NEURAL INFORMATION PROCESSING SYSTEMS

FinBen: An Holistic Financial Benchmark for Large Language Models

Qianqian Xie[a,b], Weiguang Han[b] , Zhengyu Chen[b] , Ruoyu Xiang[a], Xiao Zhang[a] , Yueru He[a] , Mengxi Xiao[b], Dong Li[b], Yongfu Dai[g], Duanyu Feng[g], Yijing Xu[a], Haoqiang Kang[e], Ziyan Kuang[l], Chenhan Yuan[c], Kailai Yang[c], Zheheng Luo[c], Tianlin Zhang[c], Zhiwei Liu[c], Guojun Xiong[j], Zhiyang Deng[i], Yuechen Jiang[i], Zhiyuan Yao[i], Haohang Li[i], Yangyang Yu[i,*], Gang Hu[h], Jiajia Huang[k], Xiao-Yang Liu[e,*], Alejandro Lopez-Lira[d,*], Benyou Wang[f], Yanzhao Lai[m], Hao Wang[g], Min Peng[b,*], Sophia Ananiadou[c,*], Jimin Huang[a,*]

[a]The Fin AI, [b]Wuhan University, [c]The University of Manchester, [d]University of Florida, [e]Columbia University, [f]The Chinese University of Hong Kong, Shenzhen, [g]Sichuan University, [h]Yunnan University, [i]Stevens Institute of Technology [j]Stony Brook University, [k]Nanjing Audit University, [l]Jiangxi Normal University, [m]Southwest Jiaotong University

SUPPORTED BY FINOS

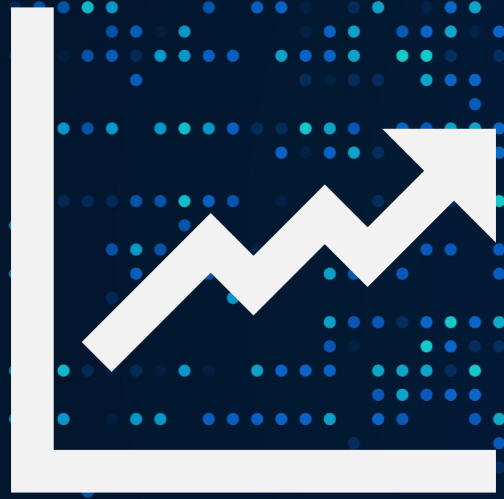
Introduction

1. Background
2. Existing Method
3. Contributions

Background

- The advanced LLMs exhibit **remarkable capabilities** on financial text analysis and prediction tasks

- The fast development of LLM highlights the need for **financial evaluation benchmarks**



Existing Method

- General-domain benchmarks: MMLU, HELM, BIG-bench
- Financial-domain evaluation benchmarks: FLUE, BBTCFLEB, PIXIU
- Challenges:

  Limited Evaluation Tasks: primarily focus on Financial NLP Tasks, ignoring forecasting, risk management, and decision-making, etc.

Table 1: Comparison of different financial benchmarks based on the number of tasks and datasets used, and the task number distribution across various aspects including information extraction (IE), textual analysis (TA), question answering (QA), text generation (TG), risk management (RM), forecasting (FO), decision-making (DM), and spanish (SP).

Benchmark	Language	Dataset	Task	IE	TA	QA	TG	RM	FO	DM	SP
CFBenchmark	Chinese	8	7	1	3	✗	3	✗	✗	✗	✗
Fin-Eva	Chinese	1	1	✗	✗	1	✗	✗	✗	✗	✗
PIXIU	English	15	8	1	3	1	1	1	1	✗	✗
FinanceBench	English	1	1	✗	✗	1	✗	✗	✗	✗	✗
BizBench	English	8	5	2	✗	2	1	✗	✗	✗	✗
FinBen	English,Spanish	42	24	6	8	3	1	4	1	1	6

