as the ground truth.

[2] https://www.lbr.lu

[3] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

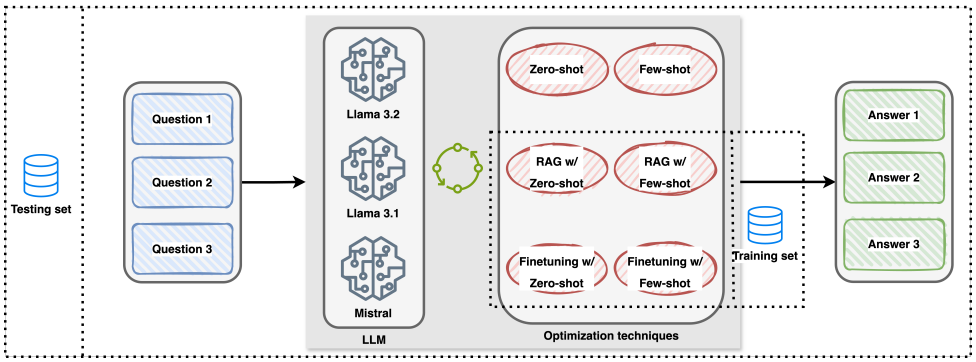

Figure 2: Workflow of experimental structure.

With zero-shot learning and few-shot learning, LLMs will directly return the answers. We deploy RAG and fine-tuning in conjunction with the same prompts as used in zero-shot learning and few-shot learning for the questions. Therefore, there are six techniques in the optimization techniques part. Considering the LLMs, in total, we have 18 different combinations of LLMs and optimization techniques, which constitute a comprehensive evaluation of how LLMs can be adapted to tackle financial analysis tasks.

4.3 Evaluation Metrics

The inference tasks of this study not only emphasise text generation, but also highlight the importance of the correctness of mathematical calculations related to financial ratios. Therefore, to fully evaluate the effectiveness of the model, we apply a set of evaluation metrics across the four research questions.

Completion rate: In this study, we particularly define a metric named completion rate for the research questions 1. For Question 1 to Question 3, we require the LLMs to summarise the required values in JSON format. Therefore, it it vital for a qualified answer to have this complete JSON to present the required calculated or forecasted values of corresponding questions. The completion rate is defined in equation 1.

$$R = \frac{\sum_{i=1}^{N}(\mathrm{A}_i \cdot \mathrm{B}_i \cdot \mathrm{C}_i)}{N} \tag{1}$$

where, N means the total number of generated answers, A_i represents whether the i-th answer contains a valid JSON format. It is 1 if valid, otherwise 0. B_i indicates whether the JSON contains all the required fields. It is 1 if all fields are present, otherwise 0. C_i checks if the values of the fields in the JSON are numbers (either integers or floats). It is 1 if all values are numeric, otherwise 0.

Recall-Oriented Understudy for Gisting Evaluation(ROUGE): ROUGE can measure the degree of overlap between the generated answers and the reference answers in terms of n-grams or the longest common subsequence, with particular emphasis on coverage (Lin, 2004). In this study, we employed ROUGE-L to evaluate the calculation steps of financial ratios or the reasoning behind predictions, as it not only assesses whether the generated text covers the reference content but also pays special attention to whether the answers are provided in sequence.

Symmetric Mean Absolute Percentage Error (sMAPE): sMAPE measures the percentage error relative to the actual value(see equation 2) and avoids the problem of infinite values when actual values are zero, making it more reliable in such cases.

$$\text{sMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\frac{|y_i|+|\hat{y}_i|}{2}} \times 100 \tag{2}$$

where, y_i is the actual value for the i-th data point, $\hat{y}_i$ is the predicted value for the i-th data point, n is the total number of data points.

