Financial Statement Analysis with Large Language Models

Alex G. Kim, Maximilian Muhn, and Valeri V. Nikolaev

(WP,2024)

石宛青

（武汉大学金融系）

2025 年 04 月 18 日

Motivation

- **LLMs Current Strengths: Strong in Text Tasks**
  - Summarization, report generation, sentiment analysis (Bernard et al., 2023)
  - All textual domain and require specialized training or fine-tuning of the model
- **Outstanding Limitation: Numerical analysis and judgment**
  - computation, numerical understanding, human-like judgment(Brown et al., 2020)
  - LLMs lack targeted training; capability remains unclear
- **Research Focus: Financial Statement Analysis (FSA)**
  - Quantitative task centered on numerical data
  - requires reasoning and judgment
  - Core task for financial analysts

Question

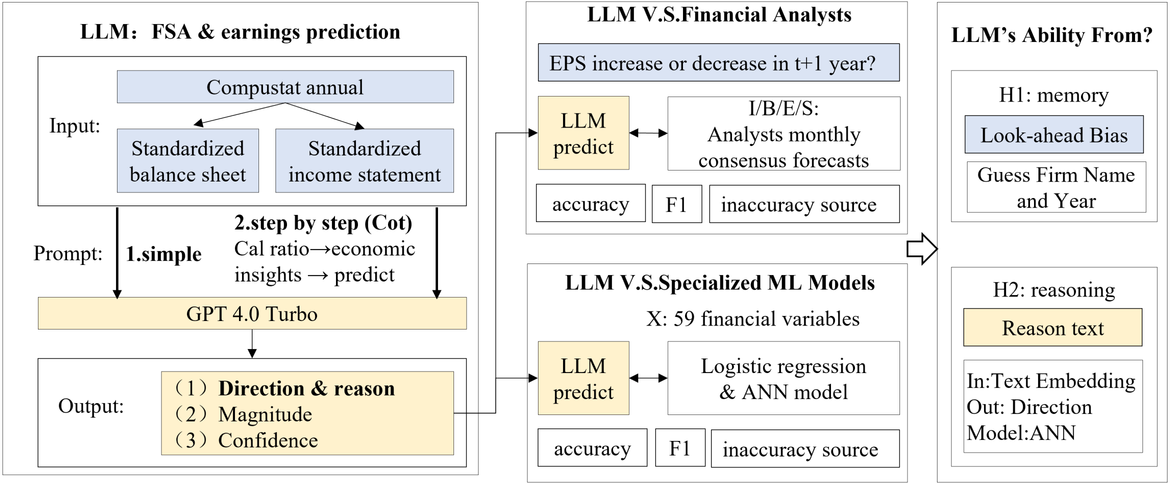
- Can LLM perform FSA earnings prediction from numbers alone, like professional human experts ?
  - better than human: F1-score:54.48% V.S.60.90%
  - better than ML model: F1-score:61.62% V.S.63.45%

- Where LLM' s prediction ability from?
  - H1: Memorization —LLM "remember" historical company patterns
  - H2: Reasoning —process numeric inputs and generate economically valuable insights——Yes

Contribution

- contributes to literature on FSA
  - prior: human analysts
  - expand: first to provide large-scale evidence on LLM with purely numbers
- contributes to complementarities between humans and machines in finance
  - prior: specialized ML models in lending,stock analyses(Cao et al.,2024)
  - expand: LLMs & human experts
- contributes to literature on earnings prediction based on fundamental analysis
  - prior: trained on 12,000+ XBRL-based variables (Chen et al., 2022)
  - expand: novel approach: derive fundamental insights about future performance from FSA and predict
- contributes to limits of LLMs-outside native domain:quantitative analysis task

Design



LLM: FSA & earnings prediction
Input:
Compustat annual
Standardized balance sheet
Standardized income statement
Prompt:
1.simple
2.step by step (Cot)
Cal ratio→economic insights → predict
GPT 4.0 Turbo
Output:
（1） Direction & reason
（2） Magnitude
（3） Confidence
LLM V.S.Financial Analysts
EPS increase or decrease in t+1 year?
LLM predict
I/B/E/S: Analysts monthly consensus forecasts
accuracy
F1
inaccuracy source
LLM V.S.Specialized ML Models
X: 59 financial variables
LLM predict
Logistic regression & ANN model
accuracy
F1
inaccuracy source
LLM's Ability From?
H1: memory
Look-ahead Bias
Guess Firm Name and Year
H2: reasoning
Reason text
In:Text Embedding
Out: Direction
Model:ANN