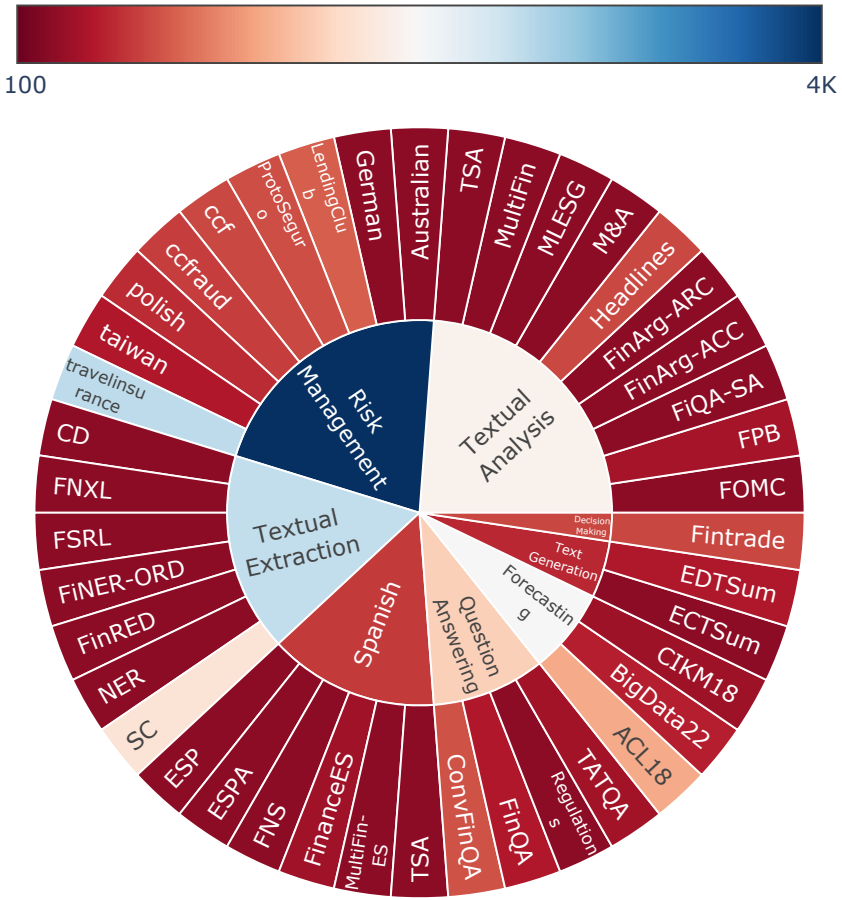
Contributions

★ First Comprehensive FinLLMs Benchmark

- 42 datasets spanning **24** financial tasks
- Organized into **8** Categories: Information extraction (IE), textual analysis (TA), question answering (QA), text generation (TG), risk management (RM), forecasting (FO), decision-making (DM), and Spanish (SP)

★ Features

- ***New Tasks:*** Introduces a significantly larger number of tasks and datasets
- ***Broader Coverage:*** Covering seven aspects of the financial sector. The first benchmark to include the evaluation of stock trading
- ***New Evaluation Strategy:*** The first benchmark to include agent-based evaluation and Retrieval-Augmented Generation (RAG) based evaluation
- ***Novel Datasets:*** Two novel open-source datasets of QA and stock trading tasks
- ***Empowering Financial LLMs Research:*** Hosted the first shared task focused on financial LLMs at the FinNLP-AgentScen workshop during IJCAI-2024, attracting 35 registrations and 12 teams.



FinBen

1. Taxonomy
2. Data Sources
3. Tasks
4. Evaluations

Taxonomy

- **Information Extraction:** focuses on identifying key entities and relationships within financial documents, transforming unstructured data into structured insights
- **Textual Analysis:** delves into content and sentiment analysis of financial texts, aiding in market trend understanding
- **Question Answering:** evaluates the model's ability to comprehend and respond to financial queries
- **Text Generation:** assesses the production of coherent financial text
- **Risk Management:** involves evaluating creditworthiness, detecting fraud, and ensuring regulatory compliance
- **Forecasting:** predicts future financial trends, enabling strategic responses to market dynamics
- **Decision-Making:** assesses the model’s proficiency in making informed financial decisions, such as developing trading strategies and optimizing investment portfolios
- **Spanish:** evaluates the model’s multilingual capabilities, particularly in low-resource languages.

