HKUST IEMS
INSTITUTE FOR EMERGING MARKET STUDIES

FALL 2024 **no.87**

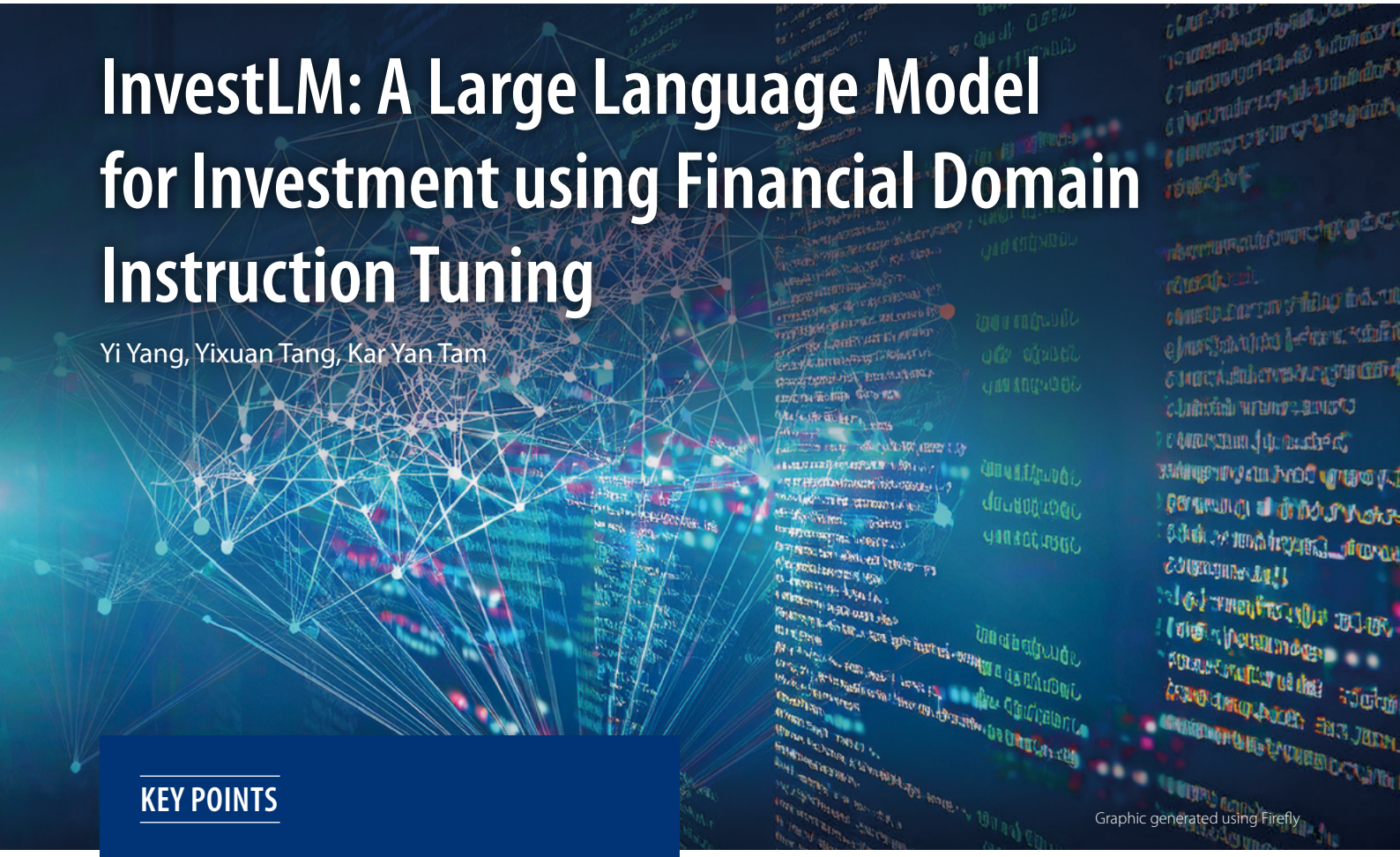
iems.ust.hk

PDF & Additional Materials

THOUGHT LEADERSHIP BRIEF

InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning

Yi Yang, Yixuan Tang, Kar Yan Tam



Graphic generated using Firefly

KEY POINTS

- We present InvestLM, a state-of-the-art Large Language Model for the financial domain, tuned on the LLaMA65B foundation model, using a set of manually crafted instruction datasets covering diverse financial and investment related topics.
- InvestLM shows strong capabilities in understanding financial text and offers helpful insights in response to investment-related inquiries comparable to GPT-3.5, GPT-4, and Claude-2.
- Applying a diverse set of high-quality, domain-specific instructions to train an LLM is more effective in enhancing its capabilities for handling domain-specific tasks than using a large volume of general-purpose instructions.

