FinAgentBench: A Benchmark Dataset for Agentic Retrieval in Financial Question Answering

Anonymous Author(s)
Affiliation
Address
email

Abstract

information retrieval (IR) is critical in the financial domain, where investors must identify relevant information from large collections of documents. Traditional IR methods—whether sparse or dense—often fall short in retrieval accuracy, as it requires not only capturing semantic similarity but also performing fine-grained reasoning over document structure and domain-specific knowledge. Recent advances in large language models (LLMs) have opened up new opportunities for *retrieval with multi-step reasoning*, where the model ranks passages through iterative reasoning about which information is most relevant to a given query. However, there exists no benchmark to evaluate such capabilities in the financial domain. To address this gap, we introduce FINAGENTBENCH, the first large-scale benchmark for evaluating retrieval with multi-step reasoning in finance – a setting we term *agentic retrieval*. The benchmark consists of 3,429 expert-annotated examples on S&P-100 listed firms and assesses whether LLM agents can (1) identify the most relevant document type among candidates, and (2) pinpoint the key passage within the selected document. Our evaluation framework explicitly separates these two reasoning steps to address context limitations. This design enables to provide a quantitative basis for understanding retrieval-centric LLM behavior in finance. We evaluate a suite of state-of-the-art models and further demonstrated how targeted fine-tuning can significantly improve agentic retrieval performance.

1 Introduction

Information Retrieval (IR) is a foundational research field that studies how to effectively search for and retrieve relevant information from large-scale text collections [21, 23]. Its practical importance in real-world applications has made it one of the central and long-standing areas in computer science since the early days of computing [15]. IR has evolved from sparse, term-frequency–based methods [24], which rely on exact keyword matches, to dense neural retrieval models that embed text into continuous latent spaces to capture deeper semantic [10]. In finance, accurate retrieval is critical, as investors depend on precise access to vast filings and reports to make high-stakes, time-sensitive decisions. To support this need, finance-specific IR benchmarks have been developed [8, 6], enabling rigorous evaluation of both sparse and dense retrievers in this complex, data-rich domain.

However, recent studies reveal persistent accuracy ceilings for both sparse and dense retrieval methods [12, 2], especially in domains that demand fine-grained understanding and structured reasoning over complex documents. To overcome these limitations, recent research has turned to large language models (LLMs), which bring strong language understanding and the ability to process long contexts [7, 13, 29]. One promising direction is *generative retrieval*, where an LLM generates the index of the most relevant document given a query and document collection [16]. These results show impressive gains in performance, indicating a shift toward retrieval systems that embed deeper reasoning capabilities [1, 18].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

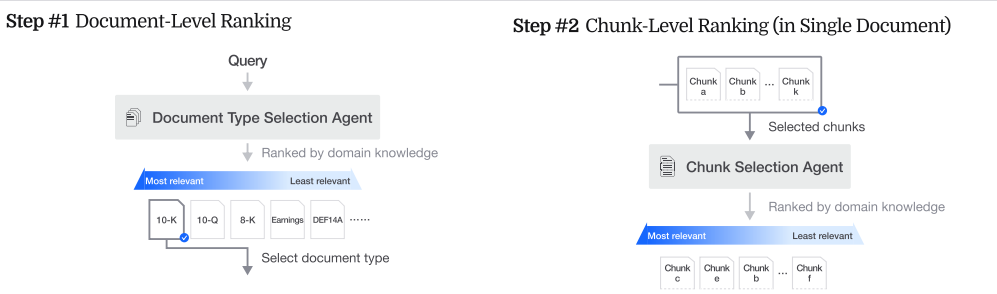

Figure 1: Agentic retrieval pipeline in FINAGENTBENCH. The process consists of two stages. (*Top*) Given a natural-language query, the **Document Type Selection Agent** ranks five SEC filing types (10-K, 10-Q, 8-K, earnings transcripts, and DEF-14A) and selects the most relevant type. (*Bottom*) The corresponding document is segmented into paragraph-level chunks, from which the **Chunk Selection Agent** identifies the top-k passages.