5 Results analysis

5.1 Answers completion

Fig. 3 highlights clear distinctions in the performance of the three LLMs across optimization strategies. Llama3.1 outperforms its counterparts in 4 scenarios, particularly excelling in zero-shot learning and finetuning with few-shot learning. Llama3.2, while showing strong general performance, exhibits minor declines in completion rates under specific fine-tuning and RAG scenarios, suggesting some sensitivity to the optimization approach. Mistral, although competitive in RAG with zero-shot learning, lags significantly behind

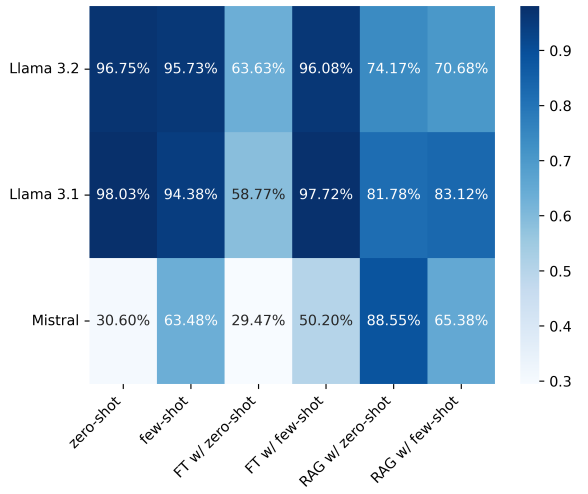

Figure 3: Distribution of completion rate over different combinations of LLMs and optimization techniques.

in other settings, indicating potential architectural or pre-training limitations in handling structured output requirements.

These results underline the importance of aligning model selection and optimization strategies with specific task requirements. Llama3.1 and Llama3.2 emerge as reliable choices for tasks demanding consistent and complete outputs, while Mistral's use may be more suited to resource-constrained scenarios or specific RAG applications.

5.2 Evaluation on calculation steps

Table 2 reveals distinct performance patterns among the three LLMs across the Altman Z-score model and DuPont analysis. Llama 3.1 consistently achieves the highest overall performance, excelling particularly in fine-tuning tasks, where it demonstrates superior F1 scores for both analysis methods. Llama 3.2 performs well in structured optimization tasks but underperforms in certain retrieval-augmented generation (RAG) scenarios. Mistral, while generally weaker, shows competitive results in RAG-based tasks, particularly with the DuPont analysis.

For the Altman Z-score model, Llama 3.1 dominates in fine-tuning (87.60% F1), while Mistral performs better in zero-shot RAG tasks (75.82%). In the DuPont analysis, Llama 3.1 also leads in fine-tuning scenarios, while Mistral achieves its highest performance in RAG with zero-shot learning (89.33%), surpassing both Llama models. Across both methods, introducing few-shot examples in RAG leads to slight performance declines for most models, but Llama 3.1 maintains its lead.

5.3 Financial Metric Calculation Accuracy

Fig 4 shows significant variation in model performance across datasets, ratios, and optimization configurations. Llama 3.2 demonstrates the most notable improvement in the Altman Z-score Model, reducing sMAPE from 186.8 (zero-shot) to 135.0 (RAG with few-shot). Similarly, Llama 3.1 shows effective enhancement in the Working Capital/Total Assets ratio, where sMAPE improves from 96.1 to 75.9 with few-shot learning. In contrast, Mistral displays inconsistencies, particularly in ratios like Earnings Before Interest and Tax/Total Assets, where RAG with zero-shot leads to a high sMAPE of 191.1, indicating limited benefit from additional vector database information.

RAG with few-shot consistently emerges as the most reliable method, particularly for complex financial prediction tasks. However, ratios involving equity and earnings, such as Market Value of Equity/Total Liabilities and Earnings Before Interest and Tax/Total Assets, remain challenging due to their sensitivity to financial volatility. High sMAPE values, such as 196.3 (Llama 3.1) and 161.7 (Mistral) for equity-related ratios, highlight the need for improved approaches.

While the overall sMAPE is high, for certain ratios like total sales/total assests (Compustat&IBES), all the LLMs perform well, which means LLMs indeed have potential to anaylze the financial statements.

