AVENIBENCH: Accessible and Versatile Evaluation of Finance Intelligence

Mateusz Klimaszewski[1,2] **Pinzhen Chen**[1,2] **Liane Guillou**[1,2]
Ioannis Papaioannou[1] **Barry Haddow**[1,2] **Alexandra Birch**[1,2]
[1]Aveni.ai [2]University of Edinburgh
Correspondence: mateusz@aveni.ai

Abstract

Over the last few years, there has been great interest in applying large language models (LLMs) to problems in the finance industry, and the field needs a robust LLM benchmark to support this work. Current financial LLM benchmarks contain simple tasks which are not representative of real use cases and have test sets with licences that do not allow commercial use. In response, we release AVENIBENCH, a permissively licensed benchmark that tests a group of six key finance-related skills: tabular reasoning, numerical reasoning, question answering, long context modelling, summarisation and dialogue. We refactor the test sets to ensure that metrics are comparable, providing a unified framework. Furthermore, AVENIBENCH introduces two task difficulty modes, easy and hard, enabling scalable evaluation based on real-world deployment needs. We use our benchmark to evaluate a diverse set of 20 widely used LLMs, from small open-weight models to proprietary systems like GPT-4. This evaluation initiates our public leaderboard, providing valuable insights for future academic research and commercial development.[1]

1 Introduction

Large language models (LLMs) have the potential to automate and enhance labour-intensive processes across a wide range of industries. Finance, as a service industry, is a key sector where LLMs can have a significant impact, due to its large user base (e.g. commercial banking), opportunities for profitability (e.g. investment decisions), and stringent regulatory requirements (e.g. privacy and fairness). Due to the complicated nature of many financial tasks, and the high risks associated with making errors, LLMs developed for the finance domain must be rigorously evaluated prior to deployment. To support this, a number of benchmarks have been proposed, including FinBen (Xie et al., 2024), FLUE (Shah et al., 2022), BizBench (Koncel-Kedziorski et al., 2023), InsightBench (Sahu et al., 2024), and UCFE (Yang et al., 2024).

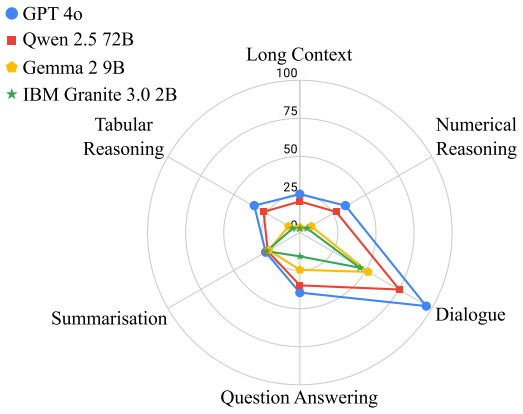

Figure 1: Overview of current capabilities of LLMs on AVENIBENCH. We pick a representative language model for each group/type. See more fine-grained analysis in Table 3.

We find that whilst many existing benchmarks provide good coverage of financial natural language processing (FinNLP) tasks, they are limited in their usefulness for evaluating real-world commercial LLM systems. Specifically, these benchmarks 1) typically adopt a wide range of multiple pre-existing NLP and machine learning datasets with little thought as to their suitability for LLMs (e.g. named entity recognition or sentiment analysis); 2) provide limited insight into the difficulty of tasks or examples; 3) have inconsistent score ranges across diverse test sets; and 4) often include data under restrictive licences making them unfit for commercial purposes, which undermines their value as financial LLMs are going to be heavily used by industry (Li et al., 2023; Nie et al., 2024).

In this paper, we directly address each of these limitations by re-examining existing financial test sets, making appropriate modifications, and filtering out those with a restrictive licence. Our contributions are as follows:

[1] https://huggingface.co/aveni-ai

- We introduce AVENIBENCH, an **open and fully permissive** benchmark for evaluating LLMs in the **finance domain**.
- We format existing datasets and adapt them in order to **unify metrics**. Thanks to this, each dataset has a corresponding and easy-to-compare leading metric, and a ranking-based aggregation.
- Some existing benchmarks proved to be either too easy or too difficult. Considering the scarcity of evaluation resources, we craft **easy and hard** modes that can be chosen based on a downstream use case.
- We evaluate 20 models, from efficient 1B LLMs to large closed-source systems like GPT-4 to present a full picture of performance on AVENIBENCH and a starting point for our leaderboard.