In the financial domain, this shift is particularly relevant. Financial documents are typically long and dense, often requiring multiple steps of reasoning: first to identify which document type (e.g., 10-K, earnings transcript) best matches the information need, and then to locate the specific evidence within it. As such, simple one-shot retrieval is often insufficient—accurate performance hinges on multi-step reasoning that reflects how professionals actually search for information. Despite the practical importance of such multi-step reasoning, no benchmark currently exists to evaluate LLMs in this setting. This leaves open critical questions about whether LLMs can serve as effective retrieval agents in high-stakes domains like finance—where precision, explainability, and structured navigation are essential.

To address this gap, we introduce the first large-scale benchmark, FINAGENTBENCH, for evaluating generative retrieval systems in the financial domain, focusing on a setting we term *agentic retrieval*. Unlike prior benchmarks that assess retrieval in a single stage, our proposed benchmark evaluates the ability of LLM agents to reason and retrieve the relevant information through a two-stage pipeline: (1) identifying which document to retrieve, evaluated via document-level ranking, and (2) determining which part of the document to focus on, evaluated via chunk-level ranking. The benchmark contains 3,429 samples and is constructed from real-world financial documents paired with expert-written queries and annotations, reflecting authentic use cases faced by professional investors. By capturing both retrieval accuracy and reasoning depth, it provides a foundation for systematically analyzing the strengths and limitations of LLM-based generative retrieval in high-stakes domains like finance.

The main contribution of this work is three-fold:

- We propose FINAGENTBENCH, to the best of our knowledge the first large-scale benchmark for evaluating agentic retrieval in finance, featuring 3,429 expert-annotated samples across document- and chunk-level ranking tasks.
- We evaluate a range of state-of-the-art LLMs on our benchmark, revealing their performance in accuracy and reasoning when applied to real-world financial retrieval scenarios.
- We further investigate the impact of fine-tuning the LLMs on agentic retrieval tasks, demonstrating that targeted supervision can substantially improve both document selection and chunk-level reasoning performance.

2 FINAGENTBENCH Dataset

In this section, we describe the concept of FINAGENTBENCH, a large-scale benchmark for evaluating agentic retrieval in finance. To retrieve accurate information, financial retrieval requires multi-step reasoning due to both the volume of data and the regularity of financial disclosures [5]. When leveraging the reasoning capabilities of large language models to improve accuracy in retrieval, the extensive length and redundancy of financial documents—where even a single 10-K can exceed hundreds of pages—pose a significant challenge, making it inefficient to process all content without any filtering or prioritization [19]. Therefore, to maintain efficiency, a system should first select the document type most likely to contain the answer—feasible information because filings follow predictable conventions, with different information consistently organized by document type (e.g.,

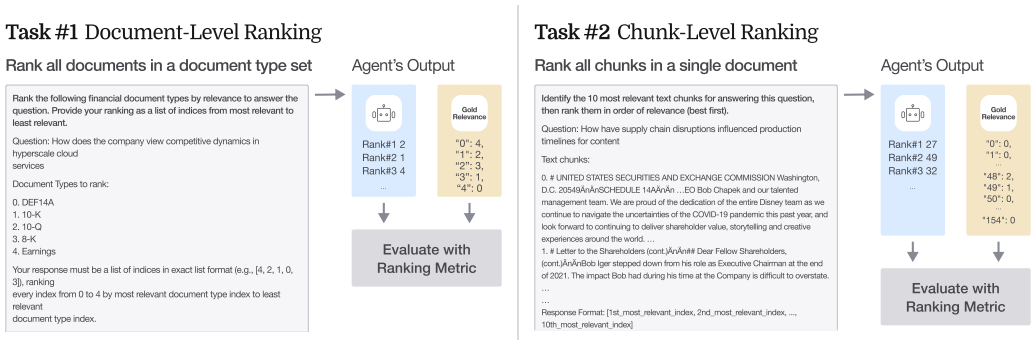

Figure 2: Examples of the two retrieval tasks in FINAGENTBENCH. The benchmark comprises: (*Left*) Document-level ranking, where the agent ranks five SEC document types based on their relevance to the input query. All document types are ordered by relevance, and the full ranking is used as the gold label during evaluation. (*Right*) Chunk-level ranking, where the agent selects and orders the top-k most relevant passages from the selected document. Each chunk is annotated with a ground-truth relevance label—0 (irrelevant), 1 (partially relevant), or 2 (directly relevant)—and predictions are evaluated using MRR, MAP, and nDCG.