5.4 Bankruptcy Prediction

Table 3 reveals ignificant variability in LLM performance for bankruptcy prediction, with results heavily influenced by the optimization strategy. Llama 3.2 shows the most consistent performance in bankruptcy prediction, particularly with zero-shot learning, achieving up to 82% accuracy and 0.62 AUC for DuPont analysis. However, its performance declines under few-shot learning and fine-tuning, highlighting the limitations of these methods. Llama 3.1 underperforms overall but demonstrates potential in combining retrieval-based techniques with few-shot training, achieving an AUC of 0.76 for the Altman Z-score model. Mistral delivers mixed results, with competitive zero-shot accuracy but poor fine-tuning performance, particularly for DuPont analysis.

Overall, Llama 3.2 is the most reliable model for bankruptcy prediction, but its variability across optimization methods underscores the need for more

Table 2: ROUGE-L comparison of different combinations of LLMs and optimization techniques

		Altman Zscore Model			DuPond analysis		
		Recall	Precision	F1 score	Recall	Precision	F1 score
Llama 3.2	zero-shot	31.80%	35.90%	33.06%	30.90%	39.17%	34.34%
	few-shot	12.27%	62.93%	19.21%	8.14%	70.58%	13.10%
	FT w/ zero-shot	79.70%	90.50%	84.30%	88.18%	92.48%	89.92%
	FT w/ few-shot	50.08%	80.87%	56.53%	85.16%	88.79%	86.89%
	RAG w/ zero-shot	75.27%	69.59%	69.69%	59.15%	58.61%	56.29%
	RAG w/ few-shot	43.94%	52.46%	46.51%	59.29%	63.48%	58.96%
Llama 3.1	zero-shot	29.50%	41.73%	31.36%	31.99%	47.08%	35.73%
	few-shot	60.36%	79.94%	68.49%	48.19%	68.97%	55.79%
	FT w/ zero-shot	82.77%	93.70%	**87.60%**	88.01%	**95.58%**	**91.50%**
	FT w/ few-shot	70.46%	90.18%	78.78%	**89.27%**	93.45%	91.26%
	RAG w/ zero-shot	**83.00%**	87.57%	84.73%	88.93%	67.81%	75.35%
	RAG w/ few-shot	68.44%	79.77%	73.57%	80.36%	85.65%	82.02%
Mistral	zero-shot	30.62%	52.89%	37.74%	24.74%	36.94%	29.28%
	few-shot	36.73%	41.92%	38.43%	85.14%	78.33%	80.74%
	FT w/ zero-shot	66.12%	**96.94%**	78.04%	86.08%	95.32%	90.32%
	FT w/ few-shot	34.64%	54.25%	42.10%	85.09%	88.11%	86.08%
	RAG w/ zero-shot	73.31%	80.40%	75.82%	88.10%	90.82%	89.33%
	RAG w/ few-shot	52.13%	79.41%	62.77%	84.72%	91.60%	87.99%