2 Related Work

FinBen (Xie et al., 2024) is the most extensive among the existing benchmarks in the FinNLP domain. It contains an impressive number of 36 tasks; however, we note that (1) not all of them are suited or formatted for LLMs and (2) the majority of the tasks are released with non-permissive licences. Moreover, due to the extensive number of tasks covered, the datasets have been adapted for LLMs but not revisited on an individual task basis.

Other comprehensive finance benchmarks include FLUE (Shah et al., 2022) which contains classification, information extraction, and question answering in the finance domain, as well as BizBench (Koncel-Kedziorski et al., 2023) which includes program synthesis to test reasoning in business and finance scenarios. Some other efforts target more specialised capabilities. FinBench (Yin et al., 2023) focuses on financial risks: credit card default, loan default, credit card fraud, and customer churn – tasks which involve processing large amounts of numerical data but little text data, and we argue are not well suited to LLMs. InsightBench (Sahu et al., 2024) evaluates LLM agents' data analytics in various business use cases. UCFE (Yang et al., 2024) is a multi-turn finance dialogue benchmark covering 17 task types, which is tailored to four distinct user groups: analysts, financial professionals, regulatory professionals, and the general public.

There is also a range of emerging datasets that focus on tabular data and mathematical reasoning – tasks that we also include in our benchmark. In particular, we highlight FinanceMATH (Zhao et al., 2024) a knowledge-intensive financial math reasoning QA dataset, and TableBench (Wu et al., 2024) a tabular QA dataset, for which financial reports make up a third of the data.

3 Benchmark

3.1 Datasets

In AVENIBENCH we include eight datasets that represent a group of six finance-relevant skills: **Tabular Reasoning (TR)**, **Numerical Reasoning (NR)**, **Question Answering (QA)**, **Long Context (LC) Modelling**, **Summarisation (Sum)** and **Dialogue (D)**. Each of the datasets covers at least one of the skills and has a permissive licence that allows for commercial use. Table 1 provides statistics on the number of evaluation examples for each of the datasets (post-filtering, details Section 3.2).

Banking77 [D] (Casanueva et al., 2020) is a fine-grained intent detection dataset for the banking domain, designed to evaluate the classification of user intents in a task-oriented dialogue (ToD) setting.

NLU++ [D] (Casanueva et al., 2022) presents two challenging and realistic ToD tasks for the banking and hotel domains: *multiple-intent detection* (identifying multiple intents in a single utterance) and *slot labelling* (identifying slot values in the utterance). We use exclusively the multi-intent detection task within the banking subset.

FinQA [QA] (Chen et al., 2021) is a QA dataset designed to evaluate numerical reasoning over financial reports, with questions written by experts.

ConvFinQA [QA] (Chen et al., 2022) extends FinQA to construct a multi-turn question answering dataset framed in a conversational setting.

ECTSum [Sum] (Mukherjee et al., 2022) is a long-document summarisation dataset for the specific task of bullet point summarisation of Earnings Calls transcripts. We include the extractive subset.

MultiHiertt [LC, NR, TR] (Zhao et al., 2022) is a QA dataset designed to assess numerical reasoning over long unstructured financial texts containing multiple tables, many of which are hierarchical.

TAT-QA [NR, TR] (Zhu et al., 2021) is a QA dataset combining text and tabular data extracted from financial reports, again requiring numerical reasoning. Unlike in MultiHiertt, most tables in TAT-QA have a flat structure.

Dataset	Test Size	Licence
Banking77	3,080	CC BY 4.0
NLU++$_{\text{EASY}}$	496	CC BY 4.0
NLU++$_{\text{HARD}}$	496	CC BY 4.0
FinQA (from ConvFinQA)	530	MIT
ConvFinQA	1,483	MIT
ECTSum	495	GPL 3.0
MultiHiertt$_{\text{EASY}}$	150	MIT
MultiHiertt$_{\text{HARD}}$	1,007	MIT
TAT-QA	1,663	CC BY 4.0
TAT-HQA	824	Apache 2.0

Table 1: Benchmark details: evaluation examples per dataset & mode, and corresponding licence.

Dataset	Metric
Banking77	Accuracy
NLU++	Accuracy
FinQA	NI
ConvFinQA	NI
ECTSum	RougeL
MultiHiertt	NI
TAT-QA	LM
TAT-HQA	LM

Table 2: Metrics derived for each dataset in the benchmark. NI stands for *numerical identity* accuracy and LM stands for *list match* accuracy.

TAT-HQA [NR, TR] (Li et al., 2022) is a modified version of TAT-QA, where hypothetical facts are added to each question, overriding the facts presented in the report.

