

Evaluating Large Language Models (LLMs) in Financial NLP: A Comparative Study on Financial Report Analysis

Md Talha Mohsin

Department of Finance & Operations Management, University of
Tulsa, 800 S Tucker Dr, Tulsa, 74104, OK, USA .

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide variety of Financial Natural Language Processing (FinNLP) tasks. However, systematic comparisons among widely used LLMs remain under-explored. Given the rapid advancement and growing influence of LLMs in financial analysis, this study conducts a thorough comparative evaluation of five leading LLMs, GPT, Claude, Perplexity, Gemini and DeepSeek, using 10-K filings from the "Magnificent Seven" technology companies. We create a set of domain-specific prompts and then use three methodologies to evaluate model performance: human annotation, automated lexical-semantic metrics (ROUGE, Cosine Similarity, Jaccard), and model behavior diagnostics (prompt-level variance and across-model similarity). The results show that GPT gives the most coherent, semantically aligned, and contextually relevant answers; followed by Claude and Perplexity. Gemini and DeepSeek, on the other hand, have more variability and less agreement. Also, the similarity and stability of outputs change from company to company and over time, showing that they are sensitive to how prompts are written and what source material is used.

Keywords: Large Language Models (LLMs), Natural Language Processing (NLP), Financial Text Analysis, 10-K, AI in Finance, Generative AI, Explainable AI (XAI)

1 Introduction

Artificial Intelligence (AI) has rapidly transformed how information is processed, analyzed, and interpreted across a multitude of industries. Among its most groundbreaking developments is the emergence of Large Language Models (LLMs)—a class of AI systems trained on vast corpora of text to generate human-like responses. Recently, LLMs have garnered significant attention in both academia and business (Bommasani et al., 2022, Zhao et al., 2023, Wei et al., 2022). These models are being used more and more in professional fields for their effectiveness, accuracy and rapidness.

As the need for scalable, low-cost tools for strategy evaluation grows, LLMs are being used to understand financial disclosures, study market sentiment, and put together unstructured data (Daimi and Iqbal, 2024). For example, (A. Kim et al., 2025) states how LLMs can work wonders because they can predict changes in company

earnings better than human analysts can, even when there aren't any clear narratives or industry-specific signals. They can also show how people feel about a company and any possible biases that might affect how investors act and how prices change (Nakagawa et al., 2024). Additionally, LLMs can create financial digests by condensing large amounts of text into clear, actionable insights (Lazarev and Sedov, 2024); this leads to better trading methods with higher Sharpe ratios (A. Kim et al., 2025). In these situations, it is very important to know how to use domain-specific language and make products that can be understood. Because of this, LLMs are now judged not only on their fluency or coherence, but also on their organized, expert-level analysis in complex Financial Natural Language Processing (FinNLP) tasks. A lot of natural language processing tasks, like summarizing, answering questions, and semantic document analysis, have been done very well by LLMs. But not much is known about how different LLMs act in high-stakes, domain-specific situations like company financial disclosures. This is especially worrying in the sense that in the financial world, regulators, institutional investors, and analysts depend on complicated textual disclosures heavily to make decisions. Transformer-based LLMs, which are pre-trained on huge datasets and fine-tuned to spot complex language patterns (T. T. Kim et al., 2025), have sped up recent progress in natural language processing (NLP). In finance, NLP methods have mostly been used for classifying emotions or recognizing entities. However, new research is looking into how well they can be used for semantic reasoning and understanding stories (S. Wu et al., 2024). Since financial reports are mostly text, it makes sense to use AI-powered tools to analyze them (Abdaljalil and Bouamor, 2021).

The Securities and Exchange Commission (SEC) requires public companies to file structured textual disclosures, like the 10-K annual report. Important soft data that quantitative predictors miss can be found in these filings (Lombardo et al., 2024). Crucial insights into the performance of a company can be gained through the examination of extensive records, particularly 10-K filings. Due to the fact that these filings encompass qualitative aspects such as strategy, risk exposure, and competitive positioning, they are excellent candidates for advanced natural language processing-based analysis. Financial statements, risk disclosures, and strategy overviews are all the components that are included in 10-K filings, which offer a structured but uncured perspective of the health of the corporation. They are, nevertheless, notoriously difficult to understand. Both human and machine analysts face difficulties when dealing with them because of their unstructured format and technical jargon. The process of text mining provides a powerful method for extracting value from these disclosures by revealing patterns, sentiment, and relational data across companies and industries (H. Kim et al., 2023).

This study address that gap by suggesting a multi-dimensional evaluation approach to see how five cutting-edge LLMs— ChatGPT-4, Perplexity, Claude 4 Opus, Gemini, and DeepSeek—look at the Business section (Item 1) of 10-K filings from the "Magnificent 7" tech companies over the last three years. We introduce a Chain-of-Thought (CoT) prompting approach that tells models to behave like financial analysts in order to simulate real-life analytical workflows. Their responses are evaluated using three distinct lenses: human-centric judgment, automated metric-based scoring, and prompt-sensitivity based behavioral analysis. Our work makes the following three contributions: (i) We test how sensitive LLM is to the design and amount of information in prompts, showing patterns of behavior across model structures. (ii) We create a benchmark that can be replicated to test LLMs' understanding of financial matters by focusing on detailed information rather than numbers, and (iii) We give academics and professionals a direction on how to use or evaluate LLMs in financial situations, mainly for strategic analysis, and information extraction The remainder of this study is structured as follows: Section 2 discusses background, Section 3 provides contextual insights into how LLMs work, Section 4 outlines the data and methodology, Section 5 presents the findings, section 6 offers discussion, and Section 7 concludes the study.

2 Related Work

2.1 Financial Text Analysis Using NLP

The systematic, and consistent quality of content analysis has made it a fundamental component of qualitative research for a long time. This is especially true in the field of finance, where content or text analysis has undergone substantial development since the introduction of Natural Language Processing (NLP). By mining massive amounts of financial text, modern natural language processing systems are now able to assist educated investment strategies, track macroeconomic signals, and improve decision-making processes within institutions (“Textual Analysis in Finance — Annual Reviews”, [n.d.](#)). When new natural language processing (NLP) and information retrieval technologies come into existence, the meaning of “new” information and the accessibility of it evolved, providing early adopters with a temporary advantage (Araci, 2019). These techniques integrate linguistic, statistical, and deep learning methodologies in order to extract value from complicated documents (L. Wang et al., 2024). Automating the extraction of structured insights from unstructured financial writings is one of the strengths of natural language processing (NLP). According to (Oyewole et al., 2024), these methods are becoming increasingly utilized in order to improve the accuracy of reporting, provide support for regulatory compliance, and perform risk signal detection in a more efficient manner.

Before NLP, Word2Vec, GloVe, and FastText were some of the first embedding models that learned word associations based on co-occurrence patterns. However, these models lacked contextual complexity in learning vocabulary. Regardless of the contexts in which a word is used, a single static vector was allocated to each individual word. Since then, this constraint has been overcome thanks to the development of contextualized embeddings through transformer-based models. These models encode words in a dynamic manner while taking into account the text that is surrounding them (Huang et al., 2023).

In recent years, the diverse use of FinNLP has entered the financial domain. From the detection of financial risks to the interpretation of narratives, these systems have grown more accurate and scalable thanks to the development of technologies such as word embeddings and sentiment analysis (Sehrawat, 2019, (L. Wang et al., 2024)). Analysts now are able to discover latent sentiment and theme shifts in real time across enormous datasets by utilizing NLP approaches (Zhou et al., 2024).