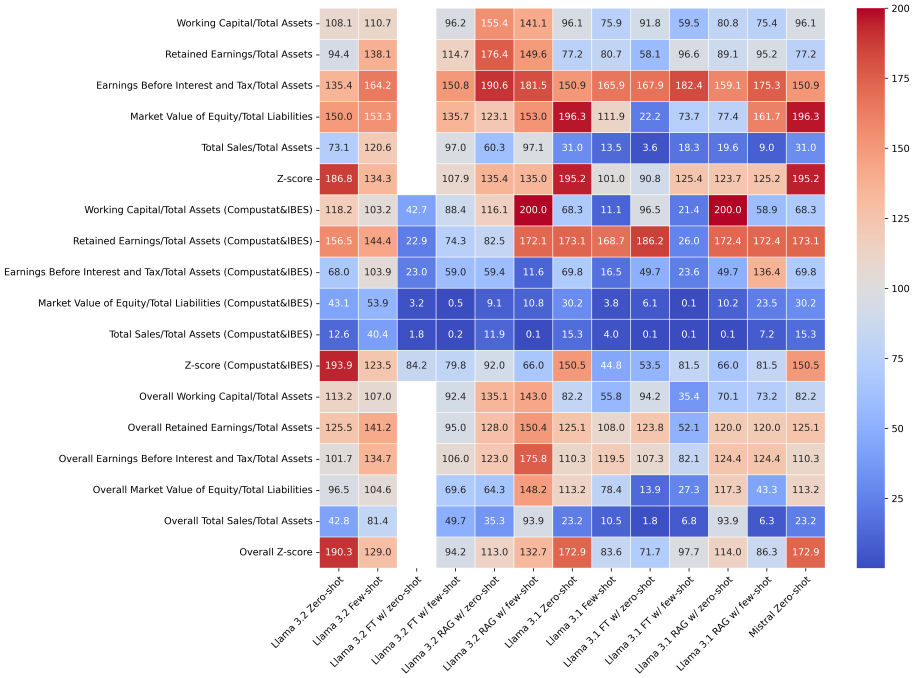

Figure 4: sMAPE for financial ratios by over different combinations of LLMs and optimization techniques. The blank area is no value due to lack of valid answers.

robust strategies tailored to financial tasks.

5.5 EBITDA and Sales Forecasting

In Table 4, we only put the best forecasting from LLMs and compare it with the forecasts from human financial expert. The financial expert achieved exceptionally low sMAPE values of 25.1 for "Next Year Sales" and 44.9 for "Next Year EBITDA," far surpassing the results obtained by all LLM configurations(B). This large gap in accuracy indicates that, despite the advances in machine learning and natural language processing, LLMs are not yet capable of matching the forecasting precision of experienced financial analysts, particularly when it comes to complex financial metrics that require nuanced judgment and domain expertise.

5.6 Resources Consumption

In this paper, we analyze the time, CPU memory, and GPU memory consumption across different models and optimization methods and reveal key performance trade-offs. The detailed records can be seen from A. Llama3.1 offers the most consistent performance, particularly in few-shot optimiza-

Table 3: Performance evaluation for bankruptcy prediction by LLMs. Slash means can't calculate the metrics due to lack of valid answer.

		Altman Zscore Model		DuPond Analysis	
		Accuracy	AUC	Accuracy	AUC
Llama 3.2	zero-shot	79%	0.61	82%	0.62
	few-shot	78%	0.36	74%	0.44
	FT w/ zero-shot	/	/	46%	0.59
	FT w/ few-shot	79%	0.52	77%	0.53
	RAG w/ zero-shot	63%	0.56	35%	0.49
	RAG w/ few-shot	64%	0.50	57%	0.30
Llama 3.1	zero-shot	66%	0.65	66%	0.59
	few-shot	61%	0.58	53%	0.61
	FT w/ zero-shot	/	/	44%	0.48
	FT w/ few-shot	73%	0.62	69%	0.58
	RAG w/ zero-shot	60%	0.65	47%	0.46
	RAG w/ few-shot	66%	0.76	51%	0.58
Mistral	zero-shot	79%	0.67	67%	0.75
	few-shot	/	/	65%	0.62
	FT w/ zero-shot	/	/	22%	0.41
	FT w/ few-shot	/	/	69%	0.30
	RAG w/ zero-shot	65%	0.61	67%	0.63
	RAG w/ few-shot	/	/	53%	0.39

Table 4: Comparison of the forecasting ability of LLMs and financial expert.

	Next Year Sales Prediction	Next Year EBITDA Prediction
Llama 3.2|zero-shot	/	129.6
Llama 3.1|few-shot	123.2	/
Expert Forecasting	**25.1**	**44.9**

tion, with the fastest response times (50 seconds). Mistral also excels in few-shot scenarios but is less effective in more complex methods. Llama3.2, while delivering high performance, requires significantly more computational resources, especially for RAG-based tasks, with response times reaching up to 600 seconds.

Regarding CPU consumption, all models exhibit similar usage, with slight increases under RAG methods, particularly for Llama3.2. However, CPU requirements are not a major constraint for any model, with usage staying below 2.5GB in most cases. GPU consumption shows more variation, with Llama3.1 consuming the most GPU memory (over 5GB), while Llama3.2 is the most resource-efficient, particularly in zero-shot and few-shot learning scenarios.