Table 1: Evaluation of reasoning LLMs on the *Document Ranking* and *Chunk Ranking* tasks.

	Document Ranking			Chunk Ranking		
Model	**nDCG@5**	**MAP@5**	**MRR@5**	**nDCG@5**	**MAP@5**	**MRR@5**
GPT-o3	0.770	0.829	0.875	0.351	0.257	0.538
Claude-Opus-4	0.773	0.840	0.875	0.418	**0.307**	**0.568**
Claude-Sonnet-4	**0.783**	**0.849**	**0.892**	**0.419**	0.296	0.567

risk factors in 10-Ks, strategic commentary in earnings calls). It should then identify the relevant chunk or passage within the selected document.

This motivates a two-stage retrieval process—document selection followed by passage selection—which we term *agentic retrieval*, as it reflects the sequential reasoning steps taken by experts. Figure 1 provides an overview of the *agentic retrieval* workflow we benchmark. At test time, the system is provided with a natural-language query q, typically issued by a professional investor about a specific firm, along with a document collection $\mathcal{D}$ containing over 35K U.S. corporate disclosures spanning from 2010 to 2024. In FINAGENTBENCH, we include 10-K, 10-Q, 8-K, earnings call transcripts, and DEF 14A proxy statements, some of the most commonly used public filings that are crucial for financial retrieval. The retrieval task follows a two-stage reasoning pipeline:

Stage 1: Document-Level Ranking. Rather than searching the entire corpus, the agent first identifies the document type most likely to contain the answer. Given the type set

$$\mathcal{T} = \{\text{10-K}, \text{10-Q}, \text{8-K}, \text{Earnings}, \text{DEF14A}\} \tag{1}$$

, the model produces a ranking over $\mathcal{T}$. This stage evaluates the model's understanding of finance-specific reporting conventions—for instance, risk factors typically appear in 10-Ks, while shareholder proposals are found in DEF-14A filings.

Stage 2: Chunk-Level Ranking. The selected document $(d_{t^\star} \in \mathcal{D})$ is split into non-overlapping passages $\mathcal{C}(d_{t^\star}) = \{c_1, \ldots, c_M\}$. The agent ranks these chunks and returns the top k passages indexed as $\langle c_{(1)}, \ldots, c_{(k)} \rangle$.

Annotations. Each query in FINAGENTBENCH is generated and annotated by domain experts. At the document-type level, a gold label t^{G} in $\mathcal{T}$ is provided with an gold ranking . At the chunk level, every chunk element $\mathcal{C}_{(i)}$ in $\mathcal{C}(d_{t^\star})$ is annotated as $\mathcal{C}^{\text{G}}_{(i)}$ with a gold relevance score. For annotating those relevance scores, we followed TREC Eval [17]: 0 (irrelevant), 1 (partially relevant), and 2 (directly relevant).

Table 2: Impact of reinforcement fine-tuning on `GPT-o4-mini` across both retrieval tasks.

	w/o fine-tuning			w/ fine-tuning		
Task	nDCG@5	MAP@5	MRR@5	nDCG@5	MAP@5	MRR@5
Document Ranking	0.758	0.826	0.872	**0.808**	**0.865**	**0.933**
Chunk Ranking	0.345	0.256	0.526	**0.371**	**0.274**	**0.587**

3 Experiments

We evaluate a range of reasoning-capable LLMs on FINAGENTBENCH to assess their ability to perform agentic retrieval in finance. Our experiments focus on two subtasks: document-type ranking and chunk-level passage selection. We also study the effect of domain-specific fine-tuning on retrieval performance.