Design-Data base

- 1986-2021, Compustat annual, 150678 observations from 15401 distinct firms
- 2 years of balance sheet and 3 years of income statement data
- unnamed， t,t-1

Panel A. Balance Sheet

Account Items	t	t-1
Cash and Short-Term Investments	11.138	17.323
Receivables	157.535	140.057
Inventories	349.811	326.411
Other current assets	27.74	12.3
Current Assets	546.224	496.091
Property, Plant, and Equitpment (Net)	90.754	89.103
Investment and Advances (equity)	32.469	31.184
Other investments	0.0	0.0
Intangible assets	115.732	123.674
Other assets	57.953	47.515
Total Asset	843.132	787.567
Debt in current liabilities	49.066	61.699
Account payable	94.357	77.99
Income taxes payable	0.0	0.0
Other current liabilities	169.163	146.208
Current liabilities	312.586	285.897
Long-term debt	0.153	0.079
Deferred taxes and investment tax credit	0.0	0.0
Other liabilities	63.192	47.937
Total Liabilities	375.931	333.913
Preferred stock	0.0	0.0
Common stock	467.201	453.654
Stockholders' equity total	467.201	453.654
Noncontrolling interest	0.0	0.0
Shareholders' Equity	467.201	453.654
Total Liabilities and Shareholders' Equity	843.132	787.567

Panel B. Income Statement

Account Items	t	t-1	t-2
Sales (net)	2030.154	1733.703	3978.711
Cost of Goods Sold	1165.555	1013.953	1153.618
Gross Profit	864.599	719.75	2825.093
Selling, General and Administrative Expenses	518.671	481.884	1852.951
Operating Income Before Depreciation	345.928	237.866	972.142
Depreciation and Amortization	110.985	100.493	160.207
Operating Income After Depreciation	234.943	137.373	811.935
Interest and related expense	21.647	27.91	10.985
Nonoperating income (excluding interest income)	22.062	1.655	-8.833
Interest income	77.543	11.887	22.783
Special items	0.0	0.0	-4.744
Pretax income	312.901	123.005	810.156
Income taxes (current)	0.0	0.0	0.0
Income taxes (deferred)	6.874	8.428	-18.459
Income taxes (other)	0.0	0.0	0.0
Income before extraordinary items and noncontrolling interest	0.0	0.0	0.0
Noncontrolling interest	0.638	0.471	0.354
Income before extraordinary items	201.412	74.438	518.834
Dividends	0.0	0.0	0.0
Income before extraordinary items for common stocks	201.412	74.438	518.834
Common Stock Equivalents - Dollar Savings	0.0	0.0	0.0
Income Before Extraordinary Items - Adjusted for Common Stock Equivalents	201.412	74.438	518.834
Extraordinary Items and Discontinued Operations	-12.366	5035.621	0.0
Net Income (Loss)	189.046	5110.059	518.834
Earnings per Share - Basic Excluding Extraordinary Items	1.47	0.54	3.82
Earnings per Share - Diluted Excluding Extraordinary Items	1.47	0.54	3.82

Design-Prompt

- **Simple**: analyze the two financial statements of a company and determine the direction of future earnings.
- **Chain-of-Thought:**take on the role of a financial analyst to perform FSA
  1. identify and describe notable changes in certain financial statement items.
  2. compute key financial ratios
  3. provide economic interpretations of the computed ratios
  4. synthesize information and predict whether earnings are likely to increase or decrease in the subsequent period

Design-LLM V.S. human or ML

- Analysts' Forecasts: forecasts of year t + 1 EPS
  - For each analyst, we use the forecast closest to the year t earnings release.
  - t+1 analyst forecasts(>=3 people) median values>real
- Specialized ML Models:Logic & ANN
  - X: 59 financial variables(Ou & Penman (1989)) exclude the price-to-earning ratio
  - X: same balance sheet and income statement variables
- Sources of Inaccuracy

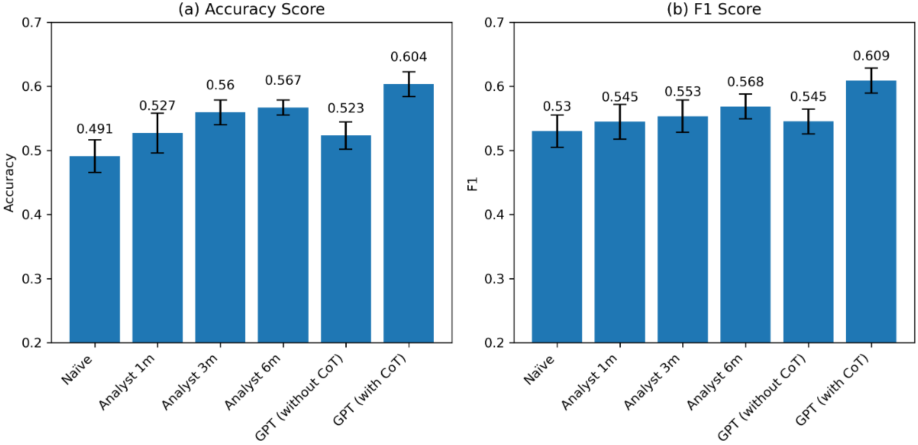
$$I(\text{Incorrect} = 1)_{it} = \beta X_{it} + \delta_{\text{year}} + \delta_{\text{ind}} + \epsilon_{it}$$

