Automated Financial Extraction from 10-K Filings: A Comparative Analysis of Open-Source LLMs in Hybrid RAG Systems

Rohit Shyamnarayan Singh
Tagliatela College of Engineering
University of New Haven
West Haven, Connecticut, United States of America
rsing12@unh.newhaven.edu

Abstract

This study evaluates three open-source large language models Llama-3-8B, Mistral-7B, and Phi-3-Mini within a standardized Retrieval-Augmented Generation (RAG) framework for financial question answering. The objective is to determine their suitability for deployment in resource-constrained and latency-sensitive environments where factual accuracy and response reliability are essential. A curated set of domain-specific reasoning and retrieval queries is used to assess performance across accuracy, hallucination rates, and generation latency. Results indicate that Llama-3-8B provides the most effective overall trade-off, achieving the highest tied accuracy (60%) alongside the lowest latency (5.86s). Mistral-7B attains comparable accuracy but incurs substantially higher latency approximately four times slower which limits its practicality for real-time use. Phi-3-Mini exhibits a zero-hallucination profile but also shows the lowest accuracy (52%) and the highest latency (55.53s), making it less suitable for time-critical applications despite its safety advantages. These findings illustrate meaningful performance differentials among current open-source models and emphasize the importance of jointly evaluating accuracy, reliability, and computational efficiency when selecting LLMs for production-grade RAG systems.

1 Introduction

The rapid proliferation of Large Language Models (LLMs) has intensified the need for systematic benchmarking to understand their capabilities and limitations across retrieval, reasoning, and knowledge-intensive tasks. Although proprietary models frequently define the upper bounds of performance, the growing maturity of open-source alternatives has made them indispensable in settings where data privacy, computational cost, deployment control, and reproducibility are critical concerns (Zhao et al., 2023; Kaddour et al., 2023). As more organizations investigate on-premises and hybrid inference pipelines, the evaluation of lightweight, resource-efficient open models has emerged as a central research priority.

Open-source LLMs such as Llama-3, Mistral, and Phi-3 reflect a broader movement toward accessible and adaptable foundation models, supported by research showing the effectiveness of model compression, distillation, and quantization for enabling high-quality reasoning on constrained hardware (Frantar et al., 2023; Dettmers et al., 2022). Concurrently, advances in Retrieval-Augmented Generation (RAG) have demonstrated the importance of grounding parametric knowledge with external evidence to mitigate hallucinations and improve factual reliability (Lewis et al., 2020; Gao et al., 2023). Given ongoing studies showing that LLMs often struggle with long-context reasoning and may fail to utilize retrieved evidence effectively (Liu et al., 2023; Xie et al., 2024), evaluating open models under identical retrieval conditions has become an essential comparative methodology.

Despite the expanding literature on model evaluation, relatively few studies directly compare open-source LLMs within a controlled RAG pipeline while simultaneously accounting for

inference latency and hallucination risk—two factors that strongly influence real-world usability. These considerations are especially relevant in high-stakes domains such as finance, healthcare, and policy analysis, where inaccurate or fabricated content may lead to harmful downstream decisions (Ji et al., 2023). Latency, in particular, serves as a practical proxy for computational efficiency, illuminating whether a model may be reasonably deployed on commodity hardware or under limited compute budgets. Likewise, hallucination frequency has been increasingly foregrounded as a core reliability metric, motivated by recent findings that smaller models do not always hallucinate less and may require explicit grounding mechanisms for stable performance (Bang et al., 2023; Huang et al., 2023).

In response to these gaps, this study conducts a controlled comparison of three widely adopted open-source models—Llama-3-8B, Mistral-7B, and Phi-3-Mini—under a unified retrieval-augmented evaluation setup. The analysis centers on three dimensions that collectively capture operational feasibility and reliability: factual accuracy and semantic relevance of generated outputs, end-to-end generation latency as a measure of computational practicality, and hallucination rate as an indicator of grounded reasoning quality. By situating these findings within the broader literature on retrieval augmentation, hallucination mitigation, and efficient LLM deployment, the results offer insight into how current-generation lightweight models perform under realistic constraints and where meaningful performance divergences emerge.

