SMU Data Science Review

Volume 9
Number 3 *Fall 2025*

Article 2

2025

Application of Open-Source Small Large Language Models for Finance report analysis

Tue Vu
SMU, vuminhtue@gmail.com

Mark Austin
AT&T, mdaustin@smu.edu

Marcel Tuijn
SMU, mtuijn@smu.edu

Application of Open-Source Small Large Language Models for Finance report analysis

Tue Vu[1], Mark Austin[2], Marcel Tujin[3]
[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX, USA
[2] VP of Data Science, AT&T, Plano, TX, USA
[3] Cox School of Business, Southern Methodist University, Dallas, TX, USA

{tuev, mdaustin, mtujin}@smu.edu

Abstract. The rapid integration of generative AI in finance introduces both opportunities and challenges, particularly when analyzing sensitive data such as Securities and Exchange Commission (SEC) filings. This study investigates the use of open-source Small Large Language Models (SLLMs), deployed locally through the Ollama and LangChain frameworks, combined with Retrieval-Augmented Generation (RAG) for extracting financial insights relevant to index performance and reporting quality. Two key objectives guide this work: (1) benchmarking multiple open-source SLLMs for sentiment analysis, multiple-choice reasoning, and financial question answering, and (2) assessing the feasibility of locally deployed SLLMs for domain-specific financial queries. A standardized set of 50 finance-related questions was applied to each SEC filing through a RAG pipeline. Six SLLMs: Gemma3, Phi4, Llama3.1, Gpt-oss, Qwen3, and DeepSeek-r1 (10b–20b parameters) were evaluated on Financial PhraseBank, FinanceBench, and the finance subset of MMLU. Performance metrics included accuracy, BLEU, and ROUGE score to measure factuality, fluency, and reasoning quality. Results reveal trade-offs between accuracy, computational efficiency, and token throughput, highlighting model-specific strengths. The findings provide a reproducible, privacy-preserving framework for deploying locally run AI systems in financial analytics on university HPC infrastructure.

1 Introduction

Recent advancement in Large Language Models (LLMs) has shown great promise of working with complex textual data accurately at large scale. These models are particularly good at three key tasks: processing large amounts of information from many different sources, grasping subtle meanings in long documents, and handling complex analytical problems that were previously too challenging for automated systems. This progress is especially important for finance and economic field, where work often requires both deep domain knowledge of the field and careful analysis of complicated and sensitive datasets. These combined requirements make financial tasks particularly difficult. As a result, many researchers have started using LLMs to handle these challenging financial problems more effectively.

Financial markets run on the timely interpretation of corporate disclosures. Each year, U.S. public companies publish thousands of SEC filings such as 8-K (current report), MD&A (Management's Discussion & Analysis), Conference Call (often called Earning/Investor call), and Risk Disclosure that contain unstructured, dense, heterogeneous narratives data which financial analysts read, analyze and convert into structured intelligence for price action. In our analysis, we evaluate every report for every company for every filing with a questionnaire containing 50-question instruments asking, for example: *Q1: Are they planning to raise product prices in the future? (1 = yes, 0 = no). (Only if Q1 = 1) How confident are you in your analysis of Q1 (1-10)? Provide the sentence(s) that led you to your conclusion for Q1. Please provide the sentence or up to three sentences that strongly support your conclusion.* These kinds of questions consisting of different Natural Language Processing (NLP)'s application such as sentiment analysis, multiple choices with the level of uncertainty and summarization technique. Performing this task at corpus scale for tenth of thousands of reports is tedious and error-prone, and beyond what traditional NLP (e.g., keyword search or regex) can reliably accomplish: answers must be grounded in context, sensitive to domain language, and accompanied by verbatim evidence. This research focuses on the underexplored intersection of finance and LLMs. Compared to mainstream NLP benchmarks, few deployed studies demonstrate LLMs as end-to-end engines for evidence-backed extraction from SEC filings across large corpora. Exploring this junction opens a research path for LLM methods tailored to regulatory text and investor workflows, where correctness, provenance, and repeatability are non-negotiable. While proprietary LLMs (e.g., ChatGPT or Perplexity) can follow detailed prompts to accomplish the task on individual documents, they face practical limitations at production scale: high token costs, daily throughput constraints, limited automation hooks, significant operator time and data privacy.

Related literature points to two converging advances. First, automation with instruction prompts following LLMs has widened the frontier for information extraction, classification, and reasoning. Second, RAG grounds model outputs in retrieved passages, reducing hallucinations and constraining context windows. Building on these ideas, our work operationalizes an open-source stack Ollama for local LLMs serving and LangChain for orchestration and running with on-premises HPC equipped with A100 GPUs to enable batch automation at scale. Utilizing SMU SuperPOD's GPU resources let us move beyond a single model toward ensembles of open-source SLLMs, echoing classical machine-learning practice to reduce bias and overfitting while improving robustness.

