SUERF
The European Money
and Finance Forum

SUERF Policy Brief
No 1008, October 2024

Large Language Models and the Future of Financial Analysis*

Alex Kim, Maximilian Muhn, and Valeri V. Nikolaev
University of Chicago Booth School of Business

Keywords: GPT4, neural network, asset pricing, earnings, direction of earnings changes, analysts, chain-of-thought, financial statement analysis, large language models

JEL codes: G12, G14, G41, M41

Abstract

We investigate whether an LLM can successfully perform financial statement analysis in a way similar to a professional human analyst. We provide standardized and anonymous financial statements to GPT4 and instruct the model to analyze them to determine the direction of future earnings. Even without any narrative or industry-specific information, the LLM outperforms financial analysts in its ability to predict earnings changes. The LLM exhibits a relative advantage over human analysts in situations when the analysts tend to struggle. Furthermore, we find that the prediction accuracy of the LLM is on par with the performance of a narrowly trained state-of-the-art ML model. LLM prediction does not stem from its training memory. Instead, we find that the LLM generates useful narrative insights about a company's future performance. Lastly, our trading strategies based on GPT's predictions yield a higher Sharpe ratio and alphas than strategies based on other models. Taken together, our results suggest that LLMs may take a central role in decision-making.

*This policy brief is based on Kim, A., Muhn, M., & Nikolaev, V. “Financial Statement Analysis with Large Language Models” SSRN Working Paper.

Introduction

The recent emergence of Large Language Models (LLMs) has sparked intense debate about the potential of generative AI. Although such tools have been widely adopted for various tasks in the financial domain, such as summarization, information extraction, or report writing, it is still unclear to what extent LLMs can play a critical role in financial markets. Therefore, in our recent research project, we investigate whether an LLM can effectively replicate or even surpass human financial analysts (as well as narrowly trained machine learning models) in predicting firms' future performance. Based on our recent SSRN working paper, this policy brief highlights the performance of LLMs in this challenging quantitative task and discusses the broader implications for financial markets and decision-making processes.

Context of the Earnings Prediction Task

Earnings prediction is a cornerstone of financial analysis and a critical input for valuation models or stock recommendations. Financial analysts forecast future earnings by scrutinizing a company's financial statements and assessing its financial health and growth potential. This process often involves a blend of quantitative analysis, contextual interpretation, and professional judgment, relying on a wealth of domain-specific knowledge and experience.

Our study examines whether an LLM like GPT-4 can perform this complex task on par with, or better than, human analysts and specialized machine learning models. We focus on a scenario where GPT is provided with only anonymized and standardized financial statements. We then aim to assess its capability to generate meaningful insights from purely numerical data.

Methodological Approach

Our research design involves presenting two primary financial statements (the balance sheet and income statement) to GPT in a standardized form. The model is tasked with analyzing these documents and determining whether the firm's earnings will rise or fall in the next period. Unlike traditional approaches that use narrative context to inform predictions, we deliberately strip any narrative context like management discussions or industry-specific insights from the model input. Therefore, we evaluate the LLM's ability to derive insights strictly based on numbers.

The approach is two-fold:

- Data Preparation: Financial statements are anonymized to prevent GPT from relying on any specific knowledge from its training data about the firm or the period. Company names are omitted, and years are replaced with generic labels (e.g., "Period t" and "Period t-1"). The format is standardized across firms to ensure uniformity.

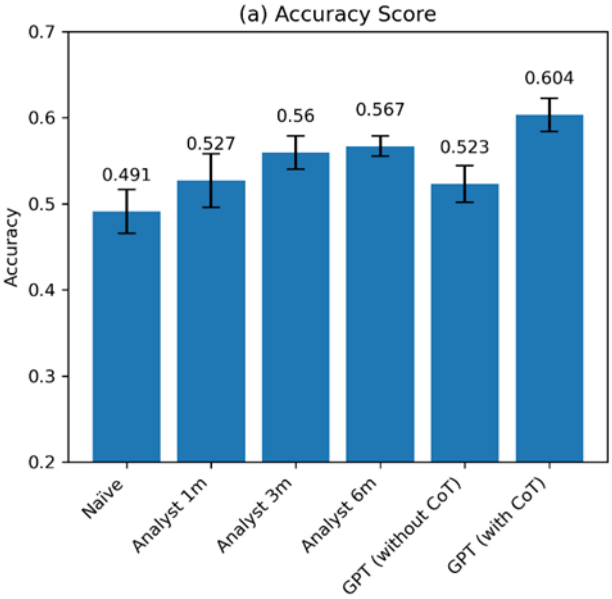
- Prompt Design: We design two types of prompts for GPT. First, a simple prompt that directly asks the model to predict future earnings direction. Second, a more sophisticated "Chain-of-Thought" (CoT) prompt that guides the model through a step-by-step process mimicking the reasoning of financial analysts. This CoT prompt instructs GPT to identify trends, calculate key financial ratios, synthesize information, and form expectations about future earnings changes.