2 Literature Review

The landscape of Large Language Models has shifted significantly from a "bigger is better" paradigm toward the optimization of smaller, more efficient architectures. Early foundational work by Kaplan et al. established scaling laws suggesting that performance correlates strictly with parameter count and training data volume. However, the release of Meta's Llama series challenged this by demonstrating that smaller models trained on significantly more tokens could outperform larger, undertrained counterparts (Touvron et al.).

The Rise of Efficient 7B-8B Models: The 7-8 billion parameter range has emerged as a "sweet spot" for local inference. Mistral AI introduced novel techniques such as Sliding Window Attention, allowing their 7B model to outperform significantly larger models like Llama-2 13B on standard benchmarks. Similarly, Meta's Llama 3 (8B) utilizes a denser tokenizer and extensive training on 15 trillion tokens to achieve state-of-the-art performance for its size class, focusing on improved reasoning and code generation capabilities.

Small Language Models (SLMs) and Data Quality: Concurrently, research into "Small Language Models" (SLMs) has focused on data quality over quantity. Microsoft's Phi series (Gunasekar et al.) posits that "textbooks are all you need," proving that highly curated, synthetic "textbook-quality" data can train sub-3B parameter models (like Phi-3 Mini) to rival much larger models in reasoning tasks. This study contributes to this evolving discourse by empirically testing these theoretical efficiency claims in a practical, resource-constrained inference setup.

3 Datasets

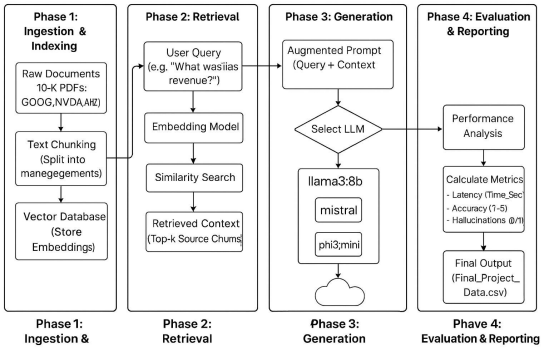
The dataset consists of Fiscal Year 2024 Form 10-K filings obtained from the U.S. Securities and Exchange Commission (SEC) EDGAR database. These filings represent the primary annual financial disclosure documents submitted by publicly traded companies and provide a comprehensive account of financial performance, business operations, risk factors, market conditions, and regulatory considerations.

For this study, filings from five major technology companies were selected: Amazon.com, Inc. (AMZN), Alphabet Inc. (GOOGL), Intel Corporation (INTC), Meta Platforms, Inc. (META) NVIDIA Corporation (NVDA)

These companies were chosen due to their scale, diverse business models, and the complexity of their financial disclosures, which collectively offer a challenging environment for retrieval-augmented question answering. Each 10-K document ranges between several hundred pages and includes dense textual structure such as management discussion and analysis (MD&A), financial statements, footnotes, and regulatory disclosures.

All filings were downloaded in PDF format directly from the EDGAR system and processed through the ingestion pipeline described in Section 3.2. The dataset serves as the exclusive knowledge base for retrieval operations, ensuring that all answers generated by the RAG system are grounded in verifiable, publicly available financial documents.

4 Methodology



4.1 System Overview

The Retrieval-Augmented Generation (RAG) system is designed to support grounded question answering over financial filings, specifically annual 10-K reports from major technology companies. The system follows a modular pipeline consisting of four stages: (1) data ingestion and indexing, (2) context retrieval, (3) answer generation using multiple open-source LLMs, and (4) performance evaluation. This architecture enables controlled comparisons across generative models while holding all retrieval components constant.

4.2 Data Ingestion and Preprocessing

The corpus consists of Fiscal Year 2024 10-K filings from five large technology firms (e.g., Alphabet, Amazon, Nvidia). All documents are imported from the SEC EDGAR database as PDF files and processed using an automated ingestion pipeline. Text is extracted from each PDF and segmented into coherent units using a sentence-window chunking strategy implemented via the Sentence Window Node Parser. Each chunk is expanded to include its surrounding context (±3 sentences), a method shown to improve retrieval quality for structurally dense financial reports.

The resulting text segments ("nodes") are stored along with metadata—including document source and windowed context—to support later attribution and traceability.

4.3 Index Construction

Two complementary indexes are constructed to support hybrid retrieval.

Dense Vector Index.