Despite progress, several gaps remain: (1) Manual review is infeasible when thousands of filings each require 50 answers; (2) Proprietary LLM APIs are cost-prohibitive at corpus scale; (3) Even with budget, rate limits and long runtimes hinder daily refreshes; (4) The questionnaire is heterogeneous, binary judgments, graded certainty, and quoted evidence, so ensemble/routing strategies are beneficial but traditionally expensive and operationally complex with closed models; (5) The issue of data privacy or dealing with protected data.

This research addresses these gaps with a RAG-centric, open-source pipeline purpose-built for SEC analysis. Our system uses LLMs to answer the standardized questionnaire instrument per report, with answers grounded in retrieved text and exported as structured CSV suitable for downstream analytics. To ensure domain

adequacy, we benchmark multiple LLMs in the finance domain using multiple datasets (e.g., *FinanceBench*, *Financial Phrasebank*, and finance-focused subsets of *Massive Multitask Language Understanding - MMLU*), and we evaluate with task-appropriate metrics: *accuracy* for categorical items and *BLEU/ROUGE* score for short, quoted spans - reflecting the instrument's qualitative and quantitative nature. The pipeline emphasizes throughput and cost control via chunking, reranking, and minimal-context prompts; automation via batch job orchestration on HPC; and quality via optional model ensembles and question-specific prompt templates.

In brief, this work contributes:

(1) A finance-focused benchmarking protocol that spans across 3 different datasets with 6 SLLMs and evaluation metrics to assess the reasoning, accuracy, and efficiency of open-source language models;

(2) A scalable RAG workflow that transforms unstructured SEC filings into reproducible, evidence-based responses using a standardized 50-question framework; and

(3) An operational blueprint for cost-efficient, high-throughput, and locally hosted SLLMs inference using the Ollama and LangChain frameworks on HPC infrastructure. By grounding generation in retrieved context, leveraging open-source language models at the 10b-20b parameter scale, and automating large-batch execution through GPU acceleration, this work aims to make large-corpus, evidence-driven financial analysis both practical and affordable, while fostering continued research at the intersection of finance and language modeling.

2 Literature Review

2.1 LLM in Finance domain study

The rapid evolution of LLMs, such as GPT, BERT, and their finance-specific variants, has significantly transformed the landscape of financial technology and analytics. LLMs are designed to process and generate human-like text, leveraging vast datasets and advanced deep learning architectures, particularly the Transformer framework. Their ability to understand context, perform zero-shot and few-shot learning, and adapt to domain-specific tasks has made them highly attractive for a range of financial applications, from sentiment analysis and risk assessment to automated report generation and customer service [1], [2]. The finance sector, characterized by complex terminology, high-stakes decision-making, and the need for real-time insights, presents unique challenges that LLMs are increasingly well-equipped to address.

One of the most prominent applications of LLMs in finance is sentiment analysis, where models like FinBERT and BloombergGPT have demonstrated superior performance over traditional machine learning and dictionary-based approaches [3], [4]. By leveraging contextual understanding, LLMs can accurately classify the sentiment of financial news, analyst reports, and social media posts, providing valuable signals for market prediction and investment strategies [5]. Recent studies show that

LLMs not only outperform earlier models in sentiment classification but also enable the development of profitable trading strategies, with generative LLMs and advanced prompting techniques further enhancing predictive accuracy and portfolio performance [6].

LLMs are increasingly useful for extracting and synthesizing information from complex financial documents, such as earnings reports, regulatory filings, and ESG disclosures [1]. Their ability to integrate narrative and quantitative data enables more nuanced risk detection, fraud prevention, and financial statement analysis. Domain-specific LLMs, fine-tuned on financial corpora, have shown promise in handling specialized terminology and providing deeper insights for auditors, analysts, and investors [7]. Additionally, LLMs facilitate automated report generation, question answering, and real-time decision support, streamlining operations and enhancing the efficiency of financial institutions [8]. The adaptability of LLMs allows them to perform multiple tasks such as summarization, keyword extraction, and planning without the need for extensive retraining, making them valuable tools for both large organizations and smaller firms with limited resources [2].

2.2 Benchmarking LLM in Finance domain

In the study by Busch et al. (2024) [9], the authors introduced the first benchmark for evaluating Large Language Models (LLMs) in Business Process Management (BPM). The authors assess seven models, both open- and closed-source, including GPT-4, Llama3, Phi-3, and Mixtral on four representative BPM tasks: activity recommendation, robotic process automation candidate identification, process question answering, and mining declarative process models. Using tailored datasets and metrics such as cosine similarity, F1-score, and ROUGE-L, the study reveals that while GPT-4 is a stable performer, smaller open-source models like Llama3-7B and Mixtral-8x7B achieve comparable or even superior results in some tasks. Findings highlight that model choice and task alignment are more critical than sheer model size, and that domain-specific benchmarks are essential for guiding organizations in selecting efficient, cost-effective LLMs for BPM applications.

