

Beyond the Reported Cutoff: Where Large Language Models Fall Short on Financial Knowledge

Agam Shah, Liqin Ye, Sebastian Jaskowski, Wei Xu, Sudheer Chava



Georgia Institute of Technology



Georgia Tech

Financial Services
Innovation Lab



Georgia
Tech.

Dr. Sudheer Chava | Director

Dr. Chava is the Alton M. Costley Chair, a Professor of Finance at Scheller College of Business at Georgia Institute of Technology, Atlanta and leads the Financial Services Innovation Lab.

Dr. Chava also serves as:

- Finance Area Coordinator at Scheller College of Business
- Director of the Masters in Quantitative and Computational Finance (MS-QCF) program at Georgia Tech
- Member of the Advisory Committee of Georgia Tech's Partnership for an Advisory Computing Environment (PACE)
- Member of a Special Committee on Payment Inclusion at the Federal Reserve Bank of Atlanta



- **25+ Industry practitioner, research, and faculty affiliates**
- **12+ Postdoc, PhD (Finance and ML)**
- **15+ MS Students (QCF, CS, ECE, etc.)**
- **100+ Undergrad Students (CS, ISyE, Math, ICT, Business, etc.)**

Major research areas at the Financial Services Innovation Lab can be categorized into three broad categories: Responsible and Sustainable Finance, AI in Finance (ML and NLP) and FinTech Innovation. Within these categories, lab members conduct research on the following topics

 Household Income, Investment, Consumption, Retirement & Debt Choices

 Banking

 Credit Risk

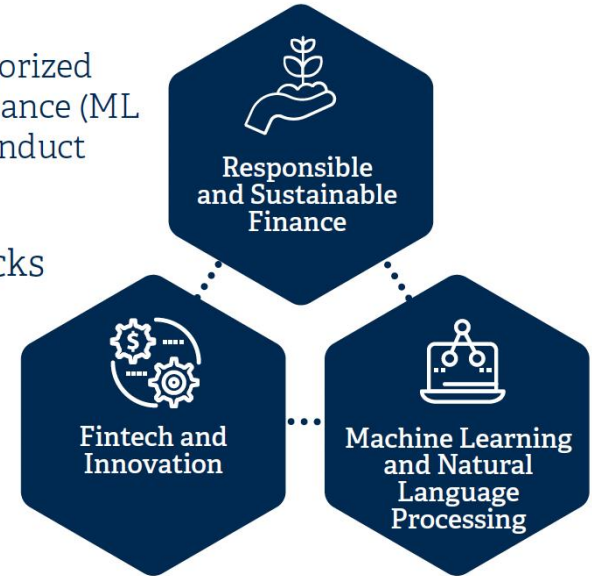
 Climate Finance and ESG

 Impact of economics shocks on households

 ML and NLP in Finance

 Financial Inclusion and Financial Literacy

 Crypto, Blockchain and Decentralised Finance



How does the lab work?

Research and Insights

Conduct cutting-edge research that generates policy relevant insights for financial markets, institutions, and consumers

Projects

Design and facilitate hands-on Experiential Learning Projects to enhance student research experience

Events

Organize events to promote dialogue between Academics, Industry Practitioners, and Regulators

Recruiting

Recruit undergraduate and graduate students as a partner

FinTech Entrepreneurship

Enable Entrepreneurship in Finance and FinTech

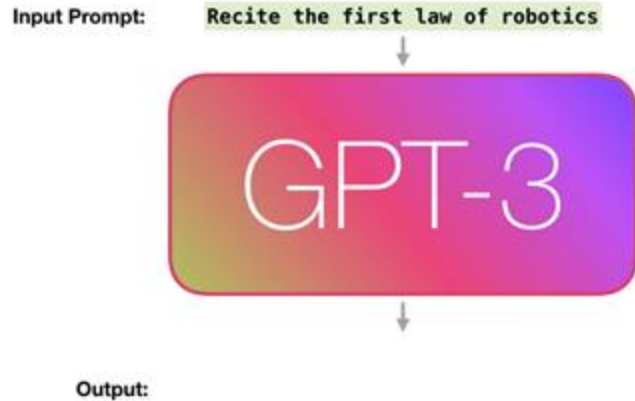
- **Usage Increase:** Non-work related chats growing from ~53% (mid-2024) to ~**70%** (mid-2025) of all consumer usage.
- 3 Major Topics (80% of all):
 - **Seeking Information**
 - **Practical Guidance**
 - **Writing**
- For “Asking/Doing/Expressing” intent:
 - ~**49%** are **Asking** (i.e. **seeking information** or clarification)
 - ~40% are Doing (i.e. asking model to produce)

LLM in Financial Domain

- ChatGPT based portfolio selection [Romanko et al., 2023]
- “*Democratizing*” financial knowledge [Yue et al., 2023]
- Trading volume decline significantly during ChatGPT outages [Chen et al., 2024]
- Offering financial advice for one-time investment (better than robo-advisors) [Oehler and Horn, 2024]
- Constructing an interdependency graph of 30 large market cap (Dow 30) companies [Chen et al., 2023]

But **hallucinations in LLM are statistically inevitable** [OpenAI, Sep 2025]

How reliable LLMs are in financial knowledge?



GIF Source: Jay Alammar

How LLMs Work

- Context comes from two sources:
 - **Retriever** (e.g., web search agent, file search agent)
 - **User/system prompt**
- **Parameters process** the context
- Output is generated **token by token**

Key Challenge

- **Parametric knowledge bias:** even with rich context, hallucinations persist
- **Opaque pre-training:** most frontier LLMs don't share pre-training corpora

Need: Evaluate parametric knowledge bias in LLMs

Motivation Example

● No Answer

● Error ≥ 10%: Hallucination

● Error < 10%: Correct

Temporal Bias

👤: What was the revenue of NEXTERA ENERGY INC in 1984?