Each text chunk is embedded using the BAAI bge-m3 embedding model, a state-of-the-art encoder designed for multilingual and domain-agnostic retrieval tasks. The embedding persisted within a ChromaDB vector store, enabling cosine-similarity search over millions of high-dimensional vectors. Sparse BM25 Index. To complement embedding-based retrieval, a sparse lexical index is created using BM25. This component improves performance on keyword-dominant financial queries, particularly those involving entity names, figures, or accounting terminology. All node objects are serialized and stored for fast BM25 reconstruction during inference.

The hybrid retrieval configuration follows recent recommendations in RAG literature showing that combining sparse and dense methods mitigates failure modes associated with either approach alone.

4.4 Query Embedding and Retrieval

At inference time, user queries are embedded using the same BGE-M3 encoder to maintain embedding-space alignment. The query vector is matched against the vector index using similarity search to identify the top-k most relevant text chunks. In parallel, the BM25 retriever computes lexical relevance scores. Both outputs are merged using a hybrid ranking strategy, and the top-k combined results are forwarded as the retrieval context.

This process yields a compact set of grounded evidence sourced directly from the 10-K filings, ensuring that all answers remain auditable and traceable to specific document segments.

4.5 Prompt Construction and Generation

The retrieved context is concatenated with the user query to form an augmented prompt. This prompt is passed to one of three open-source LLMs Llama-3-8B, Mistral-7B, or Phi-3-Mini selected at runtime. Each model is executed locally through Ollama, ensuring consistent inference conditions and complete control over system resources.

All models receive identical prompts and retrieval contexts, enabling direct, apples-to-apples comparison of their generative behavior. No model-specific prompt tuning is applied, ensuring that observed performance differences reflect inherent model capabilities rather than prompt engineering artifacts.

4.6 Evaluation Procedure

The system is evaluated on a curated set of ten domain-specific financial queries, designed to test factual recall, numerical reasoning, multi-sentence synthesis, and procedural understanding of 10-K filings. The questions include, for example:

1. What was Nvidia's total revenue for the fiscal year 2024?
2. What was the net sales revenue specifically for Amazon Web Services (AWS) in 2024?
3. Did Intel's total revenue increase or decrease in 2024 compared to 2023, and by how much?
4. What was the operating loss reported for Meta's Reality Labs segment in 2024?
5. What specific risk factors does Alphabet (Google) list regarding Generative AI and Large Language Models?
6. Who does Nvidia identify as its main competitors in the Data Center market?
7. How much did Meta spend on Research and Development (R&D) in 2024?
8. What factors does Amazon list as affecting its shipping and fulfillment costs?
9. What are the key antitrust legal proceedings mentioned by Alphabet (Google) in the 10-K?
10. Describe Intel's 'IDM 2.0' strategy and the risks associated with it.

For each model output, three quantitative metrics are recorded:

Latency (seconds): end-to-end generation time.

Accuracy: rated on a 1–5 scale capturing factual correctness and alignment with ground truth.

Hallucination Rate: binary indicator of unsupported or fabricated content.

All outputs, metadata, and timing measurements are preserved in a structured dataset (Final_Project_Data.csv), enabling reproducibility and downstream analysis.

5 Results

Model	Avg Accuracy(1-5)	Avg Latency(s)	Total Hallucinations	Sample Size	Accuracy %
llama3:8b	3	5.86	1	10	60
mistral	3	23.599	1	10	60
phi3:mini	2.6	55.526	0	10	52

Table 1

Table 1 summarizes the comparative performance of the three evaluated models—Llama-3-8B, Mistral-7B, and Phi-3-Mini—across accuracy, latency, and hallucination frequency. All models were evaluated on the same set of ten domain-specific financial queries derived from the 2024 10-K filings.

Across the three LLMs, Llama-3-8B and Mistral-7B achieved the highest accuracy, each scoring an average rating of 3.0 on a five-point scale, corresponding to 60% accuracy. In contrast, Phi-3-Mini obtained an average accuracy of 2.6, equivalent to 52%, indicating greater difficulty in producing fully correct or contextually precise responses. These results suggest that smaller parameter counts, while advantageous for efficiency, may limit semantic fidelity in complex financial contexts.