2.2 Emergence of LLMs in Financial NLP

A Large Language Model (LLM) is an AI algorithm that uses deep learning and large datasets to interpret, summarize, combine, and predict new information (“Prompt Engineering in Large Language Models”, 2024). LLMs use transfer learning as they are pre-trained on large text corpora to learn language skills and then fine-tuned for specific tasks, which leads to amazing performance (Luo & Gong, 2024; Zhang et al., 2023). Large Language Models work by turning incoming text into high-dimensional vector representations. They do this by using multilayer transformer topologies to capture the semantic and contextual relationships between words and phrases. Responses are made by predicting the next token in an autoregressive way, using the statistical patterns that have been learned. The quality of these answers depends on a number of things, such as the input prompt, which affects the context and specificity; the model’s hyperparameters, which control how it makes inferences; and the variety of the training data, which determines how much knowledge the model has (Bender et al., 2021). LLMs use probabilistic token selection methods to create outputs, which means that the same inputs and prompts might lead to different outputs in different runs (J. J. Wang and Wang, 2025). Large Language Models are gradually outperforming many

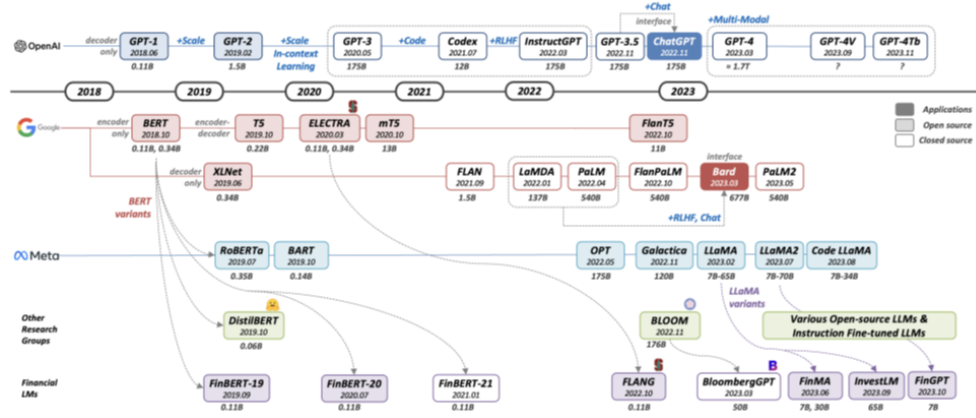


Fig. 1 Timeline showing the evolution of selected PLM/LLM releases from the general domain to the financial domain (Lee et al., 2024).

other models in a variety of NLP tasks because they can learn from extensive training sessions and find important patterns in financial data that they have never seen before (Zhang et al., 2023).

OpenAI’s GPT series and other large language models (LLMs) have made a lot of progress in natural language processing (NLP) in the last several years. The progress of these models marks a major step forward for AI to understand and create natural language. LLMs have become very good at understanding complex situations, answering questions, and creating content thanks to better processing power and more advanced algorithms. The abilities of LLMs are showing a lot of promise, especially in the finance industry (Chang et al., 2024, T. Wu et al., 2023, S. Wu et al., 2024). Many researchers believe that LLMs will change the way text is analyzed because they are easy to use, cheap, fast, and can be used for many different text analysis tasks, such as text annotation, categorization, sentiment analysis, and critical discourse analysis (Törnberg, 2023).

The fast growth of general-domain LLMs has led to the study of Financial LLMs (Fin-LLMs), which use methods like mixed-domain LLMs with prompt engineering and instruction fine-tuned LLMs with prompt engineering (Lee et al., 2024). Some of the widespread usage of these include the ability to change LLMs to classify financial news headlines (Luo and Gong, 2024), analyze financial statements and predict changes in future earnings (A. Kim et al., 2025), using for specific NLP tasks Araci, 2019. (Yu, Yao, et al., 2024) present FINCON, a multi-agent framework based on LLM that uses a manager-analyst hierarchy to help agents work together across functions to reach common goals through natural language interactions for a range of financial tasks. (J. J. Wang and Wang, 2025) focuses on building an intelligent financial data analysis system using LLM-RAG for financial analysis tasks, which provides important information for future improvements in intelligent financial data processing systems. FinBen, on the other and, is more than just a benchmark; it’s a research platform that offers new challenges, datasets, evaluation methods, and ways for the community to participate to push the boundaries of financial LLM development (Xie et al., 2024). (N. Wang et al., 2023) introduce FinGPT, a full open-source framework for financial large language models (FinLLMs). Fin-GPT stresses the need of gathering, cleaning, and preparing data for the development of open-source FinLLMs. (Yu, Li, et al., 2024) introduce FINMEM, the first LLM-based autonomous trading agent with a novel layered memory architecture and adaptable character design. Unlike previous LLM agents in finance, FINMEM has a memory module that can handle financial data from several sources. (Li et al., 2024) shows how INVESTORBENCH, a new and

broad financial benchmark, tests the reasoning and sequential decision-making skills of LLM-based agents in complex, open-ended financial situations.

2.3 Prompt Design and Sensitivity in LLMs

As LLMs get better, it’s important to understand how sensitive they are to prompts in order to make sure they work well and reliably (Anagnostidis and Bulian, 2024). Some of the current LLMs often don’t do well in the financial segment since there are big variations between general text data and financial text data (N. Wang et al., 2023). Also, LLMs usually agree with what the users think, even if it’s different from what they think, which shows a Clever Hans effect in LLMs (Anagnostidis and Bulian, 2024) and as a result, they often act like sycophants (Perez et al., 2023). For instance, (Binz and Schulz, 2023) shows how GPT-3 does several amazing things: it does vignette-based tasks as well as or better than humans, it makes good decisions based on descriptions, it beats people in a multi-armed bandit test, and it shows signs of model-based reinforcement learning. At the same time, even small changes to vignette-based tasks can throw GPT-3 off track, since it shows no signs of deliberate inquiry and does badly on a causal reasoning challenge. So, it’s really important to ask the right questions as the input prompts together with temperature settings affect how LLMs make decisions (Loya et al., 2023). LLMs can be made to work better for certain tasks by using methods including fine-tuning, in-context learning, and zero-/one-/few-shot learning (Du et al., 2023).

Prompt engineering is the methodical grouping of inputs and has become a crucial strategy for boosting the efficacy and precision of LLM models (Chen et al., 2025). The newest generation of LLMs may be pushed to do amazing things with zero-shot or few-shot performance in a lot of NLP tasks through prompt engineering (Leidinger et al., 2023). Small changes in prompt can go a long way to get the proper results. LLMs react to changes in prompts (task instructions, prompt structure, few-shot instances, and debiasing prompts) by looking at how well they do on tasks and how biased they are in social situations (Hida et al., 2024). So it is extremely important to notice how Large Language Models (LLMs) are quite sensitive to how prompts are written, which can have a big impact on how well they can give the right answers (“Benchmarking Prompt Sensitivity in Large Language Models”, 2025).

3 Language Modeling in Transformer-Based LLMs

LLMs aim to estimate the joint probability of a sequence of tokens $W = (w_1, w_2, \dots, w_n)$ as:

$$P(W) = P(w_1, w_2, \dots, w_n) = \prod_{i=1}^n P(w_i | w_1, w_2, \dots, w_{i-1})$$

Each conditional term $P(w_i | w_1, \dots, w_{i-1})$ represents the likelihood of token w_i given its preceding context. Exact computation is infeasible for long contexts, so early models simplified this using the Markov assumption:

$$P(w_i | w_1, \dots, w_{i-1}) \approx P(w_i | w_{i-(n-1)}, \dots, w_{i-1})$$

for small n , as in bigram or trigram models. However, such models fail to capture long-range dependencies.

Modern LLMs use transformer-based architectures to parameterize these conditionals with deep neural networks. Let θ represent the model parameters. For a sequence of embeddings $\mathbf{x}_1, \dots, \mathbf{x}_{i-1}$, the model computes a contextualized hidden state:

$$\mathbf{h}_i = f_{\theta}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{i-1})$$

An output layer then projects this hidden state into logits over the vocabulary:

$$\mathbf{z}_i = W_o \mathbf{h}_i + \mathbf{b}_o$$

The conditional distribution over the next token is computed using the softmax function:

$$P(w_i = v \mid w_1, \dots, w_{i-1}; \theta) = \frac{\exp(z_{i,v})}{\sum_{v'} \exp(z_{i,v'})}$$

The objective during training is to minimize the negative log-likelihood across the dataset \mathcal{D} :

$$\mathcal{L}(\theta) = - \sum_{W \in \mathcal{D}} \sum_{i=1}^{|W|} \log P(w_i \mid w_1, \dots, w_{i-1}; \theta)$$

Optimization is carried out via stochastic gradient descent or its derivatives, which enables the model to learn contextual dependencies considerably beyond the limits of fixed-length n -gram models. This end-to-end training technique allows huge language models to grasp complex semantic and syntactic links while producing fluent, coherent sequences.