Data Sources

- **Open-sourced datasets from existing studies**:
  - Designed diverse prompts by domain experts
  - Reformulated into instructions
- **Datasets from existing evaluation benchmarks**
  - **PIXIU, Flare-ES, etc.**
- **Novel datasets**
  - FinTrade**:** Integrating historical stock prices, filings data, and news data for 10 stocks over a one-year period
  - **Regulations:** Long-form question answering (QA) related to Overthe-Counter (OTC) derivatives and financial regulations within the European Union

Table 6: Summary of FinTrade dataset statistics.

Ticker	Number of News	Number of 10-K/10-Q Files	Numerical Price Data
TSLA	3,233	8	497
NFLX	965	8	497
AMZN	1,675	8	497
MSFT	1,362	8	497
AAPL	2,082	8	497
GOOG	1,144	7	497
DIS	1,445	9	497
GM	2,252	9	497
NIO	957	0	497
COIN	1,022	0	497

Tasks

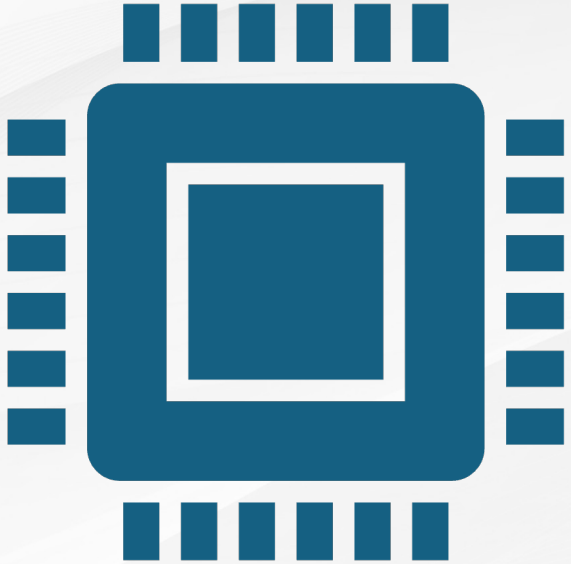
- Information Extraction:
  - **Named entity recognition:** NER, FINER-ORD; **Relation Extraction:** FINRED; **Causal Classification:** SC; **Causal detection:** CD; **Numeric Labeling:** FNXL; **Textual analogy parsing:** FSRL;
- Textual Analysis:
  - **Sentiment analysis:** the Financial Phrase Bank (FPB), FiQA-SA, TSA; **News headline classification:** Headlines; **Hawkish-Dovish classification:** FOMC; **Argument unit classification:** FinArg AUC; **Argument relation detection:** FinArg ARC; **Multi-class classification:** MultiFin; **Deal completeness classification:** MA; **ESG issue identification:** MLESG
- **Question Answering**
  - **Numerical QA:** FinQA, TATQA; **Multi-turn QA**: ConvFinQA; **Long-form QA:** Regulations;
- **Text Generation**
  - **Text summarization:** ECTSUM, EDTSUM
- **Forecasting**
  - **Stock movement prediction:** BigData22, ACL18, CIKM18;
- **Risk Management**
  - **Credit scoring:** German, Australia, LendingClub; **Fraud detection:** ccf, CCFraud; **Financial distress identification:** Polish, Taiwan; **Claim analysis:** PortoSeguro, travelinsurance
- **Trading**
  - **Stock trading (Agent-based): FinTrade**
- **Spanish**
  - **Sentiment analysis: TSA, FinanceES; Multi-class Classification: MultiFin-ES; QA: EFP, EFPA; Summarization: FNS;**

Table 2: The tasks, datasets, data statistics, and evaluation metrics included in FinBen. We use on test data for evaluation. Datasets marked with an asterisk (*) are newly constructed by us, comprisi 10.32% of the total data. EM Accuracy means the exact match accuracy.