Another study by Lu et al. (2025) [10] also introduces BizFinBench, the first large-scale benchmark tailored to real-world financial tasks, designed to evaluate the reliability of large language models (LLMs) in precision-critical domains. The benchmark comprises 6,781 annotated queries across five dimensions: "*numerical calculation, reasoning, information extraction, prediction recognition, and knowledge-based QA*", organized into nine categories such as anomalous event attribution and stock price prediction. To reduce evaluation bias, the authors propose *IteraJudge*, an iterative LLM-based judging framework. Testing 25 proprietary and open-source models, results show no model excels universally: closed-source systems like ChatGPT-o3 and Gemini-2.0-Flash lead in reasoning and tool usage, while open-source DeepSeek-r1 performs strongly in numerical tasks and entity recognition. However, models still struggle with complex, cross-concept reasoning, underscoring the need for domain-specific benchmarks. BizFinBench thus provides a rigorous foundation for advancing trustworthy financial AI.

2.3 Ability of Language models in extracting information using RAG

The RAG model has been applied in Wang et al (2025) [4] to propose an intelligent financial data analysis system that integrates LLMs with RAG to overcome the limitations of traditional statistical and rule-based methods in handling complex, unstructured financial data. The system features three modules: data preprocessing for financial standardization, vector-based storage and retrieval, and RAG-enhanced query processing. Using the NASDAQ financial fundamentals dataset (2010–2023), experiments compared baseline and optimized configurations. Results show that the gpt-3.5-turbo-1106+RAG model achieved 78.6% accuracy and 89.2% recall, significantly outperforming non-RAG baselines while reducing response time by 34.8%. Findings confirm that RAG integration enhances LLMs' domain-specific reasoning, providing a more reliable tool for financial analysts and decision-making

Soudani et al. (2024) [12] compares fine-tuning (FT) and RAG for improving language models' performance on less popular or long-tail factual knowledge. The authors evaluate twelve models across multiple datasets, analyzing factors such as fine-tuning methods, data augmentation, model size, and retriever quality. Results show that while FT enhances accuracy, it is resource-intensive and less effective for large models. RAG consistently outperforms FT, especially on low-frequency entities, and combining FT with RAG benefits smaller models. Retrieval-augmented and instruction-tuned LLMs have also been introduced to address challenges such as limited context in financial news, resulting in significant gains in accuracy and F1 scores for sentiment analysis tasks [13].

2.3 Challenges of LLM in Finance domain

Despite their transformative potential, the deployment of LLMs in finance is not without challenges. Key concerns include data privacy, model interpretability, domain adaptation, and the risk of algorithmic bias [1]. The complexity of financial language and the high stakes of financial decision-making demand rigorous evaluation, explainability, and robust governance frameworks. There is also a need for standardized financial corpora, improved explainability tools, and strategies to ensure data quality and security [7]. As research progresses, integrating LLMs with quantitative models and expert systems, as well as developing domain-specific adaptations, will be crucial for maximizing their impact while mitigating risks [1]. Overall, LLMs are poised to drive innovation in financial analysis, risk management, and customer engagement, but their responsible and effective adoption requires ongoing research and careful consideration of ethical and practical challenges.

3 Models, Data and Methodology

3.1 Data

In order to evaluate the performance and domain adaptability of SLLMs within the financial sector, we selected multiple open-source datasets that capture distinct linguistic and reasoning characteristics of financial discourse. Specifically, we draw upon Financial PhraseBank, FinanceBench and finance-related subsets of the MMLU benchmark. These datasets were selected to ensure a balanced representation of financial sentiment, reasoning, and instruction-following capabilities across a variety of sub-tasks relevant to industry applications such as market analysis, financial reporting, and investment decision support. The use of publicly available datasets not only promotes transparency and reproducibility but also facilitates consistent cross-model comparisons across open and proprietary LLMs.

FinanceBench

FinanceBench is a novel benchmark created to evaluate the ability of language models to perform financial question-answering (QA) in an "open-book" style, using external documents rather than relying solely on internal model knowledge [14]. The publicly released version [15] includes 150 fully annotated examples, drawn from a larger set of 10,231 questions about publicly traded companies, each paired with human-annotated answers, justifications and evidence strings. The questions are designed to be ecologically valid and cover a wide array of financial-document use cases from metrics-based queries (e.g., "What was the company's EPS in year X?") to domain-relevant reasoning questions (e.g., "How did change in debt/EBITDA affect the firm's credit risk?") and novel-generated items (i.e., crafted for evaluation rather than derived from existing corpora). More importantly, the benchmark not only provides the QA pairs but also links each question to the underlying source document via metadata that captures the document name (ticker, period, type), the evidence excerpt, and the page number. This structure enables RAG evaluation workflows, where one might first retrieve the relevant document or passage and then use a model to generate or select the answer from it [14]. By incorporating FinanceBench into a SLLM evaluation pipeline, researchers can assess whether compact, on-premises models can handle financial reasoning and document-grounded QA with sufficient accuracy, contextual grounding and factual consistency.