The five LLMs investigated in this study—GPT, Perplexity, Claude, Gemini, and DeepSeek—are all built on this probabilistic modeling paradigm. All of the evaluation methods we used in this study measure how closely each model approximates human-like conditional token distributions in practice, as defined by the theoretical formulation $P(w_i \mid w_{<i}; \theta)$.

4 Methodology

This study looks at how well five of the most advanced transformer-based LLMs—GPT-4, Claude 4 Opus, Gemini Pro, Perplexity, and DeepSeek—can understand the Item 1. Business Section of yearly 10-K filings from seven U.S. tech companies dubbed as 'Magnificent 7'. The method includes clearly defined phases, such as building the corpus, creating prompts, running the model, and evaluating it in several ways.

4.1 Data

The "Magnificent Seven" tech companies are Apple, Microsoft, Amazon, Alphabet, Nvidia, Meta, and Tesla. Each of these companies files an annual 10-K report, from which we extract the Item 1: Business section. This section provides a narrative overview of the company's operations, strategic positioning, and market environment. We examine three recent fiscal years—2022, 2023, and 2024—resulting in a total of 21 documents (7 companies \times 3 years). From each 10-K filing, one representative text sample is extracted for model evaluation. We took the text out of HTML files we acquired from the SEC's EDGAR system. We had to do some manual preprocessing to make sure the format was the same, get rid of boilerplate components, and make sure the input structure was the same for all the companies and models.

4.2 Prompt Engineering and Tasks Designing

We created a series of 10 open-ended interpretative questions to test LLMs' capacity to extract, combine, deduce, and interpret financial information. We kept improving the prompts until we found a good mix between how easy they were to understand, how well they could be used in different situations, and how relevant they were to the subject. The last series of questions covers things like strategic intent, business model reasoning, risk inference, stakeholder framing, and looking ahead to the future. Each

question is meant to get analytical answers instead of just extractive summaries, so the models have to make conclusions or find hidden patterns. Each company-year document got its own set of prompts, so there were no memory artifacts or context carryover. To keep previous conversations from leaking into new ones, prompting was done in fresh, separate chat rooms for each model-document pair.

Here are the questions that we engineered:

1. What are the company’s indicated strategic goals for the next two to three years? Explain why you choose your response.
2. What is the company’s competitive position, and what proof do they have that this is the case?
3. What can we guess about the company’s growth strategy based on the description of the business?
4. What problems may be seen in the company’s stated business operations, even if they aren’t spelled out?
5. What kind of business strategy best fits this company? Tell me why you chose that.
6. How many different kinds of businesses does the corporation have? Does it depend on only one part of the business, or does it have multiple areas of operation that are in balance?
7. What does the company say its value proposition is, and who are its main stakeholders?
8. What should a stakeholder or investor remember most about this company?
9. How well does the company explain its business model on a scale of 1 to 5? Tell us why you gave it that score.
10. Give this business description a score based on how forward-looking it is. Does it talk about plans for the future, or does it mostly repeat what is now happening?

4.3 Setting Up the Model

We chose five of the best closed-source LLMs to test: GPT-4 (OpenAI), Claude 4 Opus (Anthropic), Gemini Pro (Google DeepMind), Perplexity AI, and DeepSeek-V2. The architectural size, training corpora, alignment goals, and deployment APIs of these models are all different. All of the models are transformer-based autoregressive decoders that were trained on big, multi-domain corpora to make the best guesses about what the next token will be. Responses were made using the default settings for temperature and top-p sampling, unless otherwise stated. Differences in latency, verbosity, and response completeness were qualitatively noticed but not employed as key rating criteria. All models were accessed using public-facing interfaces, and they were queried using deterministic or low-temperature decoding settings to make sure that the results were the same every time.

4.4 Evaluation Framework and Scoring Metrics

We used both human and automatic evaluation metrics to check the quality, relevance, and consistency of the model’s answers. There are three parts to the evaluation framework: human annotation, metric-based comparison, and model behavior diagnostics. Five human annotators scored the responses separately.

4.4.1 Human-Centric Evaluation

We used human evaluations to compare the relevance, depth, and clarity of model results to what experts expected. Each annotator gave each answer a score from 1 to 5 on five different axes:

- Relevance: How well the model answers.
- Completeness: if the answer included all the information that was asked for.
- Clarity: How well the answer flows from word to word, how well it makes sense, and how easy it is to read.
- Conciseness: How much the answer avoids using too much wordiness.
- Factual Accuracy: How much the output match the source 10-K.

4.4.2 Metric-Based Evaluation

We used quantitative evaluation measures to compare the lexical and semantic similarity between model outputs and reference responses (or peer model outputs). These were:

- ROUGE (1, 2, L): To find out how much model outputs and reference replies have in common at the word and phrase levels.
- Jaccard Similarity: This is a different way of looking at word-level set overlap than ROUGE.
- Cosine Similarity: This is an embedding-based similarity that uses Sentence-BERT (SBERT) to find words that are semantically close.

4.4.3 Model Behavior Diagnostics

We measured the following to see how consistent and generalizable they were:

- Cosine similarity across models
- Variance in similarity by prompt

We made sure that all prompts were submitted in new sessions (new chats) for each model and document so that there was no cross-contamination of previous context. There were cases of refusal, delusion, or formatting mistakes that needed to be looked at by hand.

5 Results

5.1 Overview

In this section, we look at five cutting-edge LLMs and see how well they can produce high-quality, contextually accurate, and insightful answers to questions taken from the Business sections of different 10-K filings. The evaluation is based on three main methods: human evaluation, metric-based evaluation, and behavioral evaluation.

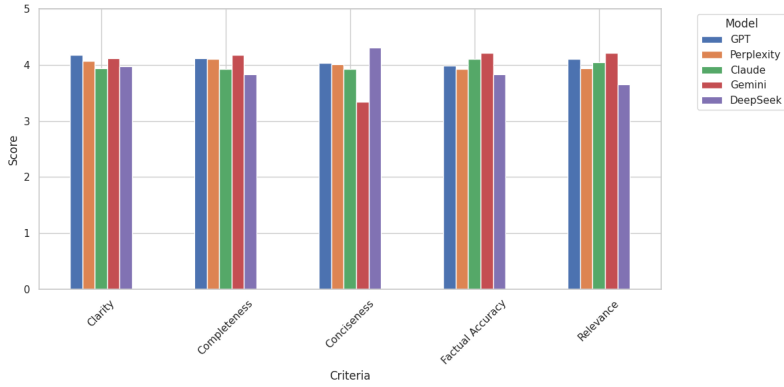
5.2 Human Annotator Evaluation

We did a structured human evaluation to systematically rate the qualitative performance of each LLM. Using a 1–5 Likert scale, five annotators scored prompt answers on five qualitative criteria. This gave us a more detailed picture of how each model answered the prompts beyond just being correct on the surface. The five criteria for evaluation are: Relevance, Completeness, Clarity, Conciseness, and Factual Accuracy. Five annotators (R1–R5) each gave a score to each answer on their own. Then, the scores were averaged across all annotators and prompts to get the final model-wise ratings. In Table 1, we can see the average scores that people gave each model on all five dimensions:

Figure 2 illustrates the results. The models that were tested showed that GPT was the best because it had the highest average score and did the best job with relevance,

Table 1 Human Evaluation Scores Across LLMs

<i>Rater</i>	Criteria	GPT	Perplexity	Claude	Gemini	DeepSeek
R1	Relevance	4.14	3.90	4.10	3.95	3.57
	Completeness	4.00	3.95	3.81	4.00	3.76
	Clarity	3.95	3.90	4.05	4.05	3.90
	Conciseness	3.81	4.00	4.14	2.95	4.24
	Factual Accuracy	4.00	3.95	4.19	4.43	3.71
R2	Relevance	4.10	3.95	3.95	4.33	3.57
	Completeness	4.24	4.29	3.86	4.29	3.90
	Clarity	4.24	4.24	4.00	4.21	3.86
	Conciseness	4.00	3.95	4.00	3.45	4.14
	Factual Accuracy	4.05	3.95	4.14	4.24	3.81
R3	Relevance	4.14	3.90	4.19	4.52	3.67
	Completeness	4.05	3.95	4.10	4.43	3.86
	Clarity	4.24	4.10	3.95	4.10	4.19
	Conciseness	4.00	4.21	3.86	3.24	4.52
	Factual Accuracy	4.10	4.10	4.19	4.57	4.00
R4	Relevance	4.05	4.00	4.10	4.38	3.57
	Completeness	4.29	4.29	3.86	4.14	3.71
	Clarity	4.14	4.14	3.76	4.29	4.00
	Conciseness	3.95	3.90	3.95	3.57	4.19
	Factual Accuracy	3.86	3.76	4.00	3.86	3.81
R5	Relevance	4.10	3.95	3.90	3.86	3.86
	Completeness	4.00	4.05	4.00	4.05	3.95
	Clarity	4.29	3.95	3.95	3.95	3.90
	Conciseness	4.43	4.00	3.67	3.52	4.43
	Factual Accuracy	3.95	3.90	4.00	3.95	3.81
Average		4.08	4.01	3.99	4.01	3.92