For the NLU++ and MultiHiertt datasets, we provide two *modes*, EASY and HARD, representing different levels of task complexity. The adaptation of the datasets is described in Section 3.2.1.

3.2 Metrics & Filtering

The selected datasets, in their initial form, have various metrics proposed in their reference implementation. However, we discovered multiple problems with using them directly to evaluate LLMs.

Firstly, when a dataset was built for BERT-based models (Devlin et al., 2019), the original evaluation regime had to be adapted. Such a change requires a modification of the dataset, which in turn impacts the metric. For example, the reference NLU++ is a multi-label dataset and the benchmark metric is F1. While we could query an LLM about each label in a binary manner (and keep F1), it would be inefficient. Therefore, we sampled distractors (more in Section 3.2.1) and cast the dataset as a multiple-choice question answering (MQA) style evaluation using accuracy instead of F1 as with MQA we eliminate the problem of class imbalance.

Secondly, we found that tasks using multiple metrics – e.g. MultiHiertt used both F1 and exact match – could easily be simplified. Reducing the dataset to have only numerical answers resulted in discarding just a few samples (e.g. the MultiHiertt dev size was slimmed from 1,044 to 1,007). This approach allows us to reduce the evaluation complexity and compare results exclusively on the numerical identity of reference and prediction.

Based on these findings, we map the datasets to unify and simplify the metrics, limiting the evaluation in AVENIBENCH to the following metrics:

- **Accuracy**: for MQA-style benchmarks.
- **Numerical identity accuracy**: compare numbers. We include a simple post-processing step to handle special signs (e.g. percentage or currency) and use numeric-based instead of string-matching comparison.
- **List match accuracy**: compare a list of possible answers (invariant to order). For such tasks, the model is expected to produce a list of answers.
- **RougeL**: for summarisation tasks (Lin, 2004).

Table 2 presents the metric used for each of the datasets in AVENIBENCH.

3.2.1 Adapting the Difficulty Ratio

Evaluation benchmarks in the financial domain are scarce; therefore, it is crucial to make use of all available resources. By default, a benchmark might be either too easy or too difficult, depending on the evaluated model size. To make use of all available data for different LLM parameter budget buckets, we split two datasets into EASY and HARD.

The NLU++ dataset with a typical number of distractors was too easy for bigger models, reaching over 90% for larger Qwen 2.5 models or Llama-3.1 70B. Therefore, to increase difficulty, we not only increased the number of distractors but also allowed them to have different lengths. The last modification allowed for a distractor to include (or be) a subset of correct labels.

On the other hand, MultiHiertt was too challenging for smaller models (e.g. OLMo 1B has a performance lower than 1%) as the dataset requires long context handling, having a range of 2-7 tables per query. We extracted an easier subset,

Model	Param.	Banking77 (0-shot)	NLU++ EASY (0-shot)	NLU++ HARD (0-shot)	FinQA (0-shot)	ConvFinQA (0-shot)	ECTSum (0-shot)	MultiHiertt EASY (2-shot)	MultiHiertt HARD (0-shot)	TAT-QA (4-shot)	TAT-HQA (4-shot)	AVG	**Borda Count** Score	**Borda Count** Rank
Proprietary LLMs														
GPT-4o	-	96.43	97.59	94.18	16.98	61.43	25.75	27.33	22.84	41.37	48.06	53.20	189	1
GPT-4o-mini	-	94.94	97.04	91.04	10.57	55.83	22.41	15.33	9.43	31.45	22.57	45.06	153	4
Open-weight LLMs														
Qwen 2.5	72B	95.27	97.85	33.02	13.58	55.43	24.61	24.00	16.48	39.63	30.95	43.08	177	2
Qwen 2.5	32B	94.81	96.51	22.04	10.94	55.36	25.16	23.33	13.21	33.19	23.67	39.82	164	3
Llama 3.1	70B	82.11	94.35	17.79	5.47	48.42	20.99	24.00	10.63	35.84	25.85	36.54	148	5
Gemma 2	27B	76.91	94.89	17.34	4.15	47.40	21.84	9.33	7.65	33.01	10.44	32.30	127	6
Qwen 2.5	7B	88.74	89.52	14.87	2.83	43.02	24.44	13.33	7.75	18.82	8.86	31.22	119	7
Mistral Nemo	12B	41.59	82.26	9.95	3.40	41.27	22.86	20.00	7.75	26.70	11.04	26.68	114	8
Mixtral v0.1	8x7B	52.89	88.98	17.11	3.77	43.83	18.32	18.00	5.06	28.80	9.22	28.60	109	9
Gemma 2	9B	57.36	87.36	11.97	5.09	44.37	23.30	0.00	6.45	25.86	9.34	27.11	107	10
Llama 3.1	8B	45.63	63.71	7.03	2.08	39.65	19.67	14.00	5.36	23.93	6.31	22.74	87	11
IBM Granite 3.0	8B	74.46	58.33	4.34	1.51	29.74	25.04	4.00	1.29	20.02	4.13	22.29	74	12
Qwen 2.5	1.5B	82.07	76.88	11.51	0.19	29.00	21.71	6.67	2.38	13.41	4.73	24.86	72	13
Mistral v0.3	7B	27.52	41.40	0.00	0.94	37.09	22.69	1.33	4.17	18.52	5.70	15.94	63	14
IBM Granite 3.0	2B	32.03	63.97	6.37	0.19	21.51	23.27	2.67	0.99	14.97	4.25	17.02	55	15
SmolLM2	1.7B	29.80	28.23	0.00	0.00	25.76	15.99	9.33	4.57	13.95	4.37	13.20	48	16
Gemma 2	2B	27.74	12.90	0.00	0.57	31.56	20.93	0.67	3.97	12.87	3.64	11.49	42	17
Llama 3.2	1B	22.11	9.14	0.00	0.00	23.40	15.08	7.33	3.48	10.22	2.43	9.32	29	18
OLMo	7B	21.14	5.11	0.00	0.00	18.81	16.07	4.00	1.79	8.90	4.49	8.03	26	19
OLMo	1.5B	20.02	16.67	0.00	0.19	3.10	17.19	4.00	0.40	9.68	1.09	7.23	23	20

