Structure First, Reason Next: Enhancing a Large Language Model using Knowledge Graph for Numerical Reasoning in Financial Documents

Aryan Mishra[1], Akash Anil[1]

[1]Department of Data Science and Engineering,
Indian Institute of Science Education and Research Bhopal, Madhya Pradesh, India

Abstract

Numerical reasoning is an important task in the analysis of financial documents. It helps in understanding and performing numerical predictions with logical conclusions for the given query seeking answers from financial texts. Recently, Large Language Models (LLMs) have shown promising results in multiple Question-Answering (Q-A) systems with the capability of logical reasoning. As documents related to finance often consist of long and complex financial contexts, LLMs appear well-suited for building high-quality automated financial question-answering systems. However, LLMs often face challenges in accurately processing the various numbers within financial reports.

Extracting numerical data from unstructured text and semi-structured tables, and reliably performing accurate calculations, remains a significant bottleneck for numerical reasoning in most state-of-the-art LLMs. Recent studies have shown that structured data augmentations, such as Knowledge Graphs (KGs), have notably improved the predictions of LLMs along with logical explanations. Thus, it is an important requirement to consider inherent structured information in financial reports while using LLMs for various financial analytics.

This paper proposes a framework to incorporate structured information using KGs along with LLM predictions for numerical reasoning tasks. The KGs are extracted using a proposed schema inherently from the document under processing. We evaluated our proposed framework over the benchmark data FinQA, using an open-source LLM, namely Llama 3.1 8B Instruct. We observed that the proposed framework improved execution accuracy by approximately 12% relative to the vanilla LLM.

1 Introduction

Numerical Reasoning in financial data refers to the analysis and interpretation of quantitative information such as revenue figures, ratios, market indicators, or statistical trends present in the financial reports (Chen et al. 2021). Although LLMs demonstrate promising reasoning capabilities in various domains, they often show limited performance when subjected to financial reasoning (Qian et al. 2025; Liu et al. 2025). The limited capability of LLMs is mainly due to the quantitative characteristics of financial data incorporating multiple paragraphs and tables with numbers, which makes it harder to exploit the inherent context (Nie et al. 2024).

A majority of the LLMs show promising reasoning capability yet they are often limited in some of the specialized domains (e.g., finance, healthcare) because they primarily learn from unstructured text data, relying on statistical co-occurrences rather than inherent relational characteristics (Tan et al. 2024). To address this limitation, recently some of the studies integrate structured information such as Knowledge Graphs (KGs) for enhancing the reasoning abilities of LLMs (Sun, Wang, and Li 2024). KGs provide semantic relationships and factual grounding and thus found to be helpful in improving reasoning performance in many domains (Wu and Tsioutsiouliklis 2024). However, to the best of our knowledge, none of these studies explicitly address numerical reasoning over financial data while capturing inherent structural aspects.

To bridge the gap in exploiting structural information in numerical reasoning for finance data, we propose a novel framework that uses inherent KG extracted using predefined schema and an open-source LLM. Figure 1 presents an end-to-end pipeline for the proposed framework. Our framework (i) preprocesses documents (including table linearization), (ii) constructs knowledge graphs using predefined schema, financial entities, and temporal relationships using few-shot prompting, (iii) performs lightweight retrieval combining semantic and structural features, and (iv) reasons using any LLM exploiting the structured input for predicting the output.

We evaluate the proposed framework[1] using Llama 3.1 8B Instruct (Llama)(Grattafiori et al. 2024) on state-of-the-art financial reasoning benchmark namely FinQA (Chen et al. 2021). Further, we systematically compare the performance of the proposed framework to the open-source Llama model. It is evident that the proposed framework using KGs considerably enhanced the performance of vanilla Llama model.

To summarize, the main contributions of this paper are:

1. Study numerical reasoning in financial data using LLMs exploiting structural information in the form of Knowledge Graphs.
2. Build an end-to-end pipeline for numerical reasoning capable of preprocessing, extracting KGs, retrieval, and en-

---

[1]Code will be released publicly upon publication.

hanced reasoning.