**Fig. 2** Human Evaluation Average Scores

thoroughness, and clarity while also being very accurate with the facts. Claude came in close behind. It had very high factual reliability and did well on most of the categories, but it wasn't quite as clear and complete as GPT. The results showed that DeepSeek was the shortest model, much shorter than the others. However, it was not as relevant or based on facts as the others. Perplexity got balanced results, doing well enough in all areas and not having any major flaws, though it did sometimes have small problems with factual correctness. Gemini lagged behind the other models because its answers were too long and its facts were not as clear, even though it spoke clearly. Overall, GPT and Claude turned out to be the most solid options. DeepSeek gave the shortest answers, and Perplexity was a good middle ground. Even though Gemini worked, it regularly did worse than its counterparts.

5.3 Metric Based Evaluation

We used the following automatic evaluation metrics to get a fair look at the lexical and semantic quality of the model responses: ROUGE (ROUGE-1, ROUGE-2, and ROUGE-L), Jaguar Similarity and Cosine Similarity (Sentence-BERT).

Table 2 Summary of Best Models and Their Scores by Metric

Metric	Best Model	Avg Score	Range
ROUGE-1	Gemini	0.56	0.20–0.62
ROUGE-2	Gemini	0.22	0.05–0.28
ROUGE-L	Gemini	0.16	0.06–0.17
Cosine Similarity	Claude	0.68	0.44–0.80
Jaccard	Gemini	0.21	0.07–0.26

Table 2 shows the strengths of each model by measure. Gemini’s ROUGE-1 score of 0.56 is 85% higher than rivals’, showing that it can copy exact words better than anyone else. It looks like semantic tasks are more competitive because Claude has a smaller lead over GPT (0.68) and Perplexity (0.71) in Cosine similarity (0.68). The Jaccard scores (Gemini: 0.21) show that Gemini is very good with wordings, and the edge in ROUGE-2 shows how good Gemini is with different phrases.

Table 3 Model Rankings Based on Average Metrics

Rank	Model	Avg R-1	Avg R-2	Avg R-L	Avg Cosine	Avg Jaccard
1	Gemini	0.56	0.22	0.16	0.63	0.22
2	GPT	0.31	0.08	0.10	0.68	0.13
3	Perplexity	0.29	0.08	0.10	0.71	0.13
4	Claude	0.27	0.08	0.09	0.67	0.12
5	DeepSeek	0.16	0.03	0.06	0.59	0.09

Table 3 shows how well the models did overall compared to the ground truth. Based on overall performance, Gemini is the best model. It has balanced numbers (R-1: 0.56, Cosine: 0.63), which makes it the safest choice. The semantics scores for GPT and Perplexity are almost the same (Cosine: 0.68 vs. 0.71), but GPT may do better on hybrid tasks because it has higher vocabulary scores (R-1: 0.31). Claude is a little behind, and DeepSeek is behind in all of the tests.

Table 4 Win Rates of Models Across Companies

Company	GPT	Perplexity	Claude	Gemini	DeepSeek
Amazon	58.60%	37.20%	25.10%	21.80%	19.20%
Apple	59.10%	38.90%	24.20%	22.30%	19.60%
Google	64.10%	40.10%	26.90%	20.90%	16.30%
Meta	61.00%	36.30%	22.30%	18.40%	16.80%
Microsoft	66.40%	41.20%	30.50%	18.60%	15.90%
Nvidia	62.50%	38.10%	23.90%	18.90%	14.70%
Tesla	62.30%	36.60%	23.20%	20.50%	19.10%

Table 4 shows how well the different LLMs usually do when put up against each other. This table gives us a general idea of how well the LLMs do when compared against each other. GPT comes out on top as the best model; depending on the

company, it wins an average of 58 to 66% of its matches. It is most dominant at Microsoft (66.4% win rate) and least dominant at Amazon (58.6% win rate), but it is still better than others. It looks like Perplexity is the best rival; its win rates range from 36-41%, which suggests it’s the second-best model overall. Claude is in the middle, with win rates in the low to mid-20s. Gemini and DeepSeek, on the other hand, are far behind, with win rates that never go above 22% in any company comparison.

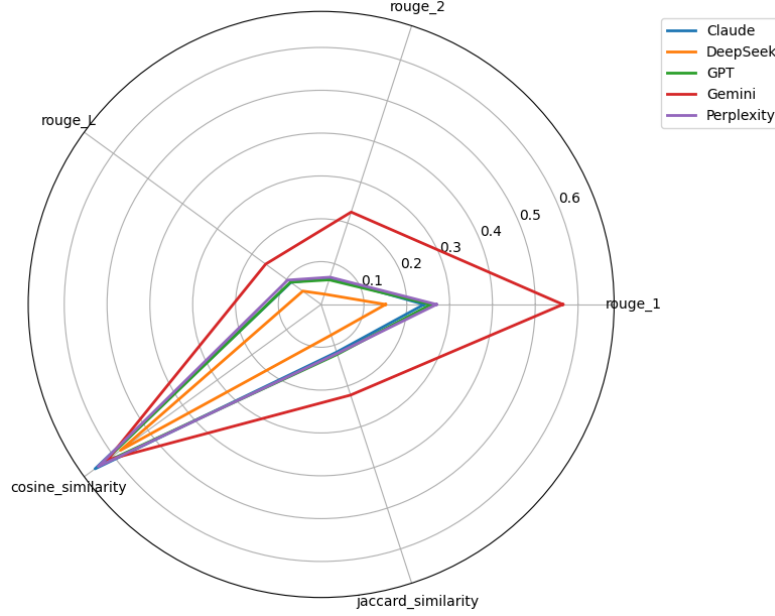


Fig. 3 Radar Chart – Average Similarity Metrics (vs Ground Truth)

Figure 3 shows how the LLMs fair on different evaluation metrics. All of the cosine similarity scores are high (≥ 0.85), with Claude (0.89) and Perplexity (0.88) slightly ahead of the rest. This means that the answers are strongly semantically aligned with the ground truth. ROUGE-1 puts Gemini ahead (0.32), while Claude, GPT, and Perplexity all score around 0.25, which means that there is some word-level agreement. ROUGE-2 and ROUGE-L also favor Gemini (0.18 and 0.27, respectively), which means that the sequences are more similar; the rest score less than 0.10. Gemini has the best Jaccard similarity score (0.41), while the others are all around 0.30, which means that there is less token-level overlap. (5) Except for cosine similarity (0.86), DeepSeek regularly does worse than other search engines. Gemini is better at lexical fidelity overall, while Claude and GPT put more emphasis on semantic coherence than surface resemblance.

5.4 Behavioral Diagnostics

We examined Across-Model Cosine Similarity and Prompt-Level Response Variance to find out how consistent and generalizable each model is across different inquiries.