Data	Task	Test	Evaluation	License
NER (Alvarado et al., 2015)	named entity recognition	980	Entity F1	CC BY-SA 3.0
FiNER-ORD (Shah et al., 2023b)	named entity recognition	1080	Entity F1	CC BY-NC 4.0
FinRED (Sharma et al., 2022)	relation extraction	1,068	F1, Entity F1	Public
SC (Mariko et al., 2020)	causal classification	8,630	F1,Entity F1	CC BY 4.0
CD (Mariko et al., 2020)	causal detection	226	F1,Entity F1	CC BY 4.0
FNXL (Sharma et al., 2023)	numeric labeling	318	F1,EM Accuracy	Public
FSRL (Lamm et al., 2018)	textual analogy parsing	97	F1, EM Accuracy	MIT License
FPB (Malo et al., 2014)	sentiment analysis	970	F1, Accuracy	CC BY-SA 3.0
FiQA-SA (Maia et al., 2018)	sentiment analysis	235	F1	Public
TSA (Cortis et al., 2017)	sentiment analysis	561	F1, Accuracy	CC BY-NC-SA 4.0
Headlines (Sinha and Khandait, 2021)	news headline classification	2,283	Avg F1	CC BY-SA 3.0
FOMC (Shah et al., 2023a)	hawkish-dovish classification	496	F1, Accuracy	CC BY-NC 4.0
FinArg-ACC (Sy et al., 2023)	argument unit classification	969	F1, Accuracy	CC BY-NC-SA 4.0
FinArg-ARC (Sy et al., 2023)	argument relation classification	496	F1, Accuracy	CC BY-NC-SA 4.0
MultiFin (Jørgensen et al., 2023)	multi-class classification	690	F1, Accuracy	Public
MA (Yang et al., 2020a)	deal completeness classification	500	F1, Accuracy	Public
MLESG (Chen et al., 2023a)	ESG Issue Identification	300	F1, Accuracy	CC BY-NC-ND
FinQA (Chen et al., 2021)	question answering	1,147	EM Accuracy	MIT License
TATQA (Zhu et al., 2021)	question answering	1,668	F1, EM Accuracy	MIT License
*Regulations	long-form question answering	254	ROUGE, BERTScore	Public
ConvFinQA (Chen et al., 2022b)	multi-turn question answering	1,490	EM Accuracy	MIT License
ECTSum (Mukherjee et al., 2022)	text summarization	495	ROUGE, BERTScore, BARTScore	Public
EDTSum (Xie et al., 2023b)	text summarization	2000	ROUGE, BERTScore, BARTScore	Public
BigData22 (Soun et al., 2022)	stock movement prediction	1,470	Accuracy, MCC	Public
ACL18 (Xu and Cohen, 2018)	stock movement prediction	3,720	Accuracy, MCC	MIT License
CIKM18 (Wu et al., 2018)	stock movement prediction	1,140	Accuracy, MCC	Public
German (Hofmann, 1994)	credit scoring	1000	F1, MCC	CC BY 4.0
Australian (Quinlan, [n. d.])	credit scoring	690	F1, MCC	CC BY 4.0
LendingClub (Feng et al., 2023)	credit scoring	2,690	F1, MCC	CC0 1.0
ccf (Feng et al., 2023)	fraud detection	2,278	F1, MCC	(DbCL) v1.0
ccfraud (Feng et al., 2023)	fraud detection	2,097	F1, MCC	Public
polish (Feng et al., 2023)	financial distress identification	1,736	F1, MCC	CC BY 4.0
taiwan (Feng et al., 2023)	financial distress identification	1,364	F1, MCC	CC BY 4.0
ProtoSeguro (Feng et al., 2023)	claim analysis	2,381	F1, MCC	Public
travelinsurance (Feng et al., 2023)	claim analysis	3,800	F1, MCC	(ODbL) v1.0
*FinTrade	stock trading	3,384	CR, SR, DV, AV, MD	MIT License
MultiFin	multi-class classification	2,066	F1,Accuracy	MIT License
FNS-2023	text summarization	232	ROUGE,BERTScore,BARTScore	Public
EFP	question answering	37	F1,Accuracy	Public
EFPA	question answering	228	F1,Accuracy	Public
TSA	sentiment analysis	3,892	F1,Accuracy	Public
FinanceES	sentiment analysis	7,980	F1,Accuracy	Public