  - X_{it}: asset size, leverage, book-to-market ratio, earnings volatility, loss indicator, and property, plant, and equipment scaled by total assets.

Design-Where LLM’ s prediction ability from?

- H1: Memorization —LLM "remember" historical company patterns
  - Can GPT Guess Firm Name and Year?
    - output Top 10 most probable firm; the most probable fiscal year
  - Analysis Outside of GPT’ s Training Window
    - 2022 data pridict 2023
- H2: Reasoning —process numeric and generate economically valuable insights
  - Information Content of Generated Text
    - Input:GPT text-BERT embedding-768 dimensional vector
    - Output:EPS direction
    - model: ANN
    - V.S. Input: variables from the two financial statements

Result-GPT vs. Human Analysts



- better than human(CoT).

Result-GPT vs. Human Analysts

Panel A. Determinants

Dep Var	I(Incorrect=1)			
	GPT (1)	Analyst 1m (2)	Analyst 3m (3)	Analyst 6m (4)
Size	-0.017***	-0.008***	-0.010***	-0.010***
	(-5.16)	(-5.72)	(-4.69)	(-4.81)
BtoM	-0.022	-0.016***	-0.012**	-0.012**
	(-0.99)	(-2.94)	(-2.21)	(-2.35)
Leverage	-0.145	-0.032	-0.029	-0.029
	(-1.50)	(-0.37)	(-1.40)	(-1.36)
Loss	0.193***	0.141***	0.146***	0.145***
	(4.76)	(7.02)	(6.90)	(6.09)
Earnings Volatility	0.236***	0.169***	0.160***	0.132**
	(2.69)	(4.08)	(3.46)	(2.47)
PP&E	0.133*	0.041	0.036*	0.031
	(1.67)	(1.18)	(1.71)	(1.25)
Year FE	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes
Adjusted R2	0.08	0.027	0.032	0.029
N	37,736	37,736	37,736	37,736

- analysts tend to be relatively better at dealing with these complex financial circumstances than GPT

Result-GPT vs. Human Analysts

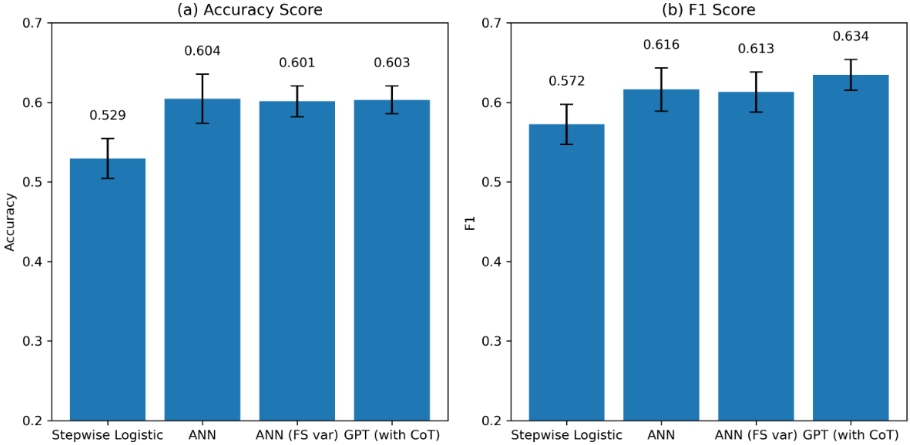
$$I(\text{Increase} = 1)_{it} = \beta_1 \text{Pred\_GPT}_{it} + \beta_2 \text{Pred\_Analyst}_{it} + \delta_{\text{year}} + \delta_{\text{ind}} + \epsilon_{it}$$

Panel B. Incremental Informativeness

Dep Var	I(Increase=1)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GPT	0.182***				0.170***	0.151**	0.152**
	(2.99)				(2.67)	(2.35)	(2.30)
Analyst 1m		0.073***			0.110**		
		(3.11)			(2.43)		
Analyst 3m			0.098***			0.122***	
			(4.02)			(3.49)	
Analyst 6m				0.100***			0.124***
				(4.05)			(3.62)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R2	0.07	0.025	0.043	0.044	0.089	0.091	0.091
N	37,736	37,736	37,736	37,736	37,736	37,736	37,736

- Analysts are not useless.