Table 3: Leaderboard of the evaluated LLMs. The final ranking was established using Borda Count.

which one could expect smaller models to handle, although it is still challenging considering other skills required to solve this dataset: NR and TR. The derived setups are as follows:

- NLU++$_{\text{EASY}}$: 4 options, each of the 3 distractors has the same length as an answer.
- NLU++$_{\text{HARD}}$: 10 options, each of the 9 distractors has a length between 1 and the length of answers.
- MultiHiertt$_{\text{EASY}}$: a subset of queries with at most 3 tables and length of max 4,096 tokens (as per Mistral-7B-v0.3). Additionally, this mode has a few-shot setup (constant examples—**2 shortest** from the training dataset to reduce long context problems that small models might encounter).
- MultiHiertt$_{\text{HARD}}$: zero-shot, has all the samples that might require extremely long context reasoning over multiple tables.

4 Leaderboard

We present the evaluation results on AVENIBENCH in Table 3. We evaluate the models using the `lm-eval-harness` (Gao et al., 2024), which provides a standardised framework for querying LLMs for MQA and generation-based tasks. The scores are normalised following the normalisation of the OpenLLM Leaderboard.[2] To avoid problems with balancing different metrics and handling performance outliers, instead of a naive arithmetic average over the scores, we rank the models using a task-level Borda Count method (Colombo et al., 2022). The Borda Count method assigns points per rank position in each task and, based on the final sum of points, establishes the ranking.

We benchmark 18 open-weight base LLMs and include GPT-4o and GPT-4o-mini for reference. GPT models are instruction-tuned, so we require a direct answer via a system prompt. For a detailed list of evaluated models, see Table 4 in Appendix A. Among open-weight LLMs, the Qwen family outperform the field at all different sizes. The 32B and 1.5B are competitive or even better against bigger models, as Qwen 2.5 32B outperforms Llama 3.1 70B and Qwen 2.5 1.5B has an impressive performance when compared against many models in the 7-9B parameter range.

5 Conclusion and Future Work

In summary, we scrutinised existing FinNLP test sets, modified and adapted data, tasks, and metrics, and finally presented a permissive AVENIBENCH. To ensure that it continues to be useful to the community, we aim to regularly review, adjust (if necessary), and incorporate new tests as they become available. We plan to ingest AVENIBENCH into `lm-eval-harness` to facilitate public contributions that could extend the leaderboard to support missing multilingual and multi-modal evaluations.

[2]Details: OpenLLM Leaderboard documentation

Limitations

AVENIBENCH is based on existing datasets which cover a range of tasks that are relevant to the evaluation of finance LLMs. Whilst the six skill categories in our benchmark cover many of the central tasks that an LLM might be expected to perform, this coverage is far from exhaustive owing to the limited availability of datasets with permissive licences.

We have focused solely on the inclusion of English datasets. Although suitable datasets likely exist in other languages, in our review of available datasets the majority that we found were only available for English. Additionally, in the current state, we restrict the benchmark to text-only tasks, which is a limitation considering the growing popularity of multi-modal LLMs (Bai et al., 2023; Chen et al., 2024; Steiner et al., 2024).