3.1 Experimental Setup

We benchmark three commercial LLMs—`GPT-o3`, `Claude-Opus-4`, and `Claude-Sonnet-4`—using zero-shot prompting. For each query, models are prompted to complete two tasks: (1) rank document types from a set of five SEC filing categories, and (2) rank paragraph-level chunks from the selected document. We split both the document-type ranking and chunk-level relevance tasks into training and evaluation sets using an 80/20 split. To assess the impact of domain adaptation, we additionally evaluate `GPT-o4-mini` before and after reinforcement fine-tuning provided by OpenAI [1] on held-out randomly sampled 10% of training splits from FINAGENTBENCH.

Performance is measured using standard ranking metrics measure by top-5 results: normalized Discounted Cumulative Gain (nDCG), Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). For chunk-level evaluation, we report metrics over the top ranked passages, evaluated against expert-annotated graded relevance scores. All models operate in a retrieval-only setting without access to external tools or retrieval augmentation.

3.2 Results

Task Performance. Table 1 compares reasoning-capable LLMs on document- and chunk-level ranking tasks. At the document level, all models perform substantially above random, indicating that LLMs possess strong priors over financial reporting structure. The best-performing model reaches near-perfect retrieval, while smaller gaps across models suggest that architectural or scale differences still matter for capturing subtle type-level cues. In contrast, chunk-level ranking proves more challenging: overall scores are lower, reflecting the complexity of fine-grained retrieval within long disclosures. Here, the two Claude variants perform comparably, with one slightly stronger in overall ranking quality and the other leading in precision-oriented metrics. This highlights that while LLMs can reliably handle document-type discrimination, chunk-level reasoning remains a demanding setting that exposes model differences more sharply.

Impact of Fine-Tuning. Table 2 evaluates the impact of reinforcement fine-tuning on `GPT-o4-mini` across both retrieval stages. Fine-tuning yields substantial gains: for document-type ranking, nDCG improves from 0.758 to 0.808, and MRR increases from 0.872 to 0.933. Similar improvements are observed in chunk ranking, with MRR rising from 0.526 to 0.587. These results demonstrate that domain-specific supervision significantly enhances retrieval accuracy, both in selecting the correct document type and in pinpointing relevant information within the document. This highlights the importance of adapting LLMs to financial reasoning tasks through task-aligned training signals.

4 Conclusion

We introduced FINAGENTBENCH, a large-scale benchmark designed to evaluate agentic retrieval capabilities of large language models in the high-stakes domain of finance. The benchmark simulates realistic investor queries over a diverse set of corporate filings, requiring models to reason over both document types and intra-document content.

FINAGENTBENCH provides a new foundation for studying end-to-end retrieval behaviors in complex domains. Future work may explore enhancing agentic retrieval performance and joint modeling of retrieval and generation for investment decision support.

[1] OpenAI Reinforcement Fine-Tuning

References

[1] Hongru Cai, Yongqi Li, Ruifeng Yuan, Wenjie Wang, Zhen Zhang, Wenjie Li, and Tat-Seng Chua. Exploring training and inference scaling laws in generative retrieval. *arXiv preprint arXiv:2503.18941*, 2025.

[2] Brian J Chan, Chao-Ting Chen, Jui-Hung Cheng, and Hen-Hsen Huang. Don't do rag: When cache-augmented generation is all you need for knowledge tasks. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 893–897, 2025.

[3] Sanchit Chanana, Hyung Won Chung, Nan Du, Jeffrey Zhao, et al. Don't do RAG: When cache-augmented generation is all you need. *arXiv preprint arXiv:2412.15605*, 2024.

[4] Zhiyu Chen, Wenhu Chen, Sameena Shah, et al. Finqa: A dataset of numerical reasoning over financial data. In *EMNLP*, 2021.

[5] Jaeyoung Choe, Jihoon Kim, and Woohwan Jung. Hierarchical retrieval with evidence curation for open-domain financial question answering on standardized documents. *arXiv preprint arXiv:2505.20368*, 2025.

[6] Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy yong Sohn, and Alejandro Lopez-Lira. Finder: Financial dataset for question answering and evaluating retrieval-augmented generation. *arXiv preprint arXiv:2504.15800*, 2025.