By focusing on earnings predictions, we can benchmark GPT's performance against both human analysts and machine learning models explicitly trained for earnings prediction.

Comparative Performance: GPT-4 vs. Human Analysts

The results show that GPT, when using the CoT prompt, can outperform human analysts in predicting the direction of future earnings. Human analysts, who typically use a combination of quantitative data and qualitative insights, achieve a median accuracy of 52.7% when predicting earnings changes one month after the release of earnings (see Figure 1). As time passes and financial analysts obtain access to more information, this median accuracy increases to about 56.7% a quarter or two later. However, GPT, relying solely on numerical data of the initial set of financial statements and guided by the CoT prompt, achieves an accuracy of 60.4%.

Figure 1. GPT-4 vs. Analysts



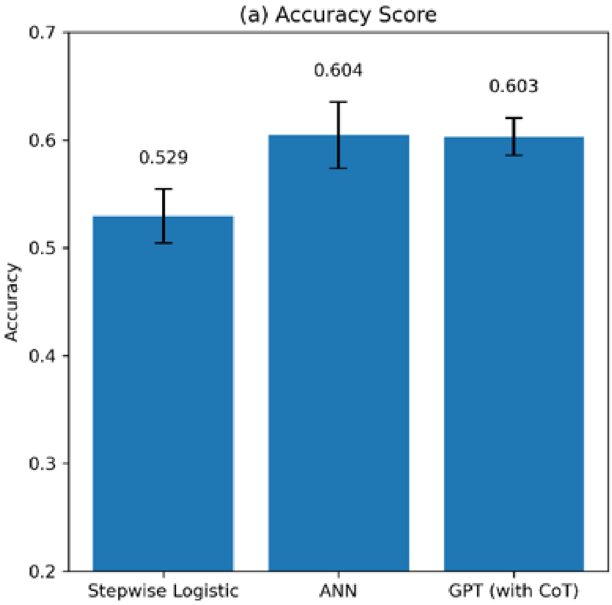
These findings suggest that LLMs like GPT can replicate, and even exceed, the performance of human analysts in fundamental financial analysis. Notably, GPT performs better in scenarios where human biases or inefficiencies might skew the analysis, demonstrating its potential to deliver more consistent and rational evaluations. However, we also show that there are complementarities between the two approaches.

Comparative Performance: GPT-4 vs. Machine Learning Models

We also compared GPT's performance with that of specialized machine learning models, such as a Stepwise Logistic Regression and an Artificial Neural Network (ANN), explicitly trained for earnings prediction. The Stepwise Logistic Regression, using 59 predictors, achieved an accuracy of 52.9%, consistent with human analysts' performance levels. The ANN, which uses the same financial data, demonstrated a higher accuracy of 60.4%—similar to GPT-4's performance (Figure 2a). Thus, a general-purpose LLM like GPT performs on par with state-of-the-art machine learning models, even though GPT was not trained to predict firms' earnings.

Further analyses suggest that there are complementarities between the two approaches. Training a new ANN on the text outputs obtained from GPT's predictions, together with financial data, achieves an even higher predictive performance (accuracy of about 63.2%). This finding suggests that while specialized models are optimized for specific tasks and data sets, GPT-4's ability to generalize from broad knowledge and reason through complex, unfamiliar situations sometimes gives it an edge.

Figure 2. GPT-4 vs. Machine Learning Models



Why Do LLMs Succeed in Financial Analysis?