ISSUE

Large language models (LLMs) have significantly changed the paradigm of natural language processing and hold great potential for artificial general intelligence. Several financial domain LLMs have been developed with the hope of analysing vast volumes of financial texts and enhancing investment and financial decision-making for investors and financial professionals. However, three challenges may hinder the broad development and adoption of financial domain LLMs. First, BloombergGPT, a foundation model with 50 billion parameters trained on Bloomberg's proprietary data, is not publicly available. Second, while other commercialized LLMs such as ChatGPT and Claude-2 are accessible via API, their model parameters are not publicly available either, making it expensive to investigate their financial task capability. Third, the research community has released several LLMs fine-tuned on financial NLP tasks, however, these models exhibit poor performance when generalizing to financial NLP tasks beyond their instructed tasks.

ASSESSMENT

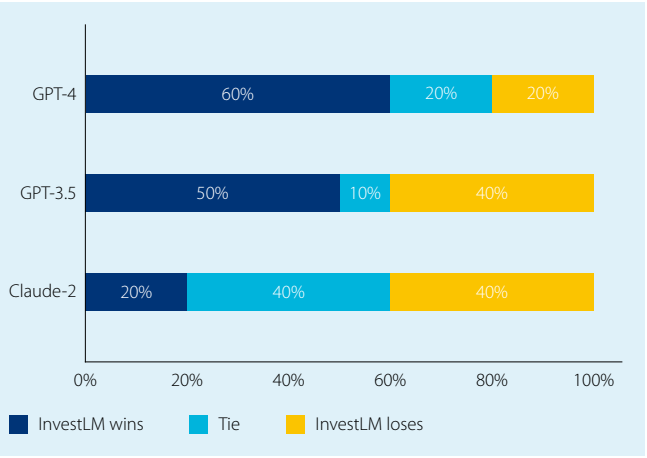
We develop Hong Kong's first open-source LLM for financial generative AI (GenAI) applications, capable of generating investment-related, human-like responses comparable to those of well-known commercial chatbots, including OpenAI's ChatGPT. InvestLM is trained on the LLaMA 65B, using a carefully curated instruction dataset related to finance and investment (Fig. 1). We evaluate InvestLM's utility in providing helpful investment advice by collaborating and interviewing a group of six financial experts, including hedge fund managers and research analysts.

Figure 1. Financial Domain Instruction Dataset

	Size	Input len.	Resp. len.
All	**1,335**	**152.9**	**145.5**
Stackexchange	205	19.4	296.2
CFA	329	125.6	157.4
Academic Journals	200	169.3	74.8
Textbooks	200	128.9	136.6
SEC Filings	80	316.2	88.2
Financial NLP tasks	200	325.9	74.5
Investments	119	72.7	144.3

We manually write 30 test questions that are related to financial markets and investment. For each question, we generate a single response from InvestLM and the three commercial models. We then ask the financial experts to compare InvestLM responses to each of the baselines and label which response is better, or whether neither response is significantly better than the other. In addition to the expert evaluation, we also conduct a GPT-4 evaluation, following the same protocol used in (Zhou et al., 2023). Specifically, we send GPT-4 with exactly the same instructions and data annotations, and ask GPT-4 which response is better or whether neither response is significantly better than the other.

Figure 2. Expert Evaluation



The expert evaluation and GPT-4 evaluation results are presented in Figure 2 and Figure 3. These results indicate that financial experts rate InvestLM's responses as either comparable to or better than those of the GPT-3.5 and GPT-4 models. This expert assessment aligns with GPT-4's own evaluation, which also prefers InvestLM's responses. While financial experts tend to favor Claude-2's responses over InvestLM most of the time, GPT-4 shows a preference for InvestLM's responses. Overall, it is encouraging to observe that our domain-specific instruction tuning effectively generates helpful answers to investment-related questions, especially considering that its foundational model, LLaMA, frequently produces hallucinations (introducing numbers that are not even mentioned in the news). In contrast, InvestLM's responses are grounded in the information presented in the news and reflect logical reasoning with consideration of risks. This highlights the value of domain instruction tuning.