3. Systematically compare the results against suitable baseline.

Rest of the paper is organized as follows. Section 2 presents related studies which is followed by a detailed discussion on the proposed framework in Section 3. In Section 4, we discuss the experimental setup and Section 5 discusses the performance of proposed framework compared with the baseline. Section 6 concludes the paper. Further, Section 7 presents a limitation overview.

2 Related Work

The FinQA benchmark (Chen et al. 2021) formalized numerical question answering over financial documents and shown various analytics. In a similar direction, authors of APOLLO in (Sun et al. 2024) introduced number-aware negative sampling and (Li et al. 2023) incorporates dynamic reranking in financial reasoning. Recently, (Qian et al. 2025) proposed a domain specific finetuning framework that enhances reasoning capability. Critically,(Qian et al. 2025) observe that even heavily fine-tuned models (FinR1 (Liu et al. 2025), Dianjin-R1 (Zhu et al. 2025)) exhibit performance degradation on longer and complex documents, falling below base models. Recently, (Srivastava et al. 2024) categorized FinQA queries into arithmetic operations such as SUM, DIFFERENCE, RATIO, CHANGE_RATIO, highlighting various challenges in multi-step domain-specific reasoning. These studies highlight major bottlenecks in numerical reasoning over financial text often considering longer and tabular structure.

To improve the reasoning using LLMs, Retrieval Augmented Generation (RAG) has garnered attention. RAG-based frameworks ground LLM outputs with external evidence (Lewis et al. 2020) to predict a more curated output and thus enhances the performance of LLMs. HybridRAG (Sarmah et al. 2024) combines vector and graph retrieval, though concatenating contexts increases cognitive load and in financial contexts, naive text chunking disrupts numerical links. Thus, our proposed framework attempts to mitigate this issue by fusing semantic and structural features in a lightweight retriever, optimizing retrieval without context concatenation overhead.

In the recent past, SubgraphRAG (Li, Miao, and Li 2025) demonstrates that lightweight MLPs with engineered graph features outperform complex graph neural networks for retrieval. In a similar direction, we aim to adapt curated domain-specific attributes such as temporal distances and entity types easily derived from FinQA dataset.

3 Proposed Framework

Figure 1 presents the proposed framework comprising of three main steps. After getting the input, the first step executes data preprocessing by table linearization and text concatenation with normalization for providing a uniform input text. We linearize tables following prior work on financial QA (Chen et al. 2021). Further, step two automatically extracts KG triplets from the input text provided from step one using a predefined schema proposed for financial reasoning.

In step three, the framework filters the unwanted triplets and performs the reasoning task using preferred LLM model.

3.1 Step 1: Document Preprocessing

Financial documents often contain hybrid data, i.e., narrative texts and semi-structured tables. To enable uniform processing, we linearize tables using template-based conversion shown below:

```
Original Table    Linearized Text
Year | Revenue

2020 | $100M   → "For 2020, revenue is $100M.

2021 | $120M   →  For 2021, revenue is $120M."
```

Text Linearization enables text-based processing though loses explicit structural temporal relationships, and thus it might be the case that entity types are not differentiated which makes the numerical formats inconsistent. This inconsistency might be alleviated using the KG construction preserving temporal relationships.

3.2 Step 2: Knowledge Graph Construction

Knowledge Graphs are one of the key features in our framework. Thus, for constructing the KGs, we leverage the financial context understanding of LLMs and follow a standard schema for generating triplets. We aim to represent the information present in the text in unambiguous format and reduce the potential extraction errors. The inherent temporal and entity relational features are also preserved in this format. The structured triplets enable better processing and understanding of the multi-hop relations present across the document. The schema for such KGs is focused on the numerical and temporal facts. The proposed schema for KG is given below:

Financial Domain Schema

```
Schema  = (subject, relation, object,
           {financial_metric_entity_type,
              company,period,value, unit})

Example:
subject: "NET_REVENUE:Entergy"
relation: "HAS_VALUE_IN_2015"
object: "5829 million USD"
financial_metric_entity_type: "NET_REVENUE"
company: "Entergy"
period: "2015"
value: "5829"
unit: "million USD"
```