Evaluations

- **21** representative general LLMs and financial LLMs
  - General LLMs:
    - Commercial APIs: ChatGPT, GPT-4, Gemini Pro
    - Open-source: LLaMA2-70B, ChatGLM3-6B, Baichuan2-6B, InternLM-7B, Falcon7B, Mixtral 8×7B, Code Llama-7B, Qwen 2 7B/72B, LLaMA3.1-8B/70B
  - Financial LLMs: FinGPT, FinMA-7B. DISCFinLLM, CFGPT, Xuanyuan 6B/70B
- **16** A100 80GB GPUs x **600** GPU hours ($51,000)



Results

Overall Results

- **Information Extraction & Textual Analysis Results**
  - GPT-4 stands the best
  - FinMA-7B is the best open-source LLMs, even better than GPT-4 on Quantification tasks due to domain fine-tuning, but struggles in IE tasks.
- **Question Answering & Text Generation Results**
  - GPT-4 and Gemini-pro is the best in QA and TG tasks
  - Larger models show better performance in summarization tasks
- **Forecasting & Risk Management Results**
  - All LLMs fail to meet expected outcomes and lag behind traditional methodologies.
  - LLMs with low instruction-following abilities tend to classify all cases into a single class in significant imbalanced data of RM tasks
- **Spanish Results**
  - ChatGPT, GPT-4 and Gemini show limited performance compared with English datasets
  - Mixtral 7B performs competitively, showing that the multilingual ability can improve language-specific tasks
  - Smaller models, particularly from the LLaMA family, struggle with domain complexities

Table 3: The zero-shot and few-shot performance of different LLMs on the FinBen. All results via our evaluations are the average of three runs. "-" represents the result that is currently unable to yield due to model size or availability, and "*" represents the result from the previous paper.