Figure 3. GPT-4 Evaluation

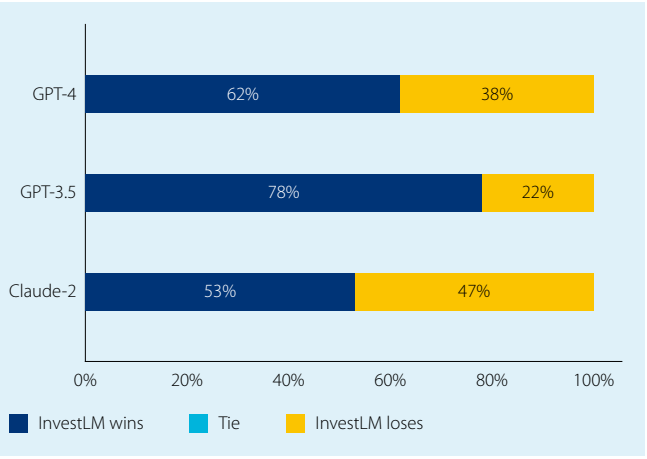


Figure 4. Different LLMs Performance on Financial NLP Benchmarks.

Dataset	Metric	LLaMA-65B	InvestLM	GPT-3.5	BloombergGPT	FinMA	GPT-4
FinSent	Micro-F1	0.71	0.79	0.75	-	0.80	**0.81**
FPB	Micro-F1	0.38	0.71	0.75	0.51	0.88	**0.90**
FOMC	Micro-F1	0.53	0.61	0.60	-	0.52	**0.73**
FiQA	Micro-F1	0.75	0.90	0.77	0.75	0.87	**0.92**
ESG	Micro-F1	0.67	**0.80**	0.64	-	0.51	0.63
FLS	Micro-F1	**0.60**	0.51	0.57	-	0.27	0.57
QA	Micro-F1	0.73	**0.81**	0.71	-	0.68	0.78
FinQA	Acc	0.23	0.29	0.47	-	0.15	**0.54**
ECTSum	Rouge-1	0.14	0.26	0.21	-	0.08	**0.30**
	Rouge-2	0.12	0.12	0.13	-	0.01	**0.15**
	Rouge-L	0.13	0.17	0.15	-	0.06	**0.20**
	CHRF++	23.65	31.53	29.79	-	6.34	**36.31**

We further evaluate InvestLM's performance on financial NLP benchmarks. We consider the following LLMs, including two instruction tuned models GPT-3.5, GPT-4 from OpenAI, two financial LLMs, BloombergGPT (a 50B foundation model) and FinMA (an instruction tuned model on LLaMA-7B), and one foundation model LLaMA-65B, upon which InvestLM is built. These results are presented in Figure 4.

When comparing InvestLM with LLaMA-65, we find that domain instruction tuning is very effective. In 8 out of 9 tasks, InvestLM outperforms LLaMA-65. Second, GPT-4 achieves the best performance in 6 out of the 9 tasks, while InvestLM achieves the best performance in 2 out of the 9 tasks, suggesting that GPT-4 is the state-of-the-art commercial LLM.

To assess the advantages of domain instruction tuning across foundation models of varying sizes, we train an InvestLM-7B model on the LLaMA-7B foundation model using our domain instruction dataset. The relative improvement brought about by domain instruction tuning is considerably more pronounced for the smaller 7B model compared to the 65B model. Domain instruction tuning improves performance by an average of 138.4% across tasks. In contrast, for the LLaMA-65B model, there's a performance increment of 28.2%. The results indicate that in scenarios where computational constraints prevent deploying a 65B model, domain instruction tuning is vital in optimizing the performance of the smaller model.

We also aim to explore whether the inclusion of general-purpose instructions can further enhance themodel's performance in domain NLP tasks. Given that the general-purpose instruction dataset encompasses instructions related to numerical reasoning and sentiment, there is potential that integrating general-purpose instructions could also improve the model's capability in financial NLP tasks. We incorporate the instruction-following data used in the fine-tuning of the Stanford Alpaca model (Taori et al., 2023) comprising 52K instructions into our domain instruction dataset. Using this augmented dataset, we train an InvestLM-7B+AlpacaInstructions model. We then evaluate the utility of generic instructions on the financial NLP benchmarks. The results (Fig 5.) lead to an interesting finding that the inclusion of generic instructions appears to negatively impact the model's generalizability on domain-specific NLP tasks. When comparing InvestLM-7B+Alpaca-Instructions (trained on the combined instruction dataset) to InvestLM-7B (trained solely on the domain instruction dataset), it's evident that InvestLM-7B consistently outperforms InvestLM-7B+Alpaca-Instructions across all Tasks. This underscores the value of our carefully curated domain instructions. This finding suggests that rather than generating a large volume of general-purpose instructions, creating a set of high-quality, domain-specific instructions can be more effective in tapping into a model's capabilities for domain tasks.