🌀: NextEra Energy Inc., formerly known as FPL Group, Inc., was founded in 1925. However, I couldn't find the revenue data for 1984 specifically. (Actual revenue is \$3940.9M)

👤: What was the revenue of NEXTERA ENERGY INC in 2006?

🌀: According to the Securities and Exchange Commission (SEC) filings, NextEra Energy, Inc. (formerly known as FPL Group, Inc.) reported revenue of \$12.4 billion in 2006. (Actual revenue is \$15710M)

👤: What was the revenue of NEXTERA ENERGY INC in 2013?

🌀: According to the NextEra Energy, Inc. 2013 Annual Report (Form 10-K), the company's revenue for 2013 was \$15.14 billion. (Actual revenue is \$15136M)

Company Size Bias

👤: What was the revenue of ROCKY MOUNTAIN CHOC FACT INC in 2018?

🌀: I'm happy to help! However, I need to clarify that there is no publicly traded company (Market Cap. ≈ 50M | Actual revenue is 34.545M)

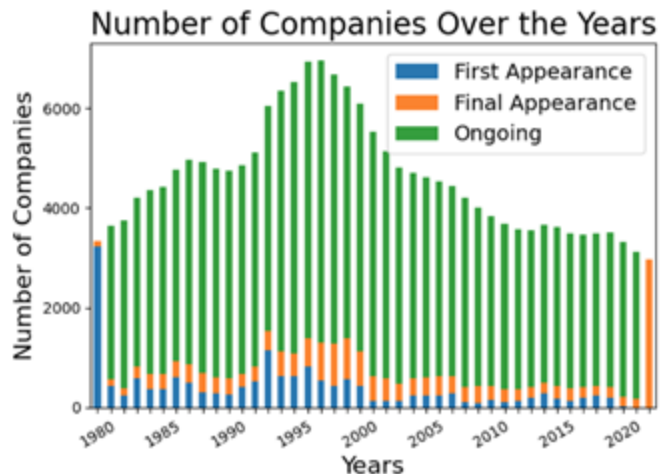
👤: What was the revenue of SPIRIT AIRLINES INC in 2018?

🌀: According to the Spirit Airlines Inc. annual report for 2018 (Form 10-K), the company's total operating revenue for 2018 was \$2.98 billion. (Market Cap. ≈ 3.95B | Actual revenue is \$3.323B)

👤: What was the revenue of MICROSOFT INC in 2018?

🌀: According to Microsoft's annual report for 2018 (fiscal year ended June 30, 2018), the company's revenue was \$110.36 billion. (Market Cap. ≈ 780B | Actual revenue is \$110.36B)

- Size
 - Full Sample: 190k observations
 - Representative Small Sample: 8,200 observations (4 brackets*50 samples*41 years)
- 17,621 unique companies
- Cross-sectional Variables
 - Market Capitalization
 - Robinhood Retail Attention
 - Bloomberg Abnormal Institutional Attention
 - SEC Access
 - Bog Index for Readability



Data (Source)	Key Variables	Years	#Data Points	#Unique Companies
Annual Financials (Compustat Capital-IQ)	Revenue	1980-2022	197,011	17,621
Market Capitalization (CRSP MSF)	Price and # Shares	1980-2022	197,011	17,621
Robinhood (Robintrack)	# Holders	2018-200	9,592	3,469
Bloomberg AIA (Bloomberg)	B-AIA	2010-2020	21,303	2,432
SEC Access (SEC-EDGAR)	# Access	2003-2017	58,519	7,581
Bog Index (Bonsall IV et al., 2017)	Bog Index	1994-2020	110,161	12,822