We then explore when LLMs like GPT are particularly effective in assessing firms' future financial performance. We find that LLMs excel in tasks involving ambiguity and complexity—situations where traditional machine learning models, which are narrowly trained, might struggle. Indeed, trading strategies based on GPT's predictions achieve a high Sharpe ratio and substantial alpha, indicating potential economic value that surpasses traditional quantitative approaches.

We find that GPTs strength lies in its ability to generate narrative-like insights from raw numerical data, effectively synthesizing information in a manner akin to human deductive reasoning. This enables the model to make accurate predictions even without specific contextual knowledge. Finally, we also rule out alternative explanations, such as the model relying on its memory when forming predictions.

Implications for Financial Professionals and Policymakers

The findings from this study have significant implications for the future of financial analysis and decision-making processes:

1. **Complementing Human Expertise:** While GPT-4 and similar LLMs show great promise in financial analysis tasks, they should be seen as complementary tools rather than replacements for human analysts. Combining LLMs with the qualitative insights and soft knowledge of human analysts yields the most accurate and comprehensive outcomes.

2. **Potential for Enhanced Productivity:** Financial professionals could adopt LLMs to improve strategic thinking and decision-making further. For example, incorporating GPT-4's predictions into trading strategies could potentially improve decisions and investment outcomes.

Conclusion

Large Language Models such as GPT represent a transformative force in financial markets. We show that they demonstrate capabilities in financial analysis that rival both human analysts and specialized state-of-the-art machine-learning models. By complementing traditional analysis methods, LLMs can provide more informed, data-driven decision-making frameworks. As AI continues to evolve, the integration of LLMs into financial markets will require careful consideration to maximize their potential while addressing their limitations.

About the authors

Alex Kim is a Ph.D. candidate in accounting at the University of Chicago, Booth School of Business. He received a Master's degree in Business Administration with a concentration in Accounting and a dual Bachelor's degree in Economics and Business Administration, Summa Cum Laude, from Seoul National University. His research mainly examines the information processing of investors in the capital market. His research has been featured in major media outlets such as Financial Times, Bloomberg, Fortune, and Forbes, and funded by Ernest R. Wish Ph.D. Research Fellowship and Fama-Miller Center for Finance Research.

Maximilian Muhn is an Associate Professor of Accounting at University of Chicago, Booth School of Business. He is broadly interested in empirical accounting research. His current work focuses on the determinants and consequences of firms' financial transparency, as well as the effects of financial market and transparency regulation. Maximilian earned a PhD in accounting from Humboldt University of Berlin and he holds an MSc and a BSc in business administration from the University of Münster, Germany. During his studies, he gained work experience in consulting (McKinsey & Company and Boston Consulting Group), auditing (KPMG and Deloitte) and management accounting (BASF and ThyssenKrupp Steel).

Valeri Nikolaev is James H. Lorie Professor of Accounting and FMC Faculty Scholar at the University of Chicago, Booth School of Business. He studies the role of financial reporting and information in capital markets and contracting. His current research focuses on the transformative effects of emerging technologies, notably Generative AI and Large Language Models, on information processing and market efficiency. Nikolaev's primary interests revolve around the potential of AI tools to assist investors in making well-informed decisions, thereby fostering more efficient and equitable capital markets. His latest scholarly contributions shed light on unstructured narrative data and emphasize the significance of context when deciphering numerical information.

Find more SUERF Policy Notes and Briefs at www.suerf.org/publications/suerf-policy-notes-and-briefs/

SUERF
The European Money and Finance Forum

SUERF is a network association of central bankers and regulators, academics, and practitioners in the financial sector. The focus of the association is on the analysis, discussion and understanding of financial markets and institutions, the monetary economy, the conduct of regulation, supervision and monetary policy.

SUERF Policy Briefs (SPBs) serve to promote SUERF Members' economic views and research findings as well as economic policy-oriented analyses. They address topical issues and propose solutions to current economic and financial challenges. SPBs serve to increase the international visibility of SUERF Members' analyses and research.

The views expressed are those of the author(s) and not necessarily those of the institution(s) the author(s) is/are affiliated with.

SUERF Secretariat
c/o OeNB
Otto-Wagner-Platz 3
A-1090 Vienna, Austria
Phone: +43-1-40420-7206
www.suerf.org • suerf@oenb.at