Table 5 shows mean Cosine Similarity of LLM Outputs by company. Using the same business-related text prompts extracted from company filings, this table compares the similarity between the outputs produced by various LLMs. By calculating cosine similarity scores (ranging from 0 to 1) between all possible pairs of the models, it reveals how consistently these AI systems interpret and respond to the same input text across various companies. It is easy to see trends in the averaged scores: some model combinations consistently provide comparable results (e.g., GPT and Claude have a high level of agreement with 0.84 similarity for Apple), while other pairs differ

Table 5 Pairwise Similarity Scores Between Models Across Companies

Company	G-P	G-C	G-G	G-D	P-C	P-G	P-D	C-G	C-D	G-Dk
Apple	0.77	0.84	0.77	0.84	0.76	0.69	0.77	0.79	0.84	0.78
Amazon	0.80	0.82	0.79	0.80	0.78	0.72	0.71	0.71	0.77	0.77
Alphabet	0.83	0.83	0.82	0.76	0.78	0.85	0.77	0.78	0.71	0.80
Meta	0.78	0.84	0.82	0.76	0.76	0.79	0.75	0.75	0.79	0.75
Microsoft	0.89	0.87	0.85	0.82	0.84	0.86	0.82	0.83	0.80	0.78
Nvidia	0.81	0.78	0.81	0.77	0.88	0.86	0.73	0.84	0.72	0.72
Tesla	0.81	0.79	0.79	0.78	0.76	0.79	0.75	0.74	0.80	0.75

more noticeably (e.g., Perplexity and Gemini have a much closer relationship with Amazon at only 0.72 similarity). The resulting content can be significantly affected by the choice of LLM, even while processing the same source material, as these changes are most likely caused by differences in the models’ training data, architectures, and algorithms.

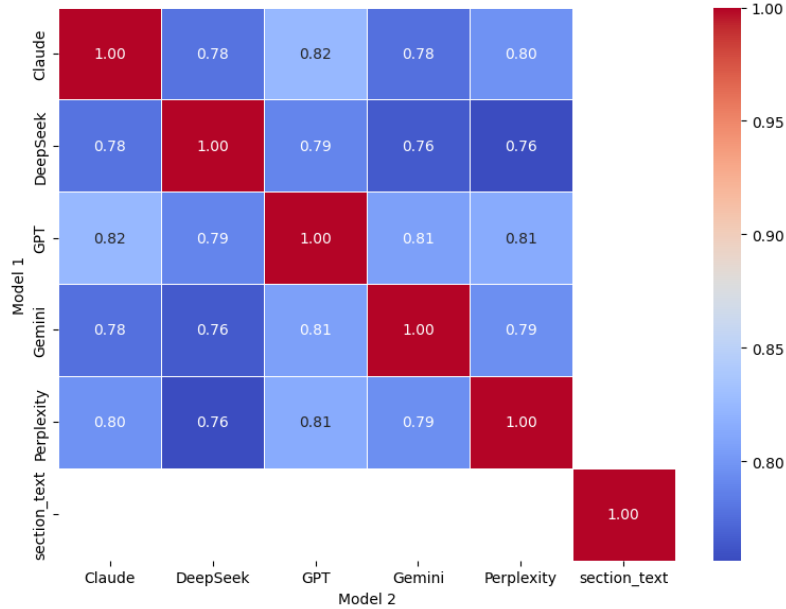
**Fig. 4** Average Cosine Similarity Between All LLMs

Figure 4 illustrates the pairwise semantic similarity between the outputs of the LLMs and the ground truth. Cosine similarity scores, ranging from 0 to 1, assess the alignment of meaning between model responses. GPT consistently exhibits the highest semantic similarity, attaining scores of 0.82 with Claude, 0.81 with Gemini and Perplexity, and 0.81 with the human reference—indicating that its outputs are most aligned with both peer models and the ground truth. We also see DeepSeek shows the least similarity, obtaining a score of 0.76 with both Gemini and the standard text, whereas Claude shows a strong alignment with the reference (0.80). The matrix shows that while DeepSeek’s outputs most significantly depart from human-generated information, GPT serves as the semantic median across all models.

Using mean cosine similarity and standard deviation, table 6 displays the consistency with which LLMs perceive identical business prompts across firms and years. The model agreement is highest for Microsoft (mean: 0.84-0.85, low standard deviation) and lowest for Amazon’s 2024 prompt (mean: 0.71, high standard deviation: 0.078). Nvidia and Tesla exhibit greater volatility in prior years, whereas Apple and Alphabet

Table 6 Cosine Similarity Scores and Variability Across Prompts

Company	Year	Prompt ID	Mean Cosine	Std. Dev
Alphabet	2022	Alphabet_2022	0.778	0.070
	2023	Alphabet_2023	0.787	0.041
	2024	Alphabet_2024	0.812	0.034
Amazon	2022	Amazon_2022	0.786	0.053
	2023	Amazon_2023	0.807	0.050
	2024	Amazon_2024	0.708	0.078
Apple	2022	Apple_2022	0.767	0.063
	2023	Apple_2023	0.774	0.055
	2024	Apple_2024	0.820	0.035
Meta	2022	Meta_2022	0.750	0.055
	2023	Meta_2023	0.789	0.038
	2024	Meta_2024	0.802	0.043
Microsoft	2022	Microsoft_2022	0.846	0.029
	2023	Microsoft_2023	0.851	0.042
	2024	Microsoft_2024	0.810	0.041
Nvidia	2023	Nvidia_2023	0.765	0.069
	2024	Nvidia_2024	0.806	0.064
	2025	Nvidia_2025	0.804	0.056
Tesla	2022	Tesla_2022	0.751	0.062
	2023	Tesla_2023	0.766	0.036
	2024	Tesla_2024	0.813	0.038

display enhanced stability over time. The findings indicate that both the input data and the specific combinations of LLMs influence the extent of output similarity.

6 Discussion

6.1 Overall

Our experiments used the LLMs to analyze 10-Ks in strict, context-isolated settings. This gave us a clear picture of LLMs relative strengths and weaknesses. From a human evaluation point of view, GPT is the best model as it consistently gives relevant, complete, and clear answers while staying factually correct, followed by Perplexity. Claude is a better choice for applications that need content validity because it is more reliable when it comes to facts, even though it is slower. DeepSeek is good at being succinct, but it sacrifices relevance and factual correctness, which makes it less useful for thorough financial research. Gemini is very flexible in its syntax, but its long-windedness and lack of conciseness make it less useful in practice.

Metrics based evaluations support the human annotation almost all of the time. Gemini gets the best scores on lexical fidelity tests (ROUGE-1/2 and Jaccard), which means it can extract information very well. But the fact that it doesn’t have very high semantic alignment and gets lower scores from people illustrates that lexical precision alone isn’t adequate for full understanding. On the other hand, Claude and Perplexity show better semantic coherence, which is in line with GPT’s balanced profile that stresses semantic depth together with enough lexical precision. The difference between lexical and semantic metrics in different models and settings shows how important it is to have a multidimensional assessment that is unique to the work at hand.

Behavioral diagnostics show that GPT and Claude agree profusely on the meaning of words, which shows that their interpretive frameworks are similar and suggests that they may have similar architecture or training. On the other hand, Gemini and DeepSeek show a lot of differences between models and over time, which suggests that they would not be able to be used consistently in critical financial areas. Temporal analyses show that changes in company disclosures effect model concordance. This shows how important it is to keep validating models when financial terms change.

These results show that LLM responses are naturally different from one another because of both the unique features of each model and the difficulties that come with each question. In finance, where high dependability and consistent interpretation are important, models that show strong semantic alignment and little response variability are very important. Our results also show that using multiple models together or hybrid frameworks can help reduce the biases of individual models and make them more resilient in real-world situations. In the end, GPT is the best and most reliable model for analyzing all kinds of financial text. It beats human judgment, automated metrics, and behavioral consistency, all while keeping operational efficiency. Gemini and Claude’s lexical accuracy is good for specific uses that need exact phrase replication, but it comes at the cost of interpretive flexibility. DeepSeek and Perplexity have similar features, but they aren’t good enough to be used as the main tools for high-stakes financial analysis.

6.2 Annotator-based rankings

According to the human rankings, GPT came out on top in all categories, with an average score of 4.08. GPT consistently came out on top in terms of relevance, completeness, and clarity, showing that it can give answers that are relevant, full of information, and clear. Notably, it kept up its competitive performance in factual correctness (4.00), which shows that having a high level of language fluency didn’t mean that the material wasn’t valid. Claude came in third overall (3.99), but it stood out for being the most reliable (4.10 average), just barely beating all the other models in this category. Claude was a little less succinct than DeepSeek and a little less complete than GPT, but it did well on all criteria, giving it a strong model in areas where accuracy is very important.