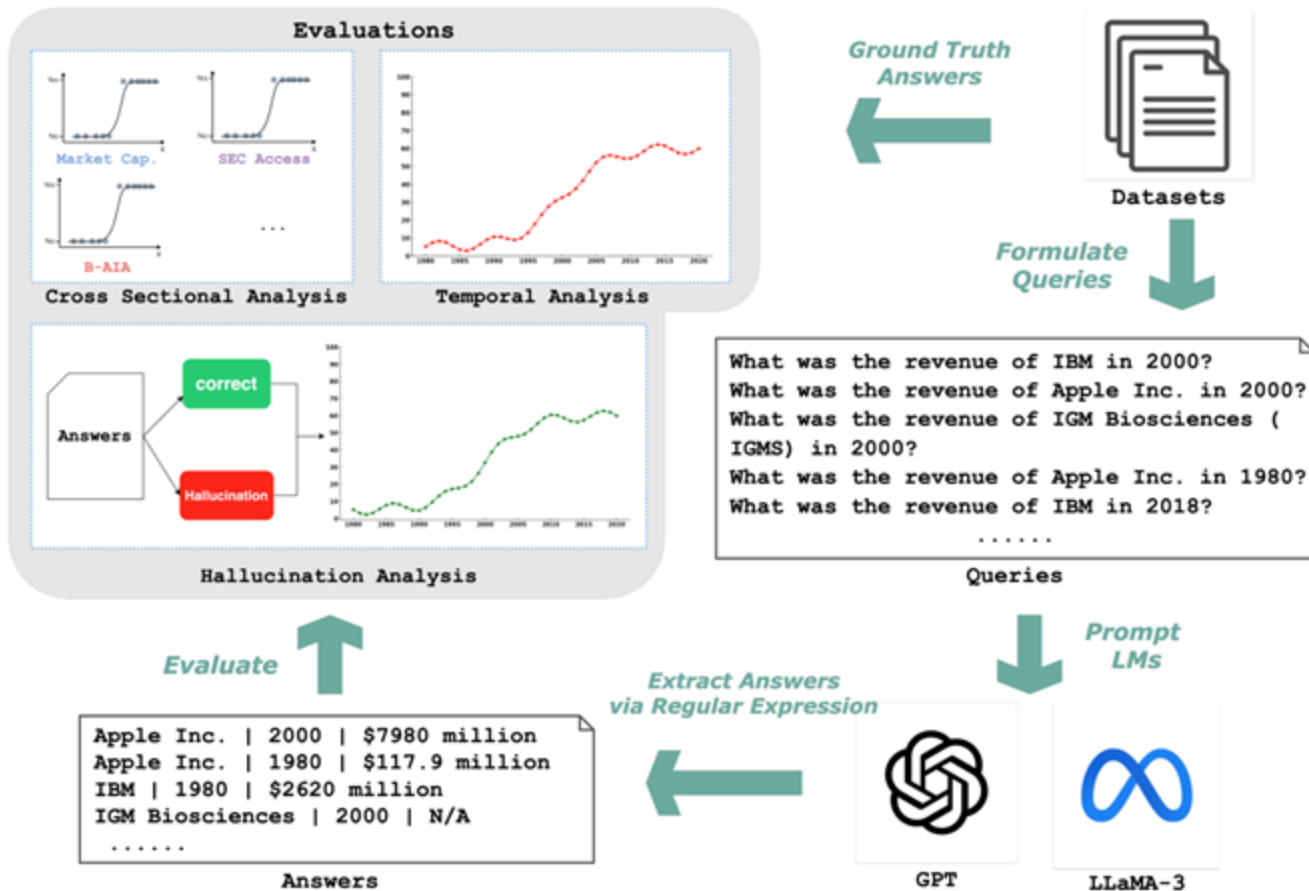
Model

- Full Sample
 - Open Weight: LLaMA-3-8B-Chat & LLaMA-3-70B-Chat
 - Closed: GPT-4o-mini & GPT-4o
- Representative Sample
 - GPT-4.5
 - DeepSeek-V3
 - Gemini-1.5-pro
 - FinMA-7B-full
- Temperature: 0.00 | Max_token: 100

Prompt

- What was the **revenue** of {company_name} in financial year {financial_year}?

Evaluation Framework



Ternary outcome variable $Y_{i,t}$ for firm i and year t based on the model's answer

$$Y_{i,t} = \begin{cases} 2 : & \text{absolute \% error} < 10\% \\ 1 : & \text{absolute \% error} \geq 10\% \\ 0 : & \text{no numerical answer} \end{cases}$$

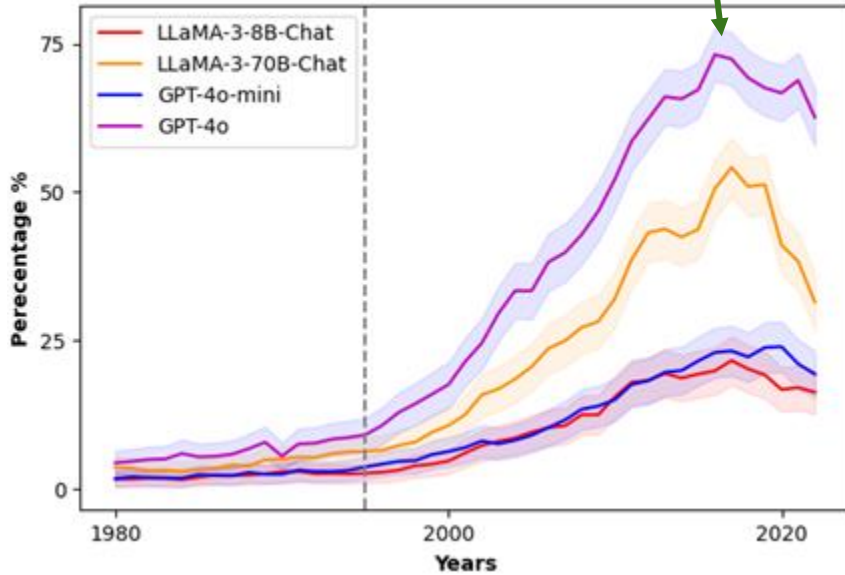
$$\text{Success Rate (T)} = \frac{\sum_{i,t=T} \mathbb{1}\{Y_{i,t}=2\}}{\sum_{i,t=T} 1}$$

$$\text{Hallucination Rate (T)} = \frac{\sum_{i,t=T} \mathbb{1}\{Y_{i,t}=1\}}{\sum_{i,t=T} \mathbb{1}\{Y_{i,t} \neq 2\}}$$