In conclusion, Llama3.1 offers the best balance of efficiency and performance for low-latency tasks, Mistral is suitable for few-shot optimization in resource-constrained settings, and Llama3.2 excels in high-quality tasks but requires more computational power, especially for complex optimization strategies like RAG.

6 Conclusion

The study demonstrates clear performance and resource trade-offs across Llama 3.2, Llama 3.1, and Mistral. Llama 3.1 achieves the highest accuracy, particularly with fine-tuning and RAG combined with few-shot learning, although it requires higher GPU memory (30% more than Llama 3.2). This makes Llama 3.1 ideal for accuracy-critical tasks where computational resources are sufficient.

Llama 3.2 balances performance and resource efficiency well, showing lower GPU and CPU usage, especially in fine-tuning and RAG. It offers a cost-effective alternative for large-scale or resource-constrained deployments, achieving competitive results with 20%–30% less GPU memory usage than Llama 3.1.

Mistral shows mixed performance, excelling in retrieval-intensive tasks but underperforming in others, particularly with zero-shot or fine-tuning optimizations. Its architecture suits tasks requiring efficiency but limits its effectiveness in general-purpose financial applications.

In summary, Llama 3.1 is best for high-accuracy tasks, particularly in RAG and few-shot setups, while Llama 3.2 is a more resource-efficient choice. Mistral performs well in retrieval-heavy tasks but struggles with accuracy in other areas. These results emphasize the need for model and optimization strategy selection based on task requirements and resource constraints, with future research focusing on hybrid approaches to further balance performance and resource usage.

References

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

Lucia Alessi and Roberto Savona. 2021. Machine learning for financial stability. In *Data Science for Economics and Finance: Methodologies and Applications*, pages 65–87. Springer International Publishing Cham.

Edward I Altman. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. Llm-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*.

CĂRUNTU Constantin and LĂPĂDUŞI MIHAELA Loredana. 2012. The analysis of the indicators which reflect the ability of companies of facing short term obligations and medium and long term maturities. *Annals-Economy Series*, 4(1):89–95.

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

Sophie Emerson, Ruairí Kennedy, Luke O'Shea, and John O'Brien. 2019. Trends and applications of machine learning in quantitative finance. In *8th international conference on economics and finance research (ICEFR 2019)*.

Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh, and David A Wood. 2023. Can artificial intelligence pass accounting certification exams? chatgpt: Cpa, cma, cia, and ea. *ChatGPT: CPA, CMA, CIA, and EA*.

Aman Gupta, Anup Shirgaonkar, Angels de Luis Balaguer, Bruno Silva, Daniel Holstein, Dawei Li, Jennifer Marsman, Leonardo O Nunes, Mahsa Rouzbahman, Morris Sharp, et al. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*.

Nicholas Harvel, Felipe Bivort Haiek, Anupriya Ankolekar, and David James Brunner. 2024. Can llms answer investment banking questions? using domain-tuned functions to improve llm performance on knowledge-intensive analytical tasks. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 125–133.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Katikapalli Subramanyam Kalyan. 2023. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048.

Alex Kim, Maximilian Muhn, and Valeri V. Nikolaev. 2024. 1. financial statement analysis with large language models. *arXiv.org*.

Long Hei Matthew Lam and Ehsan Shareghi. 2024. 2. a closer look at logical reasoning with llms: The choice of tool matters.

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yixin Liu, Avi Singh, C. D. Freeman, John Co-Reyes, and Peter J. Liu. 2023. 1. improving large language model fine-tuning for solving math problems. *arXiv.org*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519.

L. Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and S. Vassos. 2023. 2. breaking the bank with chatgpt: Few-shot text classification for finance.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Mark T Soliman. 2008. The use of dupont analysis by market participants. *The accounting review*, 83(3):823–853.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. *arXiv preprint arXiv:2403.01432*.