Now, for generating triplets across documents, we use few-shot-based prompts designed as below:

Prompt Engineering We provide comprehensive extraction rules via natural language prompt (key excerpts):

```
Extract financial facts containing:
1. Detailed metric (NET_REVENUE,
   OPERATING_EXPENSES, etc.)
2. Numerical value with units
3. Temporal qualifier (infer from context)
```
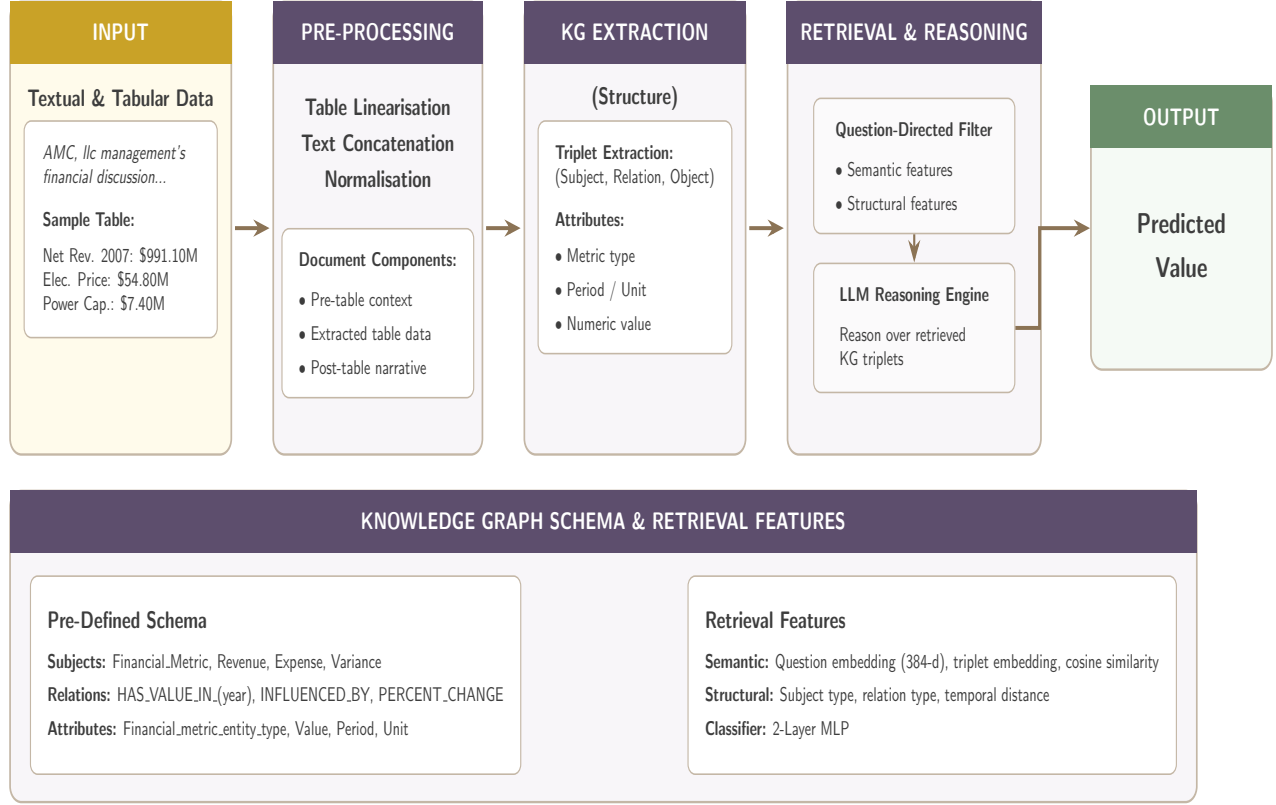
Figure 1: end-to-end pipeline for numerical reasoning in financial data with inherent knowledge graph and LLMs.

```
RULES:
- Use EXACT TEXT (no paraphrasing)
- Standardize periods: "2007", "AS_OF_2010",
  "2007-Q4", "AFTER_2015"
- Extract liberally for coverage
- Multiple periods → separate extractions

ATTRIBUTE REQUIREMENTS:
- SUBJECT: "METRIC:Company" or "METRIC"
- RELATION: "HAS_VALUE_IN_YYYY" variants
- OBJECT: "value unit" (human-readable)
- PERIOD: Standardized format
```