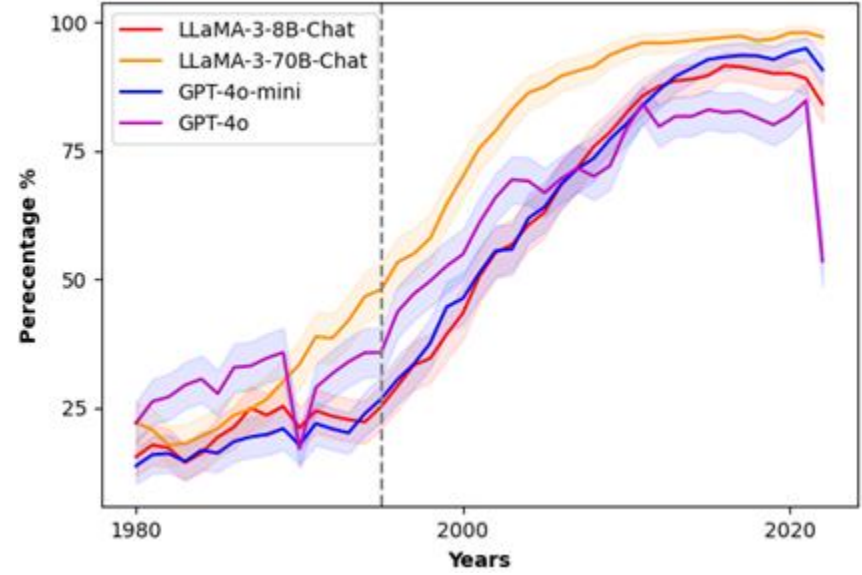
Temporal Knowledge Gap

- Stronger performance on questions from more recent years
- Hallucinations more frequent in years where accuracy is also higher
- True knowledge cutoff, echoing with “Dated Data: Tracing Knowledge Cutoffs in Large Language Models” (COLM Best Paper) [Cheng, Jeffrey et al., 2024]

Success Rate For Different Models

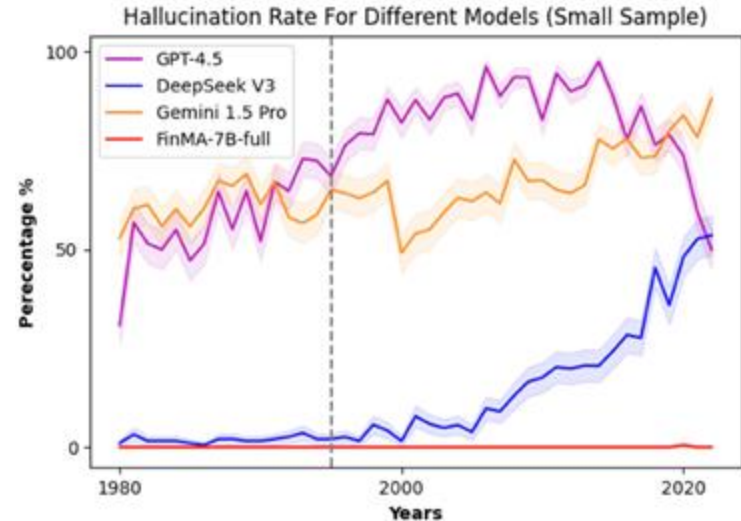
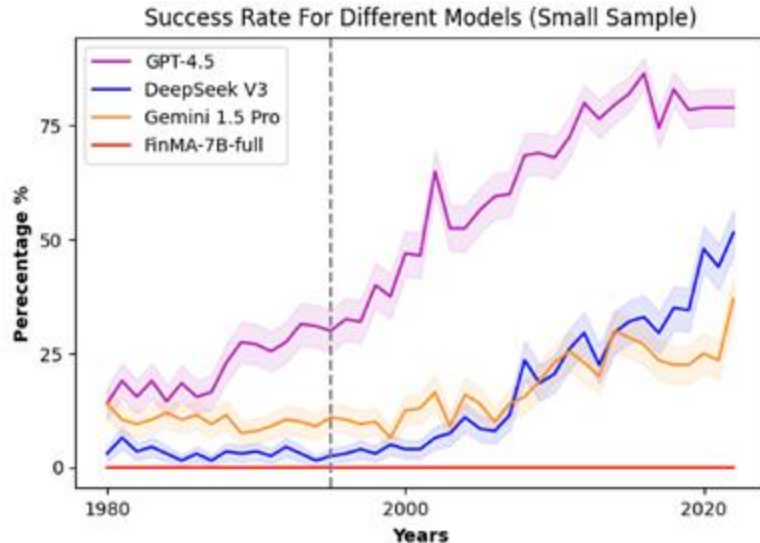


Hallucination Rate For Different Models



Temporal Knowledge Gap

- Similar trend as full sample
- GPT-4.5 has similar success rate as GPT-4o but better on hallucination
- FinMA output either no text or a number without unit



- Logistic Regression

$$\text{logit}(P(Y_{i,t} = y)) = \alpha + \beta * X_{i,t} + \delta_t * D_t + \epsilon_{i,t}$$

- Independent variables
 - Constant
 - $X_{i,t}$: CPI adjusted (Dec 2021 value) market cap or other firm-level characteristic
 - D_t : Year fixed effect (FE) dummies, D_{2011} take value 1 for observations in the year 2011 and 0 otherwise

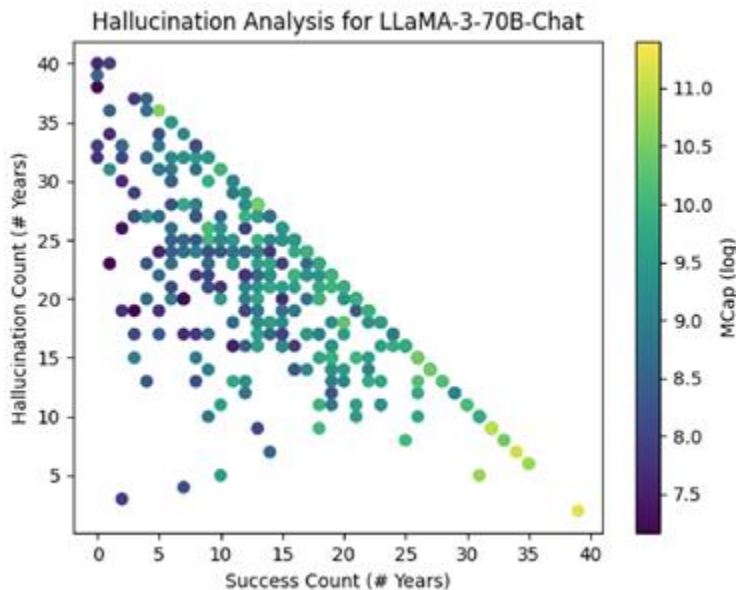
Cross-Sectional Knowledge Gap

- Enhanced performance for company with **larger MCap, higher retail attention, more SEC-Access, and higher B-AIA.**
- Lower performance for company with less readable (higher bog index) SEC filings
- **Interpretation:** For GPT-4o, 10× increase in company market cap → +0.9209 rise in log-odds of correctly answering revenue questions