Dataset	Metrics	Chat GPT	GPT 4	Gemini	LLaMA2 7B-chat	LLaMA2 70B	LLaMA3 8B	FinMA 7B	FinGPT 7b-lora	InternLM 7B	Falcon 7B	Mixtral 7B	CFGPT sft-7B-Full
NER	EntityF1	0.77*	**0.83***	0.61	0.18	0.04	0.08	0.69	0.00	0.00	0.00	0.24	0.00
FINER-ORD	EntityF1	0.28	**0.77**	0.14	0.02	0.07	0.00	0.00	0.00	0.00	0.00	0.05	0.00
FinRED	F1	0.00	**0.02**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SC	F1	0.80	0.81	0.74	0.85	0.61	0.69	0.19	0.00	**0.88**	0.67	0.83	0.15
CD	F1	0.00	0.01	**0.03**	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FNXL	EntityF1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FSRL	EntityF1	0.00	0.01	**0.03**	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FPB	F1	0.78*	0.78*	0.77	0.39	0.73	0.52	**0.88**	0.00	0.69	0.07	0.29	0.35*
	Acc	0.78*	0.76*	0.77	0.41	0.72	0.52	**0.88**	0.00	0.69	0.05	0.37	0.26*
FiQA-SA	F1	0.60	0.80	0.81	0.76	**0.83**	0.70	0.79	0.00	0.81	0.77	0.16	0.42*
TSA	RMSE↓	0.53	0.50	0.37	0.71	0.57	0.25	0.80	0.00	**0.29**	0.50	0.50	1.05
Headlines	AvgF1	0.77*	0.86*	0.78	0.72	0.63	0.60	**0.97**	0.60	0.60	0.45	0.60	0.61*
FOMC	F1	0.64	**0.71**	0.40	0.35	0.49	0.40	0.49	0.00	0.36	0.30	0.37	0.16*
	Acc	0.6	**0.69**	0.60	0.49	0.47	0.41	0.46	0.00	0.35	0.30	0.35	0.21*
FinArg-ACC	MicroF1	0.50	**0.60**	0.31	0.46	0.58	0.51	0.27	0.00	0.39	0.23	0.39	0.05
FinArg-ARC	MicroF1	0.39	0.40	**0.60**	0.27	0.36	0.28	0.08	0.00	0.33	0.32	0.57	0.05
MultiFin	MicroF1	0.59	**0.65**	0.62	0.20	0.63	0.39	0.14	0.00	0.34	0.09	0.37	0.05
MA	MicroF1	0.85	0.79	0.84	0.70	**0.86**	0.34	0.45	0.00	0.78	0.39	0.34	0.25
MLESG	MicroF1	0.25	**0.35**	0.34	0.03	0.31	0.12	0.00	0.00	0.14	0.06	0.17	0.01
FinQA	EmAcc	0.58*	**0.63***	0.00	0.00	0.06	0.00	0.04	0.00	0.00	0.00	0.00	0.00
TATQA	EmAcc	0.00*	0.13*	**0.18**	0.03	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00
Regulations	Rouge-1	0.12	0.11	-	0.24	-	0.10	0.12	0.01	0.04	0.03	-	0.14
	BertScore	0.64	0.62	-	0.65	-	0.60	0.59	0.40	0.57	0.14	-	0.57
ConvFinQA	EmAcc	0.60*	**0.76***	0.43	0.00	0.25	0.00	0.20	0.00	0.00	0.00	0.31	0.01
EDTSUM	Rouge-1	0.17	0.20	**0.39**	0.17	0.25	0.14	0.13	0.00	0.13	0.15	0.12	0.01
	BertScore	0.66	0.67	**0.72**	0.62	0.68	0.60	0.38	0.52	0.48	0.57	0.61	0.51
	BartScore	-3.64	**-3.62**	-3.87	-3.99	-3.81	-4.94	-5.71	-7.23	-4.60	-6.1	-4.47	-7.08
ECTSUM	Rouge-1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	BertScore	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	BartScore	-5.18	-5.18	-4.93	-5.18	**-4.86**	-5.18	-5.18	-5.18	-5.18	--5.18	-5.18	-5.18
BigData22	Acc	0.53	0.54	**0.55**	0.54	0.47	0.55	0.51	0.45	0.56	0.55	0.46	0.45
	MCC	-0.025	0.03	0.04	0.05	0.00	0.02	0.02	0.00	**0.08**	0.00	0.02	0.03