3.3 Step 3: Filtering Triplets and Reasoning

Similar to SubgraphRAG (Li, Miao, and Li 2025), we integrate semantic embeddings and KG embeddings to a Multi-Layer Perceptron (MLP) classifier. The MLP classifier outputs relevant triplets and works as a filter. Further, any preferred LLM may be used for numerical reasoning task with the filtered triplets.

4 Experimental Setups

In this section, we discuss the experimental setups including dataset analysis, evaluation, and baseline selection.

4.1 Dataset

In this paper we use the FinQA benchmark as the dataset. FinQA consists of 6251 training examples, 883 validation examples, and 1147 test examples. We select FinQA as it is one of the standard benchmarks used in multiple studies on financial reasoning (Zhu et al. 2025), (Liu et al. 2025), (Srivastava et al. 2024).

In particular to this paper, the schema for KG extraction is based on FinQA attributes. With little modification, it can be applied to other financial reasoning datasets.

4.2 Hyper-parameters to LLM (i.e., Llama) and MLP classifier

To generate KGs based on the above schema, we use open-source LLM, namely Llama 3.1 8B Instruct (Llama). We chose Llama because it is freely available and requires only approximately 16 GB GPU RAM. Furthermore, we use the following hyper-parameters for Llama: (i) Model: Llama 3.1 8B Instruct, (ii) Temperature: 0.2, (iii) Maximum Number of Tokens: 2048.

For MLP classifier, we used two layers and semantic features include question embedding, triplet embedding, and cosine similarity (question embedding, triplet embedding).

Binary cross-entropy was used as a loss function.

4.3 Evaluation Setup and Baselines

Similar to (Liu et al. 2025) and (Qian et al. 2025), we use execution accuracy as a performance metric. Furthermore, we follow the standard way similar to (Zhu, Wang, and Wang 2025) and choose Gemini 2.5 Pro as a Judge to evaluate. Gemini 2.5 Pro (gemini-2.5-pro) as judge, evaluates the semantic equivalence with temperature 0.0. It accounts for the format differences (e.g., 20% = 0.20), minor rounding (e.g., ±1%), unit variations (e.g., $1.2M = $1,200,000), and semantic equivalence (e.g., 20% increase = grew by 20%). We chose Gemini as a judge because it was freely available unlike GPT-4o which have been considered by some of the recent works in this direction (Qian et al. 2025).

To compare the performance by the proposed framework, we use the vanilla Llama itself. We could not directly compare with the baselines in (Qian et al. 2025) as they used proprietary version of GPT-4o for judgement and thus not comparable with our work.

5 Result and Analysis

Table 1 presents the main results for our framework using KG with Llama model. Using KGs yields a +6.41 percentage-point absolute improvement (51.93% to 58.34%), which corresponds to approximately a 12.3% relative improvement in execution accuracy. This result confirms the capability of leveraging structured information for the financial reasoning.

Method	Acc. (%)
Llama (vanilla)	51.93%
Llama + KG	58.34%

Table 1: Execution Accuracy by Llama and Llama + KG: For judging the predictions Gemini 2.5 Pro has been used in both of the settings.

5.1 Why KG Structure Helps in Numerical Reasoning?

We now analyze the limitations of LLMs using only the text inputs using the following three aspects:

- **Temporal Disambiguation:** Text inputs often contain multiple dates, and a semantic retrieval model based purely on similarity may fail to distinguish between queries such as "2020 revenue" and "2020 expenses." In contrast, a KG-based approach explicitly encodes structured attributes such as `period="2020"` and `financial_metric_type="REVENUE"`, therefore enabling precise filtering and accurate retrieval.