$X_{i,t}$	Llama-3-8B		Llama-3-70B		GPT-4o-mini		GPT-4o	
	α	β	α	β	α	β	α	β
MCap (log)	-13.0984‡	1.0546‡	-11.8915‡	1.0091‡	-15.1833‡	1.2879‡	-10.9046‡	0.9209‡
retail (std)	-1.2811‡	0.2073‡	0.1755‡	0.1036‡	-1.1033‡	0.5727‡	0.9562‡	0.2589‡
B-AIA (std)	-1.1272‡	0.0187	-0.1870‡	0.0107†	-1.0425‡	0.0217†	0.6939‡	0.0180†
SEC-Access (std)	-2.4464‡	0.0628‡	-1.5922‡	0.0782‡	-2.4840‡	0.1028‡	-0.8490‡	0.0770‡
Bog Index (std)	-2.5228‡	-0.0965‡	-1.8361‡	-0.0666‡	-2.4133‡	-0.0296‡	-1.5021‡	-0.0753‡

Hallucination

- More likely to hallucinate with larger MCap company
- Larger MCap Company: More likely to be both correct and hallucinatory
 - Consistent with *Why Hallucination Happen?* [OpenAI, Sep 2025]: modern training rewards answering even when uncertain, raising both correct answers and hallucinations



Model	Constant (α)	Beta (β)
Llama-3-8B	-6.7754‡	0.6053‡
Llama-3-70B	-2.8859‡	0.1914‡
GPT-4o-mini	-6.2188‡	0.5222‡
GPT-4o	-2.1171‡	0.0964‡

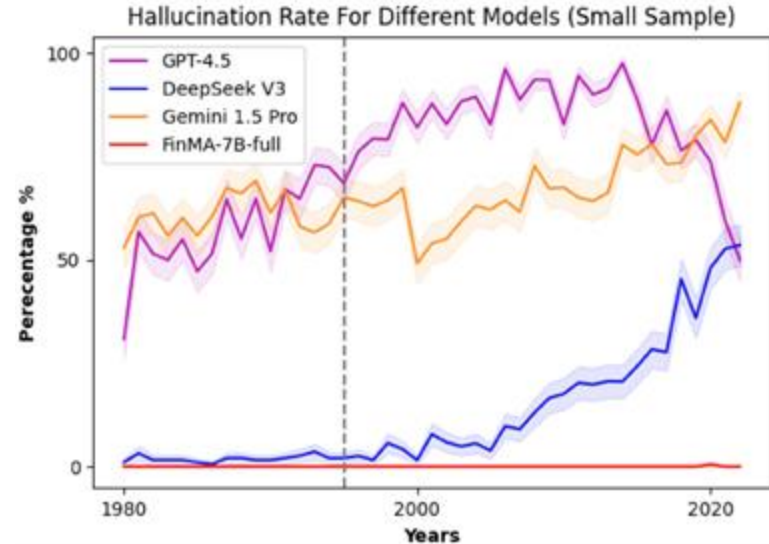
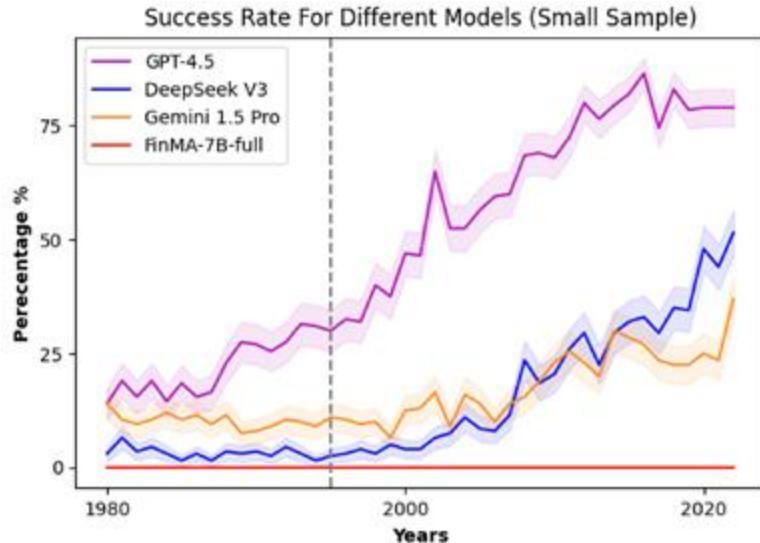
Interpretation: For GPT-4o, 10× increase in company market cap → +0.0964 rise in log-odds of correctly answering revenue questions

Robustness Check: Same Company Over Time

- 460 companies existed consistently over 43 years
- Both temporal and cross-sectional knowledge gap exist

Model	Constant (α)	Beta (β)
Llama-3-8B	-10.4140‡	0.8408‡
Llama-3-70B	-9.0624‡	0.7786‡
GPT-4o-mini	-11.0263‡	0.9567‡
GPT-4o	-8.4113‡	0.7051‡

MCap



Method

- **Chain-of-Thought, multi-turn prompt:**
 - a. show revenues for 2018-2023
 - b. ask GPT-4o to **forecast next-year revenue**
 - c. ask for a **BUY/SELL/DNK** recommendation one year further out.
- Only query after **Sept 30, 2023**, the knowledge cutoff of GPT.