ACL18	Acc	0.50	**0.52**	0.52	0.51	0.51	0.52	0.51	0.49	0.51	0.51	0.49	0.48
	MCC	0.005	0.02	**0.04**	0.01	0.01	0.02	0.03	0.00	0.02	0.00	0.00	-0.03
CIKM18	Acc	0.55	**0.57**	0.54	0.55	0.49	0.57	0.50	0.42	0.57	0.47	0.42	0.41
	MCC	0.01	0.02	0.02	-0.03	-0.07	0.03	**0.08**	0.00	-0.03	-0.06	-0.05	-0.07
German	F1	0.20	0.55	0.52	**0.57**	0.17	0.56	0.17	0.52	0.41	0.23	0.53	0.53
	MCC	-0.10	-0.02	0.00	**0.03**	0.00	0.05	0.00	0.00	-0.30	-0.07	0.00	0.00
Australian	F1	0.41	**0.74**	0.26	0.26	0.41	0.26	0.41	0.38	0.34	0.26	0.26	0.29
	MCC	0.00	**0.47**	0.00	0.00	0.00	0.00	0.00	0.11	0.13	0.00	0.00	-0.10
LendingClub	F1	0.20	0.55	0.65	**0.72**	0.17	0.10	0.61	0.00	0.59	0.02	0.61	0.05
	MCC	-0.10	-0.02	**0.19**	0.00	0.00	-0.15	0.00	0.00	0.15	-0.01	0.08	0.01
ccf	F1	0.20	0.55	0.96	0.00	0.17	0.01	0.00	1.00	**1.00**	0.10	0.00	0.00
	MCC	-0.10	-0.02	-0.01	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ccfraud	F1	0.20	0.55	**0.90**	0.25	0.17	0.36	0.01	0.00	0.57	0.62	0.48	0.03
	MCC	-0.10	-0.02	0.00	-0.16	0.00	-0.03	-0.06	0.00	-0.13	-0.02	**0.16**	0.01
polish	F1	0.20	0.55	0.86	0.92	0.17	0.83	0.92	0.30	0.92	0.76	**0.92**	0.40
	MCC	-0.10	-0.02	**0.14**	0.00	0.00	-0.06	-0.01	0.00	0.07	0.05	0.00	-0.02
taiwan	F1	0.20	0.55	**0.95**	0.95	0.17	0.26	0.95	0.60	0.95	0.00	0.95	0.70
	MCC	-0.10	-0.02	0.00	**-0.01**	0.00	-0.07	0.00	-0.02	-0.01	0.00	0.00	0.00
portoseguro	F1	0.20	0.55	0.95	0.01	0.17	0.94	0.04	**0.96**	0.96	0.95	0.72	0.00
	MCC	-0.10	-0.02	0.00	-0.05	0.00	-0.01	**0.01**	0.00	0.00	0.00	0.01	0.00
travelinsurance	F1	0.20	0.55	0.00	0.00	0.17	0.00	0.00	**0.98**	0.89	0.77	0.00	0.03
	MCC	-0.10	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	**0.12**	-0.03	0.00	0.01
MultiFin-ES	ACC	0.48	**0.60**	0.23	0.23	0.11	0.25	0.09	0.05	0.13	0.02	0.43	0.30
	F1	0.47	**0.60**	0.14	0.11	0.12	0.27	0.12	0.07	0.17	0.03	0.42	0.27
EFP	ACC	0.30	0.27	0.27	0.27	0.27	0.35	0.27	0.27	0.27	0.24	**0.41**	0.27
	F1	**0.47**	0.19	0.12	0.12	0.12	0.21	0.12	0.12	0.12	0.20	0.41	0.14
EFPA	ACC	0.31	0.34	0.25	0.26	0.20	0.35	0.25	0.26	0.25	0.23	**0.38**	0.32
	F1	0.25	0.27	0.10	0.10	0.09	0.21	0.10	0.10	0.12	0.22	**0.37**	0.18
FinanceES	ACC	0.13	0.15	0.29	0.14	0.20	0.02	0.12	0.15	0.13	0.01	**0.30**	0.05
	F1	0.08	0.09	0.16	0.13	0.23	0.03	0.16	0.18	0.20	0.02	**0.30**	0.05
TSA	ACC	0.21	0.47	0.40	0.07	0.03	0.04	0.02	0.06	0.001	0.02	**0.53**	0.07
	F1	0.24	0.46	0.44	0.04	0.06	0.07	0.04	0.10	0.002	0.04	**0.52**	0.05
FNS	Rouge-1	0.02	0.19	**0.30**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02
	Rouge-2	0.04	0.06	**0.06**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00
	Rouge-L	0.12	0.13	**0.16**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02