- **Numerical Precision:** In text-only inputs, extracting precise numbers from long documents is difficult; KGs preserve hierarchical and contextual information and can improve numerical accuracy.

- **Multi-hop Requirements:** Many related facts are usually separated by various paragraphs in financial texts. Using a KG-based solution, triplets with the same entity_type groups related triplets and improves multi-hop retrieval and reasoning.

6 Conclusion

This paper studies the effects of augmenting knowledge graphs for numerical reasoning in financial dataset. This work is motivated by the past successes of the Retrieval Augmented Generation (RAG) using structured information such as knowledge graphs. Further, we notice that there is limited prior work incorporating the inherent and natural relational structure of the financial texts. This paper at first proposes an end-to-end pipeline that can be adapted using any large language model along with harnessing the structural properties of the texts. The proposed framework exploits a predefined KG schema, which can be easily updated for various types of financial datasets.

With systematic experiments and suitable baseline we found that using structural information (e.g., KG) of the financial text improves the prediction and shows a better reasoning capability.

7 Limitations

This work has some limitations at present, which are discussed below:

- **Dataset:** We considered only a single standard benchmark dataset in financial reasoning, namely FinQA. However, there are some more available benchmarks and we intend to explore our proposed framework over these datasets in future.

- **LLM:** Due to resource constraints and proprietary solutions, we could not use multiple recently proposed LLMs. We further intend to consider more open-source LLMs and verify the effectiveness of structural characteristics in financial reasoning task.

- **KG Schema:** Although the proposed KG schema can be easily updated, we find that multiple datasets have variance in terms of many entity types and relationship types. This may be a bottleneck when the proposed schema is used for some of the localized financial datasets.

- **Baseline:** We noticed that many of the previous studies on numerical reasoning with financial dataset consider proprietary LLMs. As our research is limited to using the open-source LLMs, a majority of the past works could not be compared. Further, we were not able to easily adapt the available works to have a comparison, as their implementations were not open due to the usage of proprietary LLMs such as GPT-4o for estimating execution accuracy.