Findings

- Decision aversion for small caps: **lower MCap firms are more likely to get DNK**
- **Large-cap bias toward BUY**

Model	Constant (α)	Beta (β)
No Recommendation (DNK)	1.4804‡	-0.0654‡
BUY Recommendation	-0.5766‡	0.0751‡
SELL Recommendation	0.1006‡	-0.0103‡

Key Takeaway

- ❑ LLMs are better informed about **recent year & large / popular firms**
- ❑ More likely to hallucinate for the same
- ❑ High accuracy does not imply low risk
- ❑ Knowledge gap could lead to biased financial advice

FOMC and NLP

- **Paper: Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis (ACL 2023 main)**
 - **Why?**
 - FOMC announcements moves market
 - Chair Powell's 8-minutes speech at Jackson Hole wiped almost \$6.5T just from equity market
 - **Largest annotated dataset**
 - Meeting minutes
 - Press conferences
 - Speeches
 - **State-of-the-art performance with novel sentence-splitting method**
 - **Model successfully predicts different financial markets**
 - Stock market
 - Treasury market
 - CPI and PPI

Deciphering Central Bank Policy Implications

- Paper: Words That Unite The World: A Unified Framework for Deciphering Central Bank Communications Globally (NeurIPS 2025)

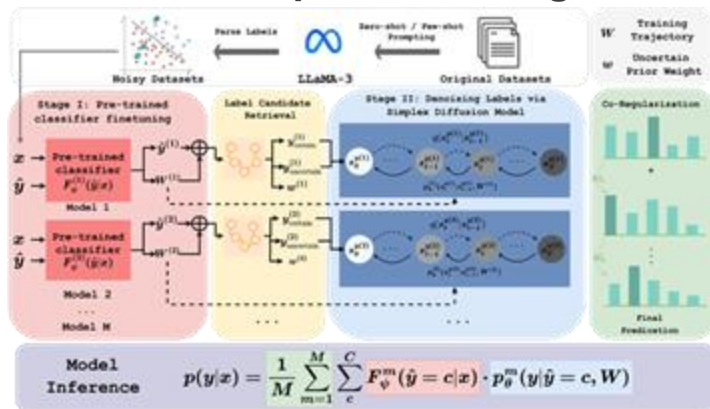


Dataset	Models	Annotations
Central Banks	25	104
Years	1996–2024	26
Scraped Sentences	380,200	6
Annotated Sentences	25,000	
Total Words	10,289,163	
Corpus Size	2,661,400	
Sentences/Year*	13,110.34	
Words/Sentence*	27.06	
	Pretrained Language Models	7
	Large Language Models	9
	Best Stance Model*	RoBERTa-Large (0.740)
	Best Temporal Model*	RoBERTa-Base (0.868)
	Best Uncertainty Model*	RoBERTa-Large (0.846)
	Benchmarking Experiments	15,075
	Few Shot	
	Few Shot + Ann. Guide	✓
	Annotation Guides	26
	Annotation Steps	6
	Tasks	
	Stance Detection	Hawkish, Dovish, Neutral, Irrelevant
	Temporal Classification	(Not) Forward Looking
	Uncertainty Estimation	(Un)certain

Bank	Base			Large			Closed-Source			Open-Source				
	MBL	FB	EBL	MBL	FB	EBL	Gen	4.1m	4.1	25	Qwen	FM	L3	L45
BoC	0.78	0.85	0.90	0.85	0.75	0.80	0.52	0.88	0.82	0.94	0.83	0.25	0.90	0.93
BoE	0.80	0.87	0.90	0.91	0.84	0.84	0.67	0.87	0.82	0.91	0.92	0.28	0.94	0.93
BoJ	0.78	0.81	0.84	0.79	0.78	0.80	0.52	0.85	0.78	0.85	0.86	0.30	0.91	0.93
BoK	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoL	0.78	0.82	0.85	0.85	0.78	0.78	0.52	0.85	0.78	0.85	0.86	0.30	0.91	0.93
BoM	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoN	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoP	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoR	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoS	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoT	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoU	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoV	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoW	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoX	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoY	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoZ	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAA	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAB	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAC	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAD	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAE	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAF	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAG	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAH	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAI	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAJ	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAK	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAL	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAM	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAN	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAO	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAP	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAQ	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAR	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAS	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAT	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAU	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAV	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAW	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAX	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAY	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoAZ	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBA	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBB	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBC	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBD	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBE	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBF	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBG	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBH	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBI	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBJ	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBK	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBL	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBM	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBN	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBO	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBP	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBQ	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBR	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBS	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBT	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBU	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBV	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBW	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBX	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBY	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoBZ	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoCA	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoCB	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88	0.85	0.90	0.90	0.28	0.91	0.94
BoCC	0.82	0.88	0.90	0.89	0.82	0.82	0.64	0.88						

Learning from LLM Noisy Labels

- Paper: Calibrating Pre-trained Language Classifiers on LLM-generated Noisy Labels via Iterative Refinement (KDD 2025)
 - Motivation: (1) LLM as annotator is **faster**, **cheaper**, and **better** than traditional labeling methods, but **NOISY**. (2) LLM-generated label noise is different from well-studied synthetic and real-world noise.
 - Method:
 - True label candidate retrieval + Label distillation
 - Simplex denoising label diffusion model