DeepSeek had the highest conciseness score (4.28), but its performance profile showed that it surrendered depth of content for brevity. It scored worse than GPT and Claude in both relevance (3.65) and factual correctness (3.83), which suggests that while its outputs were quick, they occasionally didn’t have the depth or alignment needed for financial text analysis. Perplexity did the same thing on all criterion, getting barely over 4.00 in most of them. This balance shows that the utility is generally reliable and doesn’t have any major strengths or weaknesses. But it did have small problems in factual correctness (3.93) and relevance (3.94), and sometimes GPT and Claude did better than it. Gemini had the best syntax, but it came in last overall (3.89). It did poorly in terms of conciseness (3.55) and completeness (4.06), sometimes giving long, vague answers. Some raters gave it high scores for factual accuracy (for example, R3 gave it a score of 4.57), but its overall variation implies that it is less consistent and tends to be long-winded without adding any new information.

The radar chart provides a comparative analysis of five LLMs—based on five qualitative evaluation criteria: Clarity, Completeness, Conciseness, Factual Accuracy, and Relevance. Each axis denotes a criterion, with performance scores ranging from 0 to 5, and overlapping filled areas that indicate the relative efficacy of each model. GPT (blue) consistently demonstrates a balanced and robust performance in all evaluation dimensions. It has a minor advantage in both Clarity and Relevance, which suggests that it is capable of generating responses that are meaningful, contextually appropriate, and clear. The model’s performance indicates that it is generally reliable across tasks, with no significant defects. Perplexity (green) and Claude (orange) show similar performance and are comparable, with only modest trade-offs. Claude exhibits a minor advantage in Factual Accuracy, while Perplexity maintains consistently high scores in Clarity and Completeness. Although Claude may be slightly more assertive in empirical consistency, both models seem to be well-rounded. Gemini (red) exhibits a significant decrease in Conciseness and is the least concise of the

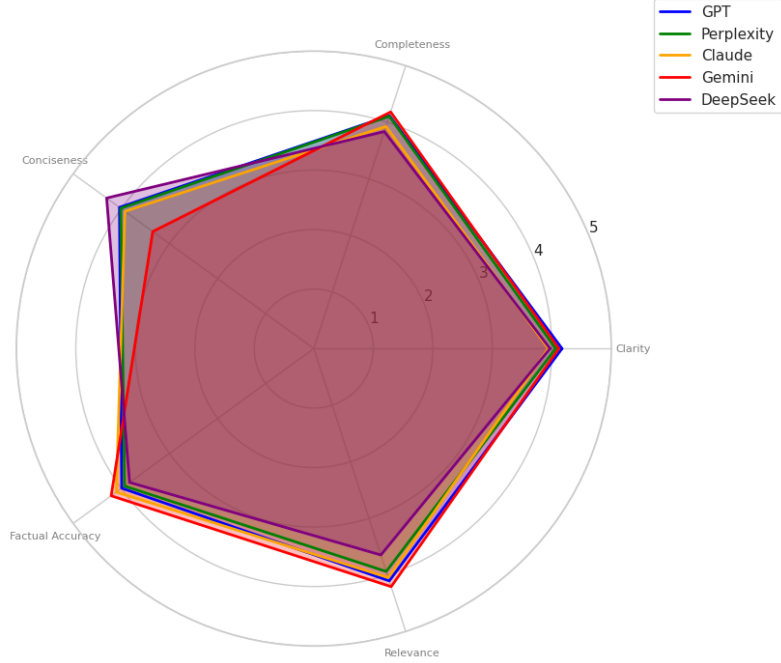


Fig. 5 Comparative analysis across the evaluation criteria

five models. We noticed Gemini’s responses were excessively verbose and lacked in brevity and impeded interpretability. Nevertheless, Gemini’s verbosity does not undermine the informational value of its responses, as evidenced by its relatively strong performance in Factual Accuracy and Relevance. As for Conciseness, DeepSeek (purple) has achieved the highest score. We noticed DeepSeek generates succinct, focused responses. However, it exhibits comparatively lower scores in Factual Accuracy and Relevance, suggesting that its brevity comes at the expense of context alignment as well as profundity. All of the The models’ somewhat parallel performance in fundamental dimensions such as Clarity, Completeness, and Factual Accuracy is indicated by the substantial overlap across the spectrums. Nonetheless, Conciseness and Relevance are the primary differentiators. Although no single model is inherently superior and each model possesses unique strengths and disadvantages, GPT is the most balanced and reliable. Gemini, Claude or GPT are more suitable options for producing outputs that are contextually rich and accurate whereas DeepSeek may be the most suitable option for applications that prioritize brevity.

Evaluator Agreement and Consistency Across the five annotators, the ratings had a reasonably low inter-rater variance, which suggests that there was a high level of agreement on what was relevant, clear, and complete in this financial context. There were times when individual raters disagreed on how to score for conciseness or factual accuracy, especially for Gemini and DeepSeek. However, the overall model-wise trends were strong and consistent. Appendix B, Table B1 has the full rater-level data and the standard deviation analysis.

6.3 Automated Metric-based Evaluation

We used a set of complimentary automated criteria to objectively measure how lexically and semantically accurate LLM outputs were compared to ground truth responses. These are ROUGE-1, ROUGE-2, and ROUGE-L; Jaccard Similarity; and Cosine Similarity.

Gemini does better than all the other models in the ROUGE metrics—ROUGE-1 (0.56), ROUGE-2 (0.22), and ROUGE-L (0.16)—which assess n-gram and sequence-level overlap and shows that the texts are very similar on the surface. Gemini also has the highest Jaccard similarity score (0.21), which shows that it can keep token-level uniqueness and intersection. For Cosine similarity, Claude has the highest average score of 0.68 and the highest maximum score of 0.80. These results show that Gemini is better at catching lexical aspects and syntactic overlaps, while Claude is better at capturing the overall meaning of expert responses (See 2).

Gemini routinely beats other models on lexical-based metrics, with ROUGE-1 scores average 0.56, which is around 85% higher than those of other models. It also has a threefold lead in ROUGE-2, which shows that it can replicate phrases better. Claude is in the lead in semantic evaluation with an average Cosine Similarity of 0.68, but this is only a little margin over GPT and Perplexity, which shows that the competition in semantic comprehension is getting stronger. The Jaccard Similarity backs up Gemini’s lexical dominance (0.21 average), making it even more reliable for tasks that need exact token reproduction. Gemini is the best model overall when all of its lexical and semantic metrics are taken into account. This makes it the best choice for applications that need both accuracy and semantic nuance. GPT and Perplexity have similar semantic strengths but slightly lower lexical scores, which means they are better for activities that focus on getting the meaning rather than copying it word for word. Claude is in the middle, while DeepSeek’s consistently lower results across metrics suggest that it is not very useful for high-precision needs (See Table 3).

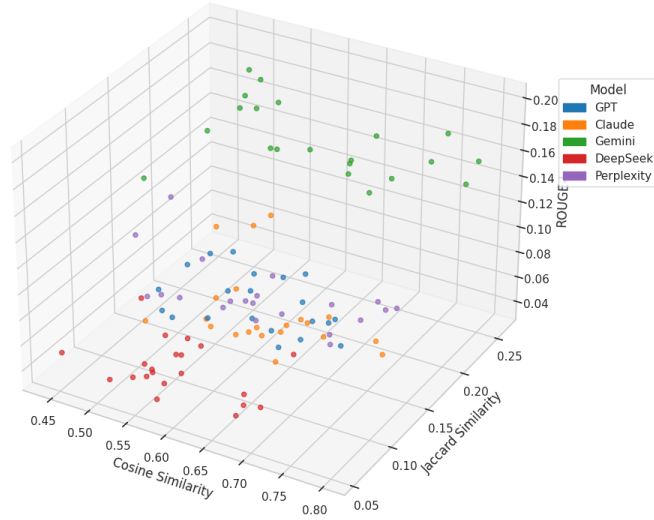


Fig. 6 3-D Plot of Similarity Matrix

The plot 6 shows how well the model works based on cosine similarity (x-axis), Jaccard similarity (y-axis), and ROUGE-L (z-axis). GPT and Claude are both in the high-performing area, with cosine values between 0.65 and 0.80, Jaccard values between 0.15 and 0.22, and ROUGE-L values up to 0.20. This shows that they are quite similar to the ground truth. DeepSeek does well linguistically (cosine \geq 0.70), but not so well structurally (ROUGE-L \leq 0.08, Jaccard 0.06–0.12). Perplexity has scores that are modest and balanced on all measures. Gemini is the most distributed and has the worst performance, with several outputs scoring less than 0.55 (cosine), 0.08 (ROUGE-L), and 0.06 (Jaccard), which means that the words and meanings are not very similar.