References

Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.-H.; Routledge, B. R.; and Wang, W. Y. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of EMNLP*, 3697–3711.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Roziere, B.; Biron, B.; Tang, B.; Chern, B.; Caucheteux, C.; Nayak, C.; Bi, C.; Marra, C.; McConnell, C.; Keller, C.; Touret, C.; Wu, C.; Wong, C.; Ferrer, C. C.; Nikolaidis, C.; Allonsius, D.; Song, D.; Pintz, D.; Livshits, D.; Wyatt, D.; Esiobu, D.; Choudhary, D.; Mahajan, D.; Garcia-Olano, D.; Perino, D.; Hupkes, D.; Lakomkin, E.; AlBadawy, E.; Lobanova, E.; Dinan, E.; Smith, E. M.; Radenovic, F.; Guzmán, F.; Zhang, F.; Synnaeve, G.; Lee, G.; Anderson, G. L.; Thattai, G.; Nail, G.; Mialon, G.; Pang, G.; Cucurell, G.; Nguyen, H.; Korevaar, H.; Xu, H.; Touvron, H.; Zarov, I.; Ibarra, I. A.; Kloumann, I.; Misra, I.; Evtimov, I.; Zhang, J.; Copet, J.; Lee, J.; Geffert, J.; Vranes, J.; Park, J.; Mahadeokar, J.; Shah, J.; van der Linde, J.; Billock, J.; Hong, J.; Lee, J.; Fu, J.; Chi, J.; Huang, J.; Liu, J.; Wang, J.; Yu, J.; Bitton, J.; Spisak, J.; Park, J.; Rocca, J.; Johnstun, J.; Saxe, J.; Jia, J.; Alwala, K. V.; Prasad, K.; Upasani, K.; Plawiak, K.; Li, K.; Heafield, K.; Stone, K.; El-Arini, K.; Iyer, K.; Malik, K.; Chiu, K.; Bhalla, K.; Lakhotia, K.; Rantala-Yeary, L.; van der Maaten, L.; Chen, L.; Tan, L.; Jenkins, L.; Martin, L.; Madaan, L.; Malo, L.; Blecher, L.; Landzaat, L.; de Oliveira, L.; Muzzi, M.; Pasupuleti, M.; Singh, M.; Paluri, M.; Kardas, M.; Tsimpoukelli, M.; Oldham, M.; Rita, M.; Pavlova, M.; Kambadur, M.; Lewis, M.; Si, M.; Singh, M. K.; Hassan, M.; Goyal, N.; Torabi, N.; Bashlykov, N.; Bogoychev, N.; Chatterji, N.; Zhang, N.; Duchenne, O.; Çelebi, O.; Alrassy, P.; Zhang, P.; Li, P.; Vasic, P.; Weng, P.; Bhargava, P.; Dubal, P.; Krishnan, P.; Koura, P. S.; Xu, P.; He, Q.; Dong, Q.; Srinivasan, R.; Ganapathy, R.; Calderer, R.; Cabral, R. S.; Stojnic, R.; Raileanu, R.; Maheswari, R.; Girdhar, R.; Patel, R.; Sauvestre, R.; Polidoro, R.; Sumbaly, R.; Taylor, R.; Silva, R.; Hou, R.; Wang, R.; Hosseini, S.; Chennabasappa, S.; Singh, S.; Bell, S.; Kim, S. S.; Edunov, S.; Nie, S.; Narang, S.; Raparthy, S.; Shen, S.; Wan, S.; Bhosale, S.; Zhang, S.; Vandenhende, S.; Batra, S.; Whitman, S.; Sootla, S.; Collot, S.; Gururangan, S.; Borodinsky, S.; Herman, T.; Fowler, T.; Sheasha, T.; Georgiou, T.; Scialom, T.; Speckbacher, T.; Mihaylov, T.; Xiao, T.; Karn, U.; Goswami, V.; Gupta, V.; Ramanathan, V.; Kerkez, V.; Gonguet, V.; Do, V.; Vogeti, V.; Albiero, V.; Petrovic, V.; Chu, W.; Xiong, W.; Fu, W.; Meers, W.; Martinet, X.; Wang, X.; Wang, X.; Tan, X. E.; Xia, X.; Xie, X.; Jia, X.; Wang, X.; Goldschlag, Y.; Gaur, Y.; Babaei, Y.; Wen, Y.; Song, Y.; Zhang, Y.; Li, Y.; Mao, Y.; Coudert, Z. D.; Yan, Z.; Chen, Z.; Papakipos, Z.; Singh, A.; Srivastava, A.; Jain, A.; Kelsey, A.; Shajnfeld, A.; Gangidi, A.; Victoria, A.; Goldstand, A.; Menon, A.; Sharma, A.; Boesenberg, A.; Baevski, A.; Feinstein, A.; Kallet, A.; Sangani, A.; Teo, A.; Yunus, A.; Lupu, A.; Alvarado, A.; Caples, A.; Gu, A.; Ho, A.; Poulton, A.; Ryan, A.; Ramchandani, A.; Dong, A.; Franco, A.; Goyal, A.; Saraf, A.; Chowdhury, A.; Gabriel, A.; Bharambe, A.; Eisenman, A.; Yazdan, A.; James, B.; Maurer, B.; Leonhardi, B.; Huang, B.; Loyd, B.; Paola, B. D.; Paranjape, B.; Liu, B.; Wu, B.; Ni, B.; Hancock, B.; Wasti, B.; Spence, B.; Stojkovic, B.; Gamido, B.; Montalvo, B.; Parker, C.; Burton, C.; Mejia, C.; Liu, C.; Wang, C.; Kim, C.; Zhou, C.; Hu, C.; Chu, C.