[7] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

[8] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

[9] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

[10] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

[11] X Li, J Jin, Y Zhou, Y Zhang, P Zhang, Y Zhu, and Z Dou. From matching to generation: A survey on generative information retrieval (2024). *URL https://api. semanticscholar. org/CorpusID*, 269303210.

[12] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Robust neural information retrieval: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2407.06992*, 2024.

[13] Kun Luo, Minghao Qin, Zheng Liu, Shitao Xiao, Jun Zhao, and Kang Liu. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment. *arXiv preprint arXiv:2408.12194*, 2024.

[14] Maurício Sousa Maia, Alberto H. F. Laender, P. Gnther, and Maria da Graça Campos Pimentel. Fiqa: A collection for aspect-based opinion mining and question answering. In *ECIR*, 2018.

[15] Mandar Mitra and BB Chaudhuri. Information retrieval from documents: A survey. *Information retrieval*, 2:141–163, 2000.

[16] Thong Nguyen and Andrew Yates. Generative retrieval as dense retrieval. *arXiv preprint arXiv:2306.11397*, 2023.

[17] Joao Palotti, Harrisen Scells, and Guido Zuccon. Trectools: an open-source python library for information retrieval practitioners involved in trec-like campaigns. SIGIR'19. ACM, 2019.

[18] Ronak Pradeep, Kai Hui, Jai Gupta, Adam D Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Q Tran. How does generative retrieval scale to millions of passages? *arXiv preprint arXiv:2305.11841*, 2023.

[19] Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. Docfinqa: A long-context financial reasoning dataset. *arXiv preprint arXiv:2401.06915*, 2024.

[20] Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR*, pages 232–241, 1994.

[21] Gerard Salton. *Automatic information organization and retrieval.* McGraw Hill Text, 1968.

[22] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 11(5):513–523, 1975.

[23] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[24] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[25] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2024.

[26] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023.

[27] Xiaoju Ye, Zhichun Wang, and Jingyuan Wang. Infinite retrieval: Attention-enhanced llms in long-context processing. *arXiv preprint arXiv:2502.12962*, 2025.

[28] Lei Zhu, Fandong Meng, Zhen Wu, Haiyang Song, and Jie Zhou. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *ACL*, 2021.

[29] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

A Related Work

Retrieval Systems. IR systems has progressed from early sparse methods like TF–IDF and BM25 [22, 20], which rely on exact lexical overlap, to dense dual encoders that map queries and documents into a shared semantic space. Notable advances such as Dense Passage Retrieval (DPR) [9] and E5-Mistral [25] have improved performance on open-domain QA by capturing deeper semantics. Despite these gains, both sparse and dense retrievers struggle with multi-hop reasoning and long-context queries—highlighting inherent limitations in fixed retrieval pipelines.

To address this, *generative retrieval* reframes the task as sequence generation, enabling models to directly produce relevant document identifiers or content [11]. Recent approaches such as CAG [3], and Infinite Retrieval [27] leverage LLMs' language modeling and attention capabilities by caching the documents to bypass traditional indexing. In parallel, agentic reasoning frameworks like ReAct [26] incorporate tool use and step-by-step reasoning, suggesting the potential of the LLMs to iteratively decide what to retrieve and why—bridging retrieval with planning and decision-making.

Finance-Retrieval Benchmarks. Retrieval in the financial domain presents unique challenges: documents are long, queries are complex, and precision is critical. Existing datasets such as FinQA [4], TAT-QA [28], and FiQA [14] address specific financial reasoning needs, while more recent corpora like FinanceBench [8] and FinDER [6] support open-domain and retrieval-augmented generation tasks. However, these benchmarks still assume retrieval as a fixed subroutine, typically relying on

vector search or keyword search—and do not evaluate whether LLMs can reason about what to retrieve or where to look within long documents.

Our proposed FINAGENTBENCH fills this gap by directly evaluating agentic retrieval in finance. It assesses whether an LLM agent can (1) generate the correct document identifier and (2) rank the most relevant chunks within that document, providing a rigorous testbed for end-to-end reasoning-driven retrieval in high-stakes settings.