Ryutaro Yamauchi, Shoushin Sonoda, Akiyoshi Sannai, and Wataru Kumagai. 2023. 3. lpml: Llm-prompting markup language for mathematical reasoning. *arXiv.org*.

Chong Zhang, Xinyi Liu, Mingyu Jin, Zhongmou Zhang, Lingyao Li, Zhengting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. 2024. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. *arXiv preprint arXiv:2407.18957*.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. 2024. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

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

Xiuqin Zhu, Lian-Xin Jiang, Yixin Gao, and Youbing Yin. 2023. 1. research on financial statement analysis methods based on machine learning.

A Appendix of resource consumption

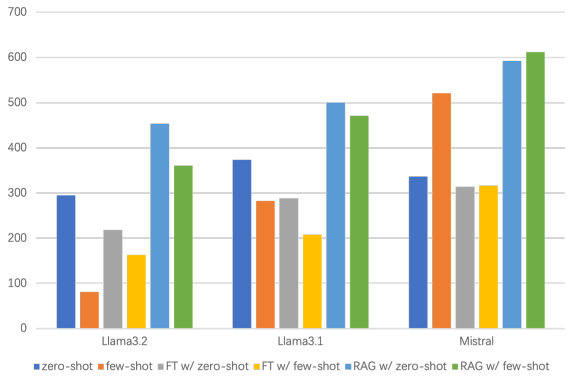

Figure 5: Average time consumption for each answer.

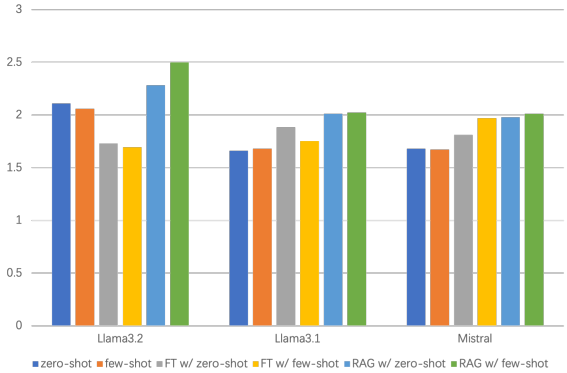

Figure 6: Average CPU memory consumption for inference.

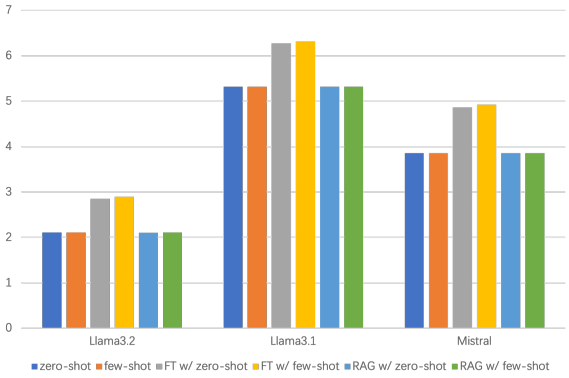

Figure 7: Average GPU consumption for inference.

B Appendix of forecasting ability

Table 5: Comparison of the forecasting ability of LLMs and financial expert.

		Next Year Sales Prediction	Next Year EBITDA Prediction
Llama 3.2	zero-shot	139.6	129.6
	few-shot	137.7	146.5
	FT w/ zero-shot	132.7	142.5
	FT w/ few-shot	137.1	146.0
	RAG w/ zero-shot	134.8	134.8
	RAG w/ few-shot	/	/
Llama 3.1	zero-shot	139.5	139.9
	few-shot	123.2	149.2
	FT w/ zero-shot	137.5	140.7
	FT w/ few-shot	138.1	152.9
	RAG w/ zero-shot	135.5	135.0
	RAG w/ few-shot	/	/
Mistral	zero-shot		
	few-shot	136.8	152.7
	FT w/ zero-shot	124.7	131.3
	FT w/ few-shot		
	RAG w/ zero-shot	139.4	130.9
	RAG w/ few-shot	/	/
Expert Forecasting		**25.1**	**44.9**