-H.; Cai, C.; Tindal, C.; Feichtenhofer, C.; Gao, C.; Civin, D.; Beaty, D.; Kreymer, D.; Li, D.; Adkins, D.; Xu, D.; Testuggine, D.; David, D.; Parikh, D.; Liskovich, D.; Foss, D.; Wang, D.; Le, D.; Holland, D.; Dowling, E.; Jamil, E.; Montgomery, E.; Presani, E.; Hahn, E.; Wood, E.; Le, E.-T.; Brinkman, E.; Arcaute, E.; Dunbar, E.; Smothers, E.; Sun, F.; Kreuk, F.; Tian, F.; Kokkinos, F.; Ozgenel, F.; Caggioni, F.; Kanayet, F.; Seide, F.; Florez, G. M.; Schwarz, G.; Badeer, G.; Swee, G.; Halpern, G.; Herman, G.; Sizov, G.; Guangyi; Zhang; Lakshminarayanan, G.; Inan, H.; Shojanazeri, H.; Zou, H.; Wang, H.; Zha, H.; Habeeb, H.; Rudolph, H.; Suk, H.; Aspegren, H.; Goldman, H.; Zhan, H.; Damlaj, I.; Molybog, I.; Tufanov, I.; Leontiadis, I.; Veliche, I.-E.; Gat, I.; Weissman, J.; Geboski, J.; Kohli, J.; Lam, J.; Asher, J.; Gaya, J.-B.; Marcus, J.; Tang, J.; Chan, J.; Zhen, J.; Reizenstein, J.; Teboul, J.; Zhong, J.; Jin, J.; Yang, J.; Cummings, J.; Carvill, J.; Shepard, J.; McPhie, J.; Torres, J.; Ginsburg, J.; Wang, J.; Wu, K.; U, K. H.; Saxena, K.; Khandelwal, K.; Zand, K.; Matosich, K.; Veeraraghavan, K.; Michelena, K.; Li, K.; Jagadeesh, K.; Huang, K.; Chawla, K.; Huang, K.; Chen, L.; Garg, L.; A, L.; Silva, L.; Bell, L.; Zhang, L.; Guo, L.; Yu, L.; Moshkovich, L.; Wehrstedt, L.; Khabsa, M.; Avalani, M.; Bhatt, M.; Mankus, M.; Hasson, M.; Lennie, M.; Reso, M.; Groshev, M.; Naumov, M.; Lathi, M.; Keneally, M.; Liu, M.; Seltzer, M. L.; Valko, M.; Restrepo, M.; Patel, M.; Vyatskov, M.; Samvelyan, M.; Clark, M.; Macey, M.; Wang, M.; Hermoso, M. J.; Metanat, M.; Rastegari, M.; Bansal, M.; Santhanam, N.; Parks, N.; White, N.; Bawa, N.; Singhal, N.; Egebo, N.; Usunier, N.; Mehta, N.; Laptev, N. P.; Dong, N.; Cheng, N.; Chernoguz, O.; Hart, O.; Salpekar, O.; Kalinli, O.; Kent, P.; Parekh, P.; Saab, P.; Balaji, P.; Rittner, P.; Bontrager, P.; Roux, P.; Dollar, P.; Zvyagina, P.; Ratanchandani, P.; Yuvraj, P.; Liang, Q.; Alao, R.; Rodriguez, R.; Ayub, R.; Murthy, R.; Nayani, R.; Mitra, R.; Parthasarathy, R.; Li, R.; Hogan, R.; Battey, R.; Wang, R.; Howes, R.; Rinott, R.; Mehta, S.; Siby, S.; Bondu, S. J.; Datta, S.; Chugh, S.; Hunt, S.; Dhillon, S.; Sidorov, S.; Pan, S.; Mahajan, S.; Verma, S.; Yamamoto, S.; Ramaswamy, S.; Lindsay, S.; Lindsay, S.; Feng, S.; Lin, S.; Zha, S. C.; Patil, S.; Shankar, S.; Zhang, S.; Zhang, S.; Wang, S.; Agarwal, S.; Sajuyigbe, S.; Chintala, S.; Max, S.; Chen, S.; Kehoe, S.; Satterfield, S.; Govindaprasad, S.; Gupta, S.; Deng, S.; Cho, S.; Virk, S.; Subramanian, S.; Choudhury, S.; Goldman, S.; Remez, T.; Glaser, T.; Best, T.; Koehler, T.; Robinson, T.; Li, T.; Zhang, T.; Matthews, T.; Chou, T.; Shaked, T.; Vontimitta, V.; Ajayi, V.; Montanez, V.; Mohan, V.; Kumar, V. S.; Mangla, V.; Ionescu, V.; Poenaru, V.; Mihailescu, V. T.; Ivanov, V.; Li, W.; Wang, W.; Jiang, W.; Bouaziz, W.; Constable, W.; Tang, X.; Wu, X.; Wang, X.; Wu, X.; Gao, X.; Kleinman, Y.; Chen, Y.; Hu, Y.; Jia, Y.; Qi, Y.; Li, Y.; Zhang, Y.; Zhang, Y.; Adi, Y.; Nam, Y.; Yu; Wang; Zhao, Y.; Hao, Y.; Qian, Y.; Li, Y.; He, Y.; Rait, Z.; DeVito, Z.; Rosnbrick, Z.; Wen, Z.; Yang, Z.; Zhao, Z.; and Ma, Z. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-Augmented Generation for

Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474.

Li, M.; Miao, S.; and Li, P. 2025. SIMPLE IS EFFECTIVE: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.20724*.

Li, X.; Zhu, Y.; Liu, S.; Ju, J.; Qu, Y.; and Cheng, G. 2023. DyRRen: A Dynamic Retriever-Reranker-Generator Model for Numerical Reasoning over Tabular and Textual Data. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*.

Liu, Z.; Guo, X.; Lou, F.; Zeng, L.; Niu, J.; Wang, Z.; Xu, J.; Cai, W.; Yang, Z.; Zhao, X.; et al. 2025. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning. *arXiv preprint arXiv:2503.16252*.

Nie, Y.; Kong, Y.; Dong, X.; Mulvey, J. M.; Poor, H. V.; Wen, Q.; and Zohren, S. 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. arXiv:2406.11903.

Qian, L.; Wang, Y.; Peng, X.; Zhou, W.; Han, Y.; Zhao, Y.; Huang, J.; Xie, Q.; and Nie, J.-Y. 2025. Fino1: On the Transferability of Reasoning-Enhanced LLMs and Reinforcement Learning to Finance. *arXiv preprint arXiv:2502.08127*.

Sarmah, B.; Hall, B.; Rao, R.; Patel, S.; Pasquali, S.; and Mehta, D. 2024. HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction. *arXiv preprint arXiv:2408.04948*.

Srivastava, P.; Malik, M.; Gupta, V.; Ganu, T.; and Roth, D. 2024. Evaluating LLMs' Mathematical Reasoning in Financial Document Question Answering. In *Findings of ACL*, 3853–3878.

Sun, J.; Zhang, H.; Lin, C.; Su, X.; Gong, Y.; and Guo, J. 2024. APOLLO: An Optimized Training Approach for Long-form Numerical Reasoning. *arXiv preprint arXiv:2212.07249*.

Sun, L.; Wang, X.; and Li, Y. 2024. Pyramid-Driven Alignment: Pyramid Principle Guided Integration of Large Language Models and Knowledge Graphs. arXiv:2410.12298.

Tan, X.; Wang, H.; Qiu, X.; Cheng, Y.; Xu, Y.; Chu, W.; and Qi, Y. 2024. Struct-X: Enhancing Large Language Models Reasoning with Structured Data. arXiv:2407.12522.

Wu, X.; and Tsioutsiouliklis, K. 2024. Thinking with Knowledge Graphs: Enhancing LLM Reasoning Through Structured Data. arXiv:2412.10654.

Zhu, J.; Chen, Q.; Dou, H.; Li, J.; Guo, L.; Chen, F.; and Zhang, C. 2025. DianJin-R1: Evaluating and Enhancing Financial Reasoning in Large Language Models. arXiv:2504.15716.

Zhu, L.; Wang, X.; and Wang, X. 2025. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. arXiv:2310.17631.