Datasets (→)	NumClaim		TREC		SemEval		20News
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot
Llama-3-70b	89.94	95.53	81.80	84.00	47.50	48.50	74.04
PLC	90.54±0.72	95.11±0.30	80.64±0.94	77.72±1.34	51.59±0.44	50.46±0.72	71.2±0.52
Co-teaching	92.25±1.17	94.45±0.53	80.24±3.51	80.42±2.55	52.45±1.43	50.90±1.36	70.95±0.64
JoCoR	92.14±0.61	93.45±0.80	82.22±0.89	82.08±0.39	53.30±1.11	53.60±0.68	70.95±0.63
NPC	90.83±0.62	95.04±0.61	79.48±1.97	78.88±1.47	50.73±1.70	47.53±1.26	70.60±0.51
DyGen	91.13±0.30	95.41±0.28	82.88±0.71	84.80±0.86	60.86±0.81	60.79±2.23	71.42±0.31
SiDyP	93.63±0.84	95.97±0.15	84.76±0.79	85.60±0.44	64.26±0.27	64.79±0.96	72.66±0.58

Risk of Blindly Following Influencer Advice

- Paper: VideoConviction: A Multimodal Benchmark for Human Conviction and Stock Market Recommendations (KDD 2025)



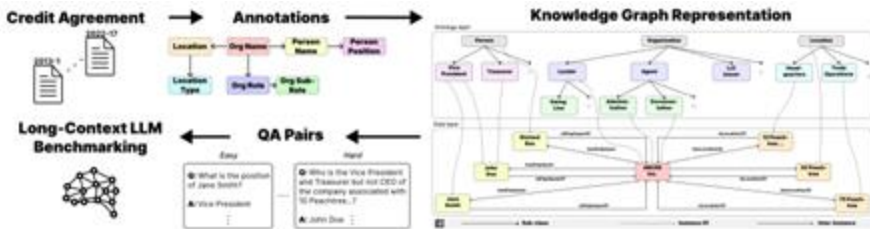
- Benchmark MLLMs and LLMs on full-length & segmented content to see if non-textual signals and reduced noise improve model performance.
- Identified model-human gap in interpreting informal, multimodal dialogue.
- Discovered inverse-influencer strategies are more profitable (*higher risk).

Extracting Information from Credit Agreements with LLMs

- **Paper:** KG-QAGen: A Knowledge-Graph-Based Framework for Systematic Question Generation and Long-Context LLM Evaluation

- Credit agreements often span dozens of pages, and while modern LLMs claim to handle entire documents at once, our study asks whether they are truly reliable at extracting the specific information that financial professionals care about
- We design a KG-based pipeline that extracts questions **at varying levels of complexity** from ground-truth annotations, enabling us to pinpoint not just whether LLMs fail, but also why they fail

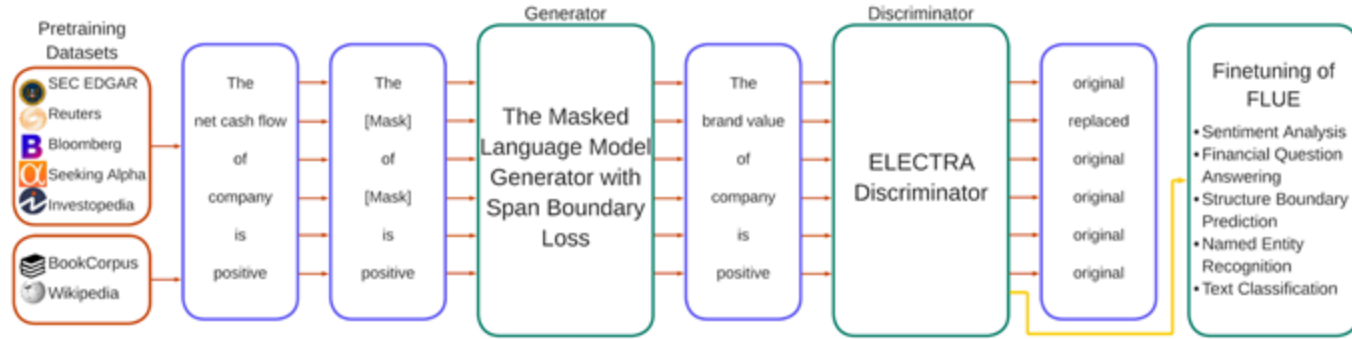
P	H	#SO	Example Template	Hop Path	Set Operation
0	1	0	What is the position of [Person Name]?	Person→Position	None
1	1	0	What are the positions of [Person Name]?	Person→Position	None
0	1	1	What position is held by both [Person A] and [Person B]?	Person→Position	$A \cap B$
0	1	2	What position does [Person A] hold that [Person B] doesn't?	Person→Position	$A \cap \sim B$
0	1	3	What position is held by [Person A] but not [Person B] or [Person C]?	Person→Position	$A \cap \sim B \cap \sim C$
1	1	1	What are the positions held by both [Person A] and [Person B]?	Person→Position	$A \cap B$
1	1	2	What are the positions held by [Person A] but not [Person B]?	Person→Position	$A \cap \sim B$
1	1	3	What are the positions held by [Person A] but not [Person B] or [Person C]?	Person→Position	$A \cap \sim B \cap \sim C$
0	2	0	Who is the [Position] of [Org]?	Org→Pos→Per	None
1	2	0	What are the roles of [Org] in the agreement where [Person] is employed?	Per→Org→Role	None
0	2	1	Who is the [Position A] and [Position B] of [Org]?	Org→Pos→Per	$A \cap B$
1	2	1	Who are both [Position A]s and [Position B]s of [Org]?	Org→Pos→Per	$A \cap B$
0	3	0	Who is the [Position] of the company associated with [Location]?	Loc→Cr→Ps→Pe	None
1	3	0	Who are the [Position]s of the company associated with [Location]?	Loc→Cr→Ps→Pe	None
0	3	1	Who is both [Position A] and [Position B] of the company associated with [Location]?	Loc→Cr→Ps→Pe	$A \cap B$
0	3	2	Who is [Position A] but not [Position B] of the company associated with [Location]?	Loc→Cr→Ps→Pe	$A \cap \sim B$
0	3	3	Who is [Position A] and [Position B] but not [Position C] of the company associated with [Location]?	Loc→Cr→Ps→Pe	$A \cap B \cap \sim C$
1	3	1	Who are both [Position A] and [Position B] of the company associated with [Location]?	Loc→Cr→Ps→Pe	$A \cap B$
1	3	2	Who are [Position A]s but not [Position B]s of the company associated with [Location]?	Loc→Cr→Ps→Pe	$A \cap \sim B$