When it comes to head-to-head success rates, assessments across the dataset demonstrate that GPT consistently does well, with victory percentages between 58% and 66%, and Microsoft leading the way at 66.4%. Perplexity is a small opponent, with a win percentage of 36% to 41%, whereas Claude’s win rate in the mid-20s shows that the competition is not very tough. Gemini has better vocabulary, but both Gemini and DeepSeek have lower relative success rates (less than 22%). This could be because GPT and Perplexity are better at understanding meaning and adapting to new situations (See Table 4).

6.4 Behavioral Outcome

The differences in lexical and semantic metrics, as well as model-specific strengths, show how important it is to choose models that are best for certain tasks. The fact that the Cosine Similarity scores only go from 0.63 to 0.71, while the lexical metrics (ROUGE-1: 0.16-0.56) have a wider range of values, shows that the semantic measures alone may not be enough to tell the difference between model performances. This shows how important it is to use multiple metrics in full LLM benchmarking.

The pairwise semantic similarity table (See Table 5) indicates that various LLMs understand the same financial inputs in quite different ways. This shows that the internal representations can be the same and different at the same time. For instance, GPT and Claude always have the best semantic alignment across all firm datasets, with cosine similarity scores above 0.8 (0.84 for Apple and 0.87 for Microsoft, for example). This strong agreement suggests that the training datasets for both models are probably the same or that their semantic embedding spaces are similar. This means that they employ similar ways of understanding when they know how to speak about money. On the other hand, combos that include Gemini and DeepSeek often show less semantic agreement, especially when comparing Perplexity and Gemini, where similarity scores might drop as low as 0.69-0.72 (for example, Amazon). Changes in model architecture, differences in pretraining data, and the goals of fine-tuning are probably to blame for the inconsistencies.

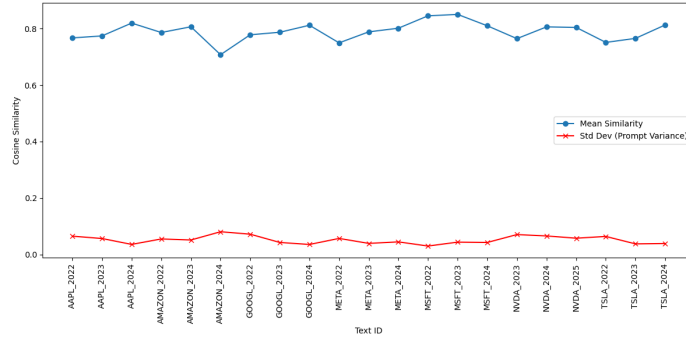


Fig. 7 Mean and Variance of Cosine Similarity

Figure 7 shows how model agreement changes from year to year and company to company. Microsoft’s filings consistently show the strongest cross-model consensus (mean cosine similarity = 0.85 with little variation), which means that timely content in this area leads to steady and strong semantic interpretations. On the other hand, Amazon’s 2024 prompts have the lowest inter-model similarity (mean 0.71, standard deviation 0.078), which means that they are either more sophisticated or feature new financial disclosures that make it harder for the models to agree (See Table 6).

7 Conclusion

LLMs have ushered in a new era of AI—machines capable of reasoning, comprehension, and decision-making. These models have revolutionized various industries by producing, comprehending, and adapting language for certain tasks. Despite their effectiveness in producing High-Parameter Models (HPMs), some LLMs have come under fire for being too complicated and difficult for the average user to grasp (Mohsin and Nasim, 2025). As critical domains such as finance are increasingly becoming reliant on these systems, it is high time knowing the reasoning behind a model’s output. By providing an in-depth, multi-faceted evaluation of five state-of-the-art transformer-based LLMs for financial text analysis, this paper provides an immediate solution to that challenge. Within a controlled experiment that distinguished between context and standardized inputs, we evaluated the model’s performance using automated similarity measurements, behavioral diagnostics, and human annotation. The findings demonstrate that GPT outshines all other generalist models in terms of factual correctness, semantic fluency, and operational robustness. Up next came Perplexity, who was more precise and consistent with his cues, although he moved more slowly and spoke more slowly than the others. No matter the criteria, Claude’s performance was consistently moderate. When compared to DeepSeek and Gemini, the latter two exhibit less generalization, more erratic answers, and inconsistent semantics.

Among other things, our findings highlight the critical fact that performance is insufficient. The simplicity, adaptability, and consistency of these models’ reasoning routes are of utmost importance when they are used to make financially consequential decisions with real-world consequences. The "black box" nature of LLMs is a major concern, notwithstanding their strength. Based on our findings, there is a potential practical solution to this lack of transparency. By examining which models perform well and providing explanations for how they operate, change, and apply to new scenarios under standard conditions, our study contributes to the ongoing discourse regarding the responsible use of LLMs. Consequences of AI-driven financial systems, methods to improve model alignment, and causal interpretability should all be part of future research that expands on this paradigm.

References

- Abdaljalil, S., & Bouamor, H. (2021). An exploration of automatic text summarization of financial reports. *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, 1–7. Retrieved July 16, 2025, from <https://aclanthology.org/2021.finnlp-1.1.pdf>
- Anagnostidis, S., & Bulian, J. (2024, August). How Susceptible are LLMs to Influence in Prompts? [arXiv:2408.11865 [cs]]. <https://doi.org/10.48550/arXiv.2408.11865>
- Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv 2019. *arXiv preprint arXiv:1908.10063*.
- Benchmarking Prompt Sensitivity in Large Language Models [ISSN: 0302-9743, 1611-3349]. (2025). In *Lecture Notes in Computer Science* (pp. 303–313). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-88714-7_29
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3 [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, 120(6). <https://doi.org/10.1073/pnas.2218523120>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2022, July). On the Opportunities and Risks of Foundation Models [arXiv:2108.07258 [cs]]. <https://doi.org/10.48550/arXiv.2108.07258>
Comment: Authored by the Center for Research on Foundation Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI). Report page with citation guidelines: <https://crfm.stanford.edu/report.html>.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models [Publisher: Association for Computing Machinery (ACM)]. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45. <https://doi.org/10.1145/3641289>
- Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models [Publisher: Elsevier]. Retrieved July 11, 2025, from [https://www.cell.com/patterns/fulltext/S2666-3899\(25\)00108-4](https://www.cell.com/patterns/fulltext/S2666-3899(25)00108-4)
- Daimi, S. A., & Iqbal, A. (2024, September). A Scalable Data-Driven Framework for Systematic Analysis of SEC 10-K Filings Using Large Language Models [arXiv:2409.17581 [cs]]. <https://doi.org/10.48550/arXiv.2409.17581>
Comment: 10 pages, 7 figures.
- Du, Y., Watkins, O., Wang, Z., Colas, C., Darrell, T., Abbeel, P., Gupta, A., & Andreas, J. (2023). Guiding pretraining in reinforcement learning with large language models. *International Conference on Machine Learning*, 8657–8677. Retrieved July 17, 2025, from <http://proceedings.mlr.press/v202/du23f.html>
- Hida, R., Kaneko, M., & Okazaki, N. (2024, July). Social Bias Evaluation for Large Language Models Requires Prompt Variations [arXiv:2407.03129 [cs]]. <https://doi.org/10.48550/arXiv.2407.03129>
- Huang, A. H., Wang, H., & Yang, Y. (2023). *FinBERT: A Large Language Model for Extracting Information from Financial Text* [Publisher: Wiley]. *Contemporary Accounting Research*, 40(2), 806–841. <https://doi.org/10.1111/1911-3846.12832>