Latency measurements reveal substantial variation in computational efficiency. Llama-3-8B produced the fastest responses, with an average end-to-end generation time of 5.86 seconds, demonstrating suitability for interactive or near–real-time applications. Mistral-7B exhibited significantly higher latency, averaging 23.60 seconds, approximately four times slower than Llama-3-8B despite equivalent accuracy. Phi-3-Mini displayed the highest latency, averaging 55.53 seconds,

indicating that model size alone does not guarantee computational efficiency when running on constrained hardware.

Hallucination behavior further differentiates the models. Phi-3-Mini produced zero hallucinations, demonstrating strong reliability when grounded context is provided. Both Llama-3-8B and Mistral-7B generated one hallucinated response each within the ten-question evaluation. The combination of low hallucination rate and reduced accuracy in Phi-3-Mini indicates a pattern of conservative but incomplete answers, whereas the two larger models occasionally attempted more detailed explanations that increased the risk of unsupported claims.

Overall, the results highlight a clear trade-off among the three models. Llama-3-8B delivers the best balance of accuracy and latency, making it the most practical choice for real-time financial questions answering under a RAG pipeline. Mistral-7B matches Llama-3-8B in correctness but is hindered by substantially slower inference speed. Phi-3-Mini demonstrates strong safety characteristics but lags in both accuracy and responsiveness, limiting its suitability for time-sensitive or analytically demanding scenarios.

6 Analysis

6.1 Accuracy–Latency Trade-offs:

The results reveal a clear divergence in the efficiency profiles of the evaluated models despite comparable accuracy scores. Both Llama-3-8B and Mistral-7B achieved the highest average accuracy rating of 3.0 (60%). However, the latency gap between them is substantial: Llama-3-8B exhibited an average response time of 5.86 seconds, whereas Mistral-7B required 23.60 seconds, representing a nearly fourfold increase without any corresponding improvement in accuracy. This discrepancy suggests that Llama-3-8B offers a more favorable balance between computational efficiency and reasoning capability, particularly for the types of fact-based and numerically oriented queries present in the financial domain.

Phi-3-Mini performed notably worse on both metrics, recording an average accuracy of 2.6 (52%) and the slowest average latency of 55.53 seconds. The combination of lower correctness and substantially higher inference time indicates that Phi-3-Mini is poorly suited for real-time or interactive settings. Its latency profile suggests that it may only be practical in batch-processing environments where throughput, rather than responsiveness, is the primary concern.

6.2 Hallucination Behavior and Safety Considerations:

A distinct pattern emerges when examining hallucination frequency. Phi-3-Mini produced zero hallucinations across all evaluated queries, in contrast to the single hallucinated response generated by both Llama-3-8B and Mistral-7B. This behavior indicates a conservative generative strategy: Phi-3-Mini appears more inclined to avoid explicit claims when contextual evidence is insufficient, reducing the likelihood of fabricating unsupported information. Although this strategy contributes to lower overall accuracy by producing underspecified or incomplete responses, it enhances the model's reliability profile in contexts where misinformation poses a significant risk.

By comparison, the larger models demonstrated a higher tendency to produce detailed answers, which likely contributed to their isolated hallucination events. This pattern aligns with prior findings in LLM safety research, where models with greater expressive capacity sometimes generate confident but incorrect statements when faced with ambiguous or weakly supported prompts.

Overall, the analysis underscores an important trade-off among correctness, responsiveness, and safety. Llama-3-8B provides the best overall balance, Mistral-7B trades speed for equivalent accuracy, and Phi-3-Mini sacrifices accuracy and efficiency in favor of more conservative, low-hallucination behavior. These tendencies should inform model selection in domains requiring strict factual grounding versus those prioritizing speed or depth of reasoning.

7 Conclusion

Our comparative evaluation yields the following insights:

Llama 3:8B emerges as the preferred choice for general-purpose applications, striking the best balance between speed and accuracy. It

demonstrates superior efficiency while maintaining top-tier accuracy within the tested cohort.

Mistral delivers strong accuracy, but its higher computational cost and latency make it less practical than Llama 3 for this specific workload.

Phi 3: Mini is most suitable for offline, safety-critical tasks where minimizing hallucinations is more important than response speed or peak accuracy.
Future work should focus on expanding both the size and diversity of the dataset to validate these conclusions across a wider range of domains.

7. References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadallah, H., ... & Zhou, X. (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., ... & Yuan, L. (2023). Textbooks are all you need.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models.

Meta AI. (2024). The Llama 3 Herd of Models.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models.