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. We also thank Ben Trevett, Nicole Nisbett, Nikolai Debono, and Proyag Pal for the discussions and feedback that led to this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Lewis Tunstall, Agustín Piqueres, Andres Marafioti, Cyril Zakka, Leandro von Werra, and Thomas Wolf. 2024. SmolLM2 - with great data, comes great performance.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*.

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan CLEMENCON. 2022. What are the best systems? new perspectives on NLP benchmarking. In *Advances in Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint*.

Granite Team. 2024. Granite 3.0 language models.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint*.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2023. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint*.

Moxin Li, Fuli Feng, Hanwang Zhang, Xiangnan He, Fengbin Zhu, and Tat-Seng Chua. 2022. Learning to imagine: Integrating counterfactual thinking in neural discrete reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics.

Mistral AI Team. 2024. Mistral NeMo.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Gaurav Sahu, Abhay Puri, Juan Rodriguez, Amirhossein Abaskohi, Mohammad Chegini, Alexandre Drouin, Perouz Taslakian, Valentina Zantedeschi, Alexandre Lacoste, David Vazquez, et al. 2024. Insightbench: Evaluating business analytics agents through multi-step insight generation. *arXiv preprint*.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, and Zhoujun Li. 2024. Tablebench: A comprehensive and complex benchmark for table question answering. *Preprint*, arXiv:2408.09174.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. FinBen: A holistic financial benchmark for large language models. *arXiv preprint*.

Yuzhe Yang, Yifei Zhang, Yan Hu, Yilin Guo, Ruoli Gan, Yueru He, Mingcong Lei, Xiao Zhang, Haining Wang, Qianqian Xie, et al. 2024. UCFE: A user-centric financial expertise benchmark for large language models. *arXiv preprint*.

Yuwei Yin, Yazheng Yang, Jian Yang, and Qi Liu. 2023. Finpt: Financial risk prediction with profile tuning on pretrained foundation models. *arXiv preprint*.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

A Evaluation details

Table 4 lists the evaluated LLMs, with references to their technical details and specific versions.

The GPT-4o models are instruction-tuned as opposed to foundation models, so we have provided a system-level prompt requiring that it generates an answer directly. Moreover, as the API does not return probabilities of prompt tokens (required for the default MQA configuration as by `lm-eval-harness`), we converted the MQA tasks configuration to generate an answer letter.

Model	Source/Version	Reference
GPT4o	gpt-4o-2024-08-06	Achiam et al. (2023)
GPT4o-mini	gpt-4o-mini-2024-07-18	Achiam et al. (2023)
Qwen 2.5 72B	Qwen/Qwen2.5-72B	Qwen Team (2024)
Qwen 2.5 32B	Qwen/Qwen2.5-32B	Qwen Team (2024)
Qwen 2.5 32B	Qwen/Qwen2.5-7B	Qwen Team (2024)
Qwen 2.5 7B	Qwen/Qwen2.5-1.5B	Qwen Team (2024)
Llama 3.2 1B	meta-llama/Llama-3.2-1B	Dubey et al. (2024)
Llama 3.1 70B	meta-llama/Llama-3.1-70B	Dubey et al. (2024)
Llama 3.1 8B	meta-llama/Llama-3.1-8B	Dubey et al. (2024)
Gemma 2 27B	google/gemma-2-27b	Gemma Team et al. (2024)
Gemma 2 9B	google/gemma-2-9b	Gemma Team et al. (2024)
Gemma 2 2B	google/gemma-2-2b	Gemma Team et al. (2024)
IBM Granite 3.0 8B	ibm-granite/granite-3.0-8b-base	Granite Team (2024)
IBM Granite 3.0 2B	ibm-granite/granite-3.0-2b-base	Granite Team (2024)
Mixtral v0.1 8x7B	mistralai/Mixtral-8x7B-v0.1	Jiang et al. (2024)
Mistral Nemo 12B	mistralai/Mistral-Nemo-Base-2407	Mistral AI Team (2024)
Mistral v0.3 7B	mistralai/Mistral-7B-v0.3	Jiang et al. (2023)
SmolLM2	HuggingFaceTB/SmolLM2-1.7B	Allal et al. (2024)
OLMo 7B	allenai/OLMo-7B-hf	Groeneveld et al. (2024)
OLMo 1.5B	allenai/OLMo-1B-0724-hf	Groeneveld et al. (2024)

Table 4: Evaluated LLM details.