- Kim, A., Muhn, M., & Nikolaev, V. (2025, February). Financial Statement Analysis with Large Language Models [arXiv:2407.17866 [q-fin]]. <https://doi.org/10.48550/arXiv.2407.17866>
 Comment: A co-author identified inconsistencies in the data and analyses while attempting to replicate past analyses from the working paper. Accordingly, we have temporarily withdrawn the working paper from circulation while we review the research findings.
- Kim, H., Lee, E., & Yoo, D. (2023). Do SEC filings indicate any trends? Evidence from the sentiment distribution of forms 10-K and 10-Q with FinBERT [Issue: 2 Publisher: Emerald Publishing Limited]. Retrieved July 11, 2025, from <https://www.emerald.com/insight/content/doi/10.1108/dta-05-2022-0215/full/html>
- Kim, T. T., Makutonin, M., Sirous, R., & Javan, R. (2025). Optimizing Large Language Models in Radiology and Mitigating Pitfalls: Prompt Engineering and Fine-tuning [Publisher: Radiological Society of North America (RSNA)]. *RadioGraphics*, 45(4). <https://doi.org/10.1148/rg.240073>
- Lazarev, A., & Sedov, D. (2024). Utilizing Modern Large Language Models (LLM) for Financial Trend Analysis and Digest Creation. *2024 6th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, 317–321. Retrieved July 15, 2025, from <https://ieeexplore.ieee.org/abstract/document/10803746/>
- Lee, J., Stevens, N., Han, S. C., & Song, M. (2024). A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*. Retrieved July 16, 2025, from <https://smallake.kr/wp-content/uploads/2024/10/2402.02315v1.pdf>
- Leidinger, A., Rooij, R. v., & Shutova, E. (2023, November). The language of prompting: What linguistic properties make a prompt successful? [arXiv:2311.01967 [cs]]. <https://doi.org/10.48550/arXiv.2311.01967>
 Comment: Accepted to EMNLP 2023 Findings.
- Li, H., Cao, Y., Yu, Y., Javaji, S. R., Deng, Z., He, Y., Jiang, Y., Zhu, Z., Subbalakshmi, K., Xiong, G., Huang, J., Qian, L., Peng, X., Xie, Q., & Suchow, J. W. (2024, December). INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-based Agent [arXiv:2412.18174 [cs]]. <https://doi.org/10.48550/arXiv.2412.18174>
- Lombardo, G., Trimigno, G., Pellegrino, M., & Cagnoni, S. (2024). Language Models Fine-Tuning for Automatic Format Reconstruction of SEC Financial Filings [Publisher: IEEE]. Retrieved July 11, 2025, from <https://ieeexplore.ieee.org/abstract/document/10445214/>
- Loya, M., Sinha, D. A., & Futrell, R. (2023). Exploring the Sensitivity of LLMs’ Decision-Making Capabilities: Insights from Prompt Variation and Hyperparameters [arXiv:2312.17476 [cs]], 3711–3716. <https://doi.org/10.18653/v1/2023.findings-emnlp.241>
 Comment: EMNLP 2023.
- Luo, W., & Gong, D. (2024, January). Pre-trained Large Language Models for Financial Sentiment Analysis [arXiv:2401.05215 [cs]]. <https://doi.org/10.48550/arXiv.2401.05215>
- Mohsin, M. T., & Nasim, N. B. (2025, March). Explaining the Unexplainable: A Systematic Review of Explainable AI in Finance [arXiv:2503.05966 [q-fin]]. <https://doi.org/10.48550/arXiv.2503.05966>
 Comment: 2 tables, 11 figures.
- Nakagawa, K., Hirano, M., & Fujimoto, Y. (2024). Evaluating company-specific biases in financial sentiment analysis using large language models. *2024 IEEE International Conference on Big Data (BigData)*, 6614–6623. Retrieved July 15, 2025, from <https://ieeexplore.ieee.org/abstract/document/10826008/>

- Oyewole, A. T., Adeoye, O. B., Addy, W. A., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Automating financial reporting with natural language processing: A review and case analysis. *World Journal of Advanced Research and Reviews*, 21(3), 575–589. Retrieved July 15, 2025, from https://www.researchgate.net/profile/Adedoyin-Oyewole-2/publication/379429762_Automating_financial_reporting_with_natural_language_processing_A_review_and_case_analysis/links/6611a44c2034097c54fb7559/Automating-financial-reporting-with-natural-language-processing-A-review-and-case-analysis.pdf
- Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., & Kadavath, S. (2023). Discovering language model behaviors with model-written evaluations. *Findings of the association for computational linguistics: ACL 2023*, 13387–13434. Retrieved July 16, 2025, from <https://aclanthology.org/2023.findings-acl.847/>
- Prompt Engineering in Large Language Models [ISSN: 2524-7565, 2524-7573]. (2024). In *Algorithms for Intelligent Systems* (pp. 387–402). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-7962-2_30
- Sehrawat, S. (2019). Learning word embeddings from 10-K filings for financial NLP tasks. Available at SSRN 3480902. Retrieved July 15, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3480902
- Textual Analysis in Finance — Annual Reviews. (n.d.). Retrieved July 16, 2025, from <https://www.annualreviews.org/content/journals/10.1146/annurev-financial-012820-032249>
- Törnberg, P. (2023). How to use large language models for text analysis. *arXiv preprint arXiv:2307.13106*. Retrieved July 16, 2025, from https://www.researchgate.net/profile/Petter-Toernberg/publication/377559012_How_to_Use_Large-Language_Models_for_Text_Analysis/links/671f7c16df4b534d4efd515e/How-to-Use-Large-Language-Models-for-Text-Analysis.pdf
- Wang, J. J., & Wang, V. X. (2025, June). Assessing Consistency and Reproducibility in the Outputs of Large Language Models: Evidence Across Diverse Finance and Accounting Tasks [arXiv:2503.16974 [q-fin]]. <https://doi.org/10.48550/arXiv.2503.16974>
Comment: 89 pages, 20 tables, 15 figures.
- Wang, L., Cheng, Y., Xiang, A., Zhang, J., & Yang, H. (2024, June). Application of Natural Language Processing in Financial Risk Detection [arXiv:2406.09765 [q-fin]]. <https://doi.org/10.48550/arXiv.2406.09765>
- Wang, N., Yang, H., & Wang, C. D. (2023, November). FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets [arXiv:2310.04793 [cs]]. <https://doi.org/10.48550/arXiv.2310.04793>
Comment: Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022, October). Emergent Abilities of Large Language Models [arXiv:2206.07682 [cs]]. <https://doi.org/10.48550/arXiv.2206.07682>
Comment: Transactions on Machine Learning Research (TMLR), 2022.
- Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2024). BloombergGPT: A large language model for finance, 2023. URL <https://arxiv.org/abs/2303.17564>.
- Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development [Publisher: IEEE]. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122–1136.

Retrieved July 18, 2025, from <https://ieeexplore.ieee.org/abstract/document/10113601/>

- Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., & Feng, D. (2024). Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37, 95716–95743. Retrieved July 11, 2025, from https://proceedings.neurips.cc/paper_files/paper/2024/hash/adb1d9fa8be4576d28703b396b82ba1b-Abstract-Datasets_and_Benchmarks_Track.html
- Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., Liu, R., Suchow, J. W., & Khashanah, K. (2024). Finmem: A performance-enhanced llm trading agent with layered memory and character design [Issue: 1]. *Proceedings of the AAAI Symposium Series*, 3, 595–597. Retrieved July 16, 2025, from <https://ojs.aaai.org/index.php/AAAI-SS/article/view/31290>
- Yu, Y., Yao, Z., Li, H., Deng, Z., Jiang, Y., Cao, Y., Chen, Z., Suchow, J., Cui, Z., & Liu, R. (2024). Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37, 137010–137045. Retrieved July 16, 2025, from https://proceedings.neurips.cc/paper_files/paper/2024/hash/f7ae4fe91d96f50abc2211f09b6a7e49-Abstract-Conference.html
- Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X.-Y. (2023). Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. *4th ACM International Conference on AI in Finance*, 349–356. <https://doi.org/10.1145/3604237.3626866>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., & Dong, Z. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2). Retrieved July 18, 2025, from https://www.researchgate.net/profile/Tang-Tianyi-3/publication/369740832_A_Survey_of_Large_Language_Models/links/665fd2e3637e4448a37dd281/A-Survey-of-Large-Language-Models.pdf
- Zhou, H., Xu, K., Bao, Q., Lou, Y., & Qian, W. (2024). Application of conversational intelligent reporting system based on artificial intelligence and large language models. *Journal of Theory and Practice of Engineering Science*, 4(03), 176–182. Retrieved July 15, 2025, from <https://centuryscipub.com/index.php/jtpes/article/view/525>