Figure 5. Performance of InvestLM-7B Trained Using Different Instruction Dataset

Dataset	Metric	LLaMA-7B	InvestLM-7B	InvestLM-7B+Alpaca-Instructions
FinSent	Micro-F1	0.53	**0.69**	0.64
FPB	Micro-F1	0.12	**0.74**	0.42
FOMC	Micro-F1	0.25	**0.40**	0.32
FIQA ABSA	Micro-F1	0.31	**0.76**	0.40
ESG	Micro-F1	0.19	**0.61**	0.48
FLS	Micro-F1	0.34	**0.53**	0.17
QA	Micro-F1	0.72	**0.84**	0.40
FinQA	Acc	0.07	**0.07**	0.03
EctSUM	Rouge-1	0.06	**0.24**	0.14
	Rouge-2	0.06	**0.10**	0.05
	Rouge-L	0.06	**0.15**	0.09
	Bert Score	0.73	**0.78**	0.75
	CHRF++	12.90	**29.18**	19.48

CONCLUSION

InvestLM shows strong capabilities in understanding financial text and typically arrives at a more concise logical investment conclusion compared to state-of-the-art commercial LLMs. Furthermore, we discover that using a small yet high-quality instruction dataset to fine-tune a large foundational model, yields a promising approach for crafting domain-specific LLMs. Finally, we find that generic instructions, like those used in Alpaca can detrimentally impact the performance of instruction-tuned models on domain tasks. This emphasizes the importance of curating domain specific instructions. Together, our findings provide insights into how to fine-tune a foundation model for a specific domain. We release the parameters of InvestLM and adopt the same licensing terms as LLaMA.

Yi Yang is an Associate professor in the Department of Information Systems, Business Statistics and Operations Management at Hong Kong University of Science and Technology. He is the Director of the Center for Business and Social Analytics (CBSA). He received his PhD in computer science from Northwestern University. He designs machine learning methods in his research to solve challenging business and Fintech problems. His work has been published in business discipline journals such as *Information Systems Research, Management Information Systems Quarterly, Journal of Marketing, Contemporary Accounting Research* and *INFORMS Journal on Computing*. His work has also been published in top-tier machine learning and natural language processing conferences such as *Annual Meeting of the Association for Computational Linguistics (ACL), Conference on Empirical Methods in Natural Language Processing (EMNLP) and International Conference on Artificial Intelligence and Statistics (AISTATS).*

Yixuan Tang is an MPhil student in Information Systems at the Hong Kong University of Science and Technology. Yixuan has cultivated her interest in Natural language processing (NLP) in Finance. She is particularly passionate about adapting Large Language Models in the Finance domain and mining finance signals from text embedding. She has published in machine learning conferences such as the Conference on Empirical Methods in Natural Language Processing (EMNLP) and the Conference on Language Modeling (COLM).

Kar Yan Tam is Vice-President for Administration and Business and Chair Professor of Information Systems, Business Statistics and Operations Management at the Hong Kong University of Science and Technology (HKUST). Prof Tam is known for his contributions in information systems and the diffusion of innovations in organizations. According to *Google Scholar*, his publications have received over 23,000 citations. Prof Tam is currently serving on the editorial board of a number of academic journals. Prof Tam also plays an active role in public services. He is a member of the Hong Kong Exchange Fund Advisory Committee of the Hong Kong Monetary Authority and the Chairperson of the Hong Kong Committee for the Pacific Economic Collaboration.

Read all HKUST IEMS Thought Leadership Briefs at http://iems.ust.hk/tlb

HKUST IEMS
INSTITUTE FOR EMERGING MARKET STUDIES

T: (852) 3469 2215
E: iems@ust.hk
W: http://iems.ust.hk
A: Lo Ka Chung Building, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon

With Support from

EY 安永