Result-GPT vs. Machine Learning Models



- better than ML(F1-score)

Result-GPT vs. Machine Learning Models

Panel B. Sources of Inaccuracy

Dep Var =	I(Incorrect=1)		
	GPT (1)	ANN (2)	Stepwise Logistic (3)
Size	-0.015***	-0.024***	-0.029***
	(-9.09)	(-11.33)	(-11.56)
BtoM	0.001	0.002	0.002
	(0.38)	(0.73)	(0.69)
Leverage	0.092***	0.085***	0.090***
	(6.30)	(5.88)	(6.02)
Loss	0.134***	0.181***	0.202***
	(9.64)	(11.35)	(12.96)
Earnings Volatility	0.040**	0.062***	0.078***
	(2.09)	(6.35)	(8.02)
PP&E	0.027*	0.016	0.02
	(1.95)	(1.53)	(1.69)
Year FE	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Estimation	OLS	OLS	OLS
Adjusted R2	0.097	0.102	0.109
N	133,830	133,830	133,830

- ML similar to LLM,can be even more sensitive

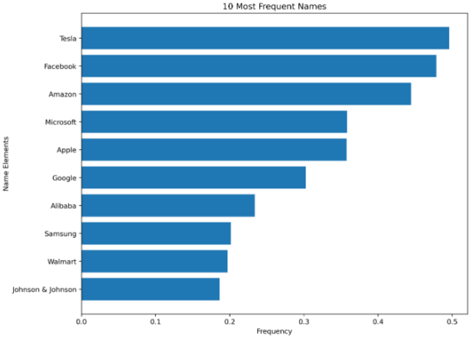
Result-GPT vs. Machine Learning Models

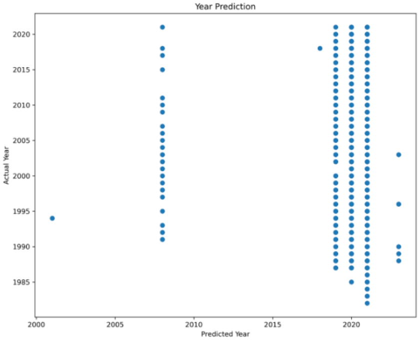
Panel C. Incremental Informativeness

Dep Var	I(Increase=1)				
	(1)	(2)	(3)	(4)	(5)
GPT	0.181***			0.170***	0.179***
	(3.43)			(2.67)	(3.35)
ANN		0.150***		0.053**	
		(3.69)		(2.44)	
Logistic			0.088***		0.068**
			(2.99)		(2.05)
Year FE	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Adjusted R2	0.056	0.051	0.032	0.061	0.06
N	133,830	133,830	133,830	133,830	133,830

- ML are not useless.

Result-GPT’ s Memory





- 2022 predict 2023 robust;H1 is not supported

Result-Reasoning-Predictive Ability of GPT-Generated Texts

	Accuracy	F1 Score	AUC
1. Baseline			
ANN with Financial Statement Variables	60.12%	61.30%	59.13%
2. Embeddings of the Generated Text			
ANN with GPT Text Embedding	58.95%	65.26%	64.22%
ANN with Adjusted Text Embedding			
ANN excl. Trend	57.11%	64.03%	63.81%
ANN excl. Ratio	55.65%	62.36%	61.89%
ANN excl. Rationale	58.88%	65.15%	64.16%
3. Text and FS Variables Together			
ANN with Embedding *and* FS Variables	63.16%	66.33%	65.90%
ANN with Adjusted Text Embedding and FS Variables			
ANN excl. Trend	62.51%	65.58%	65.50%
ANN excl. Ratio	61.77%	64.30%	63.16%
ANN excl. Rationale	62.95%	65.96%	65.59%

- highlights the value of narrative insights generated by an LLM from purely numerical numerical information.

Idea

- 替换 Y:
  - 如 LLM 是否可以识别出异常的财务报表，从而预警欺诈行为?
  - 识别信用风险，预测债券价格?
- 替换 X:
  - 其他非文本领域：K 线图，step by step 分析 K 线，预测股价走向
  - 其他数值：FSA 与宏观经济数据相结合（加入 t,t-1 年的宏观数据），提高复杂情况的预测效果
- 人与机器：LLM 在复杂情况表现不佳，设计 human+ML, 复杂时段增加 human 比重

Thanks!