Decision Making Results

- GPT-4 shows superior trading performance with a >1 Sharpe Ratio
- All LLMs show better performance than Buy & Hold
- Only Large LLMs (>70B) can generate trading actions while small models (7/13B) fail to handle multiple input sources and complex instructions

Table 4: The average trading performance (95% Confidence Interval) comparison for different LLMs across 10 stocks. The results include large LLMs only ($\geq 70B$), as models with smaller contexts have difficulty understanding the instructions and producing a static strategy of holding.

Model	CR (%)↑	SR↑	DV (%)↓	AV (%)↓	MD (%)↓
Buy & Hold	-4.00 ± 22.39	0.02 ± 0.87	3.59 ± 1.34	56.43 ± 21.00	30.67 ± 17.48
GPT-4	**28.19 ± 25.27**	**1.51 ± 1.08**	2.52 ± 1.30	39.88 ± 20.66	**18.34 ± 9.77**
GPT-4o	-5.54 ± 19.12	-0.19 ± 0.84	2.73 ± 1.30	43.62 ± 20.67	29.96 ± 18.89
GPT3.5-Turbo	4.48 ± 22.23	0.15 ± 0.82	2.84 ± 1.47	45.39 ± 23.35	28.83 ± 15.40
llama2-70B	4.02 ± 24.65	0.52 ± 1.48	2.18 ± 1.28	34.86 ± 20.38	25.55 ± 16.83
llama3-70B	-2.57 ± 22.63	-0.04 ± 1.19	2.71 ± 1.54	43.42 ± 24.65	29.31 ± 15.57
gemini	14.95 ± 28.04	1.03 ± 1.24	**2.17 ± 1.39**	**34.67 ± 22.23**	20.13 ± 11.36

Table 5: Traditional model performances on stock trading.

Model	Cumulative Return	Sharpe Ratio	Standard Deviation	Annualized Volatility	Max Drawdown
A2C	-4.2232	-0.2586	2.7522	43.6898	30.5819
PPO	-0.5586	0.0085	2.7531	43.7048	28.9496
DQN	-2.9924	-0.1656	2.7486	43.6319	31.78

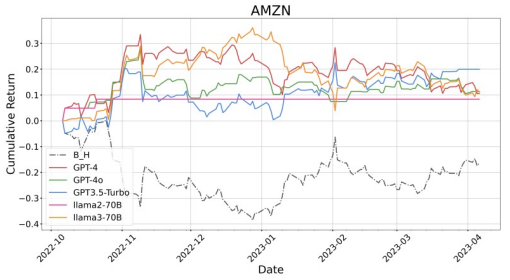


Figure 3: Accumulative Returns of LLM Trading Strategies on AMZN

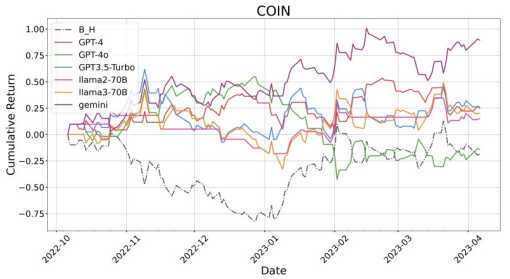


Figure 4: Accumulative Returns of LLM Trading Strategies on COIN

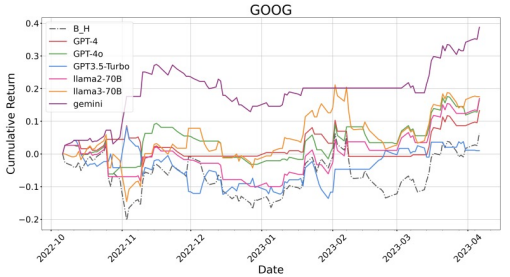


Figure 5: Accumulative Returns of LLM Trading Strategies on GOOG

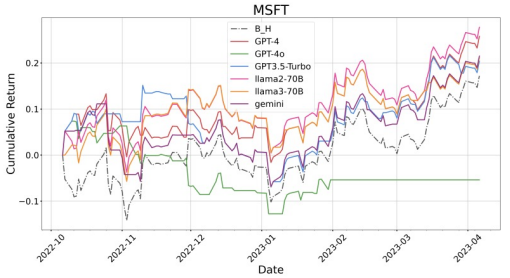


Figure 6: Accumulative Returns of LLM Trading Strategies on MSFT

Conclusion

- Introduces FinBen, the first comprehensive evaluation benchmark for LLMs in finance
- FinBen includes 42 diverse datasets spanning 24 tasks, meticulously organized to assess LLMs across 8 critical aspects: information extraction, textual analysis, question answering, text generation, risk management, forecasting, decision-making, and Spanish.
- Perform evaluations on 21 general and domain LLMs, revealing the potential and bottleneck of LLMs in financial tasks and applications
- Aims to expand FinBen to encompass additional languages and a wider array of financial trading tasks to become the domain-level benchmark

Thank You

Stay tuned with Our leaderboard

https://huggingface.co/spaces/finosfoundation/Open-Financial-LLM-Leaderboard