- Our study shows that LLMs often stumble on subtle but high-impact details, such as handling co-reference across clauses, interpreting nested roles, and understanding document structure: issues that directly affect the reliability of extracting obligations, counterparties, and risk exposures from credit agreements

FLANG and FLUE

- **Paper: When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain (EMNLP 2022 main)**



- **Advantages over FinBERT**

- Trained with large finance-specific dictionary
- Preferential phrase masking
- Span boundary objective and in-filing objective

- **FLUE**

- An open-source comprehensive suite of benchmarks for the financial domain
- Sentiment analysis (regression and classification), NER, Q&A, SBD

The **FinGT Ecosystem** is an open-source methodology, dataset, model family, and code demonstrating the best practices in industry-standard LLM-powered AI systems for finance, economics, and business

1. FinPILE

- Industrial-strength collection and parsing pipeline for large language models. This pipeline enables large-scale data scraping, collection, and transformation from rich documents and multimedia sources for LLM training.
- **Research Question:** How do we strategically choose, collect, and curate high-quality, reliable data sources for training large language models in the financial domain?

3. FERRArI (*Financial Economics Reasoning Reinforcement for AI*)

- State of the art reinforcement learning using synthetic instruction-based fine-tuning data to enhance the reasoning and decision-making capabilities of large language models in highly specific financial contexts.
- **Research Question:** How can we autonomously generate finance data needed for LM post-training (supervised fine-tuning, reinforcement learning with verified rewards, etc.) ?

2. FinLMs

- Training various sizes, types, and architectures of large language models for financial services, optimizing them for specific domains and tasks without fine-tuning, to achieve high performance in core tasks, incl. reasoning, decision-making.
- **Research Question:** How can we train a family of LMs with different data compositions and training regimes to provide Pareto-optimal performance/cost trade-offs?

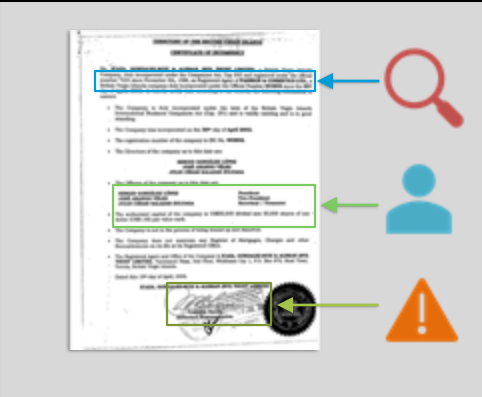
4. FLaME (*Financial Language Model Evaluation*)

- Comprehensive evaluation benchmark suite for large language models, focusing on robust and reliable machine-driven natural language evaluation.
- **Research Question:** How can we determine the most effective methods to benchmark AI financial domain performance?

Objective: Partner humans with a collaborative AI assistant to decrease the time required to identify critical information from unstructured data needed for experts in knowledge-intensive fields to discover the information they need to make high-stakes decisions for real-time tasks.



Machine-Assisted Global Intelligence Curation (MAGIC)



Current Status

Credit Rating: AAA (WAAA)

Stock Price: \$3,750.50 / \$500.25

Market Moves: Up by 1.75%

Recent Transactions: Acquired XYZ Corp

Financial Metrics

Revenue: \$75 Billion / \$10 Billion

Profit Margin: 25%

Debt Level: \$15 Billion / \$2 Billion

Market Share: 15%

General Information

Entity Name: Shenzhen Semiconductor Manufacturing

Founding Date: April 1, 2013

Employees: 54,000

Headquarters: Shenzhen, China

Industry: Semiconductor

Purpose/Users

Core Business: Semiconductor Manufacturing

Primary Clients: Global Electronics Companies

Virtual Assistant Interaction

USER: Are there any security concerns with our latest chip exports?

AGENT: There have been reports of potential interception, we're increasing our data encryption standards.

USER: How do we ensure that our technology does not fall into the wrong hands?

AGENT: We follow strict export control laws and conduct thorough vetting of our partners.

Your message...

Last Updated: 2023-11-16 22:40:19



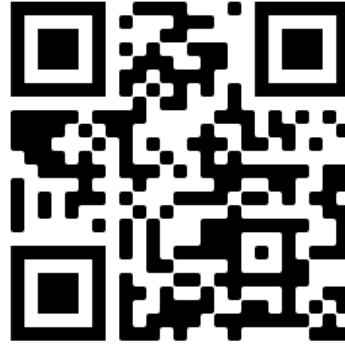
Fig 1-1, ATP 2-33.4, 2020



Follow Us



 @gtfintechlab



 fintech.gatech.edu



 LinkedIn

Q&A