The sample json format of this dataset:


```
{
"financebench_id": "financebench_id_01198",
     "company": "AMD",
     "question": "What drove revenue change as of the FY22 for AMD?",
     "context": "Net revenue for 2022 was $23.6 billion, an increase of
44% compared to 2021 net revenue of $16.4 billion. The increase in net
revenue was driven by a 64% increase in Data Center segment revenue
```

```
primarily due to higher sales of our EPYC server processors, a 21% increase
in Gaming segment revenue primarily due to higher semi-custom product
sales, and a significant increase in Embedded segment revenue from the
prior year period driven by the inclusion of Xilinx embedded product
sales.",
     "expected_output": "In 2022, AMD reported Higher sales of their EPYC
server processors, higher semi-custom product sales, and the inclusion of
Xilinx embedded product sales",
}
```


Financial PhraseBank

Financial PhraseBank [16] is one of the most widely adopted corpora for financial sentiment analysis. It consists of annotated sentences extracted from financial news articles, categorized as positive, negative, or neutral based on their sentiment toward company or market performance. The annotations were provided by financial experts to ensure high domain reliability. This dataset has total 2264 QA pairs downloaded from HuggingFace data: *https://huggingface.co/datasets/takala/financial_phrasebank* and is instrumental for benchmarking models on tasks involving sentiment classification, tone detection, and fine-grained emotion recognition in financial communications - capabilities that are essential for applications such as automated trading sentiment systems and earnings call analysis.

The sample json format for Financial PhraseBank is below:


```
{
     "index": 995,
     "text": "Metsa-Botnia will finance the payment of dividends, the
repayment of capital and the repurchase of its own shares with the funds
deriving from its divestment of the Uruguay operations, and shares in
Pohjolan Voima, and by utilising its existing financing facilities",
     "gold": "neutral",
}
```


In which the “gold” is the targeted or label for the corresponding “text” in that index

MMLU – Finance focus

The MMLU dataset [17] offers a broad test of world knowledge and reasoning abilities across 57 subjects, including finance-related domains such as economics, business ethics, and accounting. For this study, we extract and analyze a total of 8 finance-specific subjects from *Econometrics, Business Ethics, Highschool Macro/Microeconomic, Management, Marketing, Professional Accounting and Public Relation* (total 1571 QA pairs) from MMLU to assess model proficiency in financial reasoning and numerical comprehension. By incorporating MMLU’s standardized multiple-choice format, this subset provides an objective measure of factual and conceptual understanding that complements the open-ended reasoning tasks found in FinanceBench and Financial PhraseBank. MMLU data was downloaded from hugginface data: *https://huggingface.co/datasets/takala/financial_phrasebank*

The sample json format for MMLU question is below:


```
  {
    "benchmark": "MMLU",
```

```
    "subject": "macroeconomics",
    "topic": "monetary_policy",
    "difficulty": "college",
    "question": "If a central bank raises its policy interest rate, which
short-run effect is most likely on aggregate demand (AD)?",
    "choices": {
      "A": "AD increases because investment rises",
      "B": "AD increases because net exports rise",
      "C": "AD decreases because investment falls",
      "D": "AD is unchanged because monetary policy is neutral in the
short run"
    },
    "correct_answer": "C"
  }
```


Collectively, these datasets offer a comprehensive foundation for assessing SLLMs in financial applications, covering key dimensions of language understanding: instruction adherence (MMLU), sentiment interpretation (Financial PhraseBank), and analytical reasoning (FinanceBench). Their open-source nature ensures that results can be replicated, audited, and extended by future research efforts, reinforcing the methodological rigor and openness of this benchmarking study.

3.2 Small-sized Large Language Models (SLLMs)

Recent advances in LLMs have given rise to a class of more parameter-efficient models in the range of 10–20 billion-parameter that can run on more modest hardware, including local and on-premises environments. These models often described as *Small-sized LLMs (SLLMs)* - strike a balance between performance and computational cost: they are large enough to handle non-trivial reasoning, generation, and instruction-following tasks, yet small enough to be deployed on high-end workstations or domestic servers. The following paragraphs introduce six notable families in this category: Gemma, Phi, Llama, Gpt-oss, DeepSeek, and Qwen.

3.2.1 Gemma3:12b (Google)

The Gemma family, introduced in early 2024, represents Google DeepMind's open-weight line derived from the Gemini research codebase. For example, the Gemma 2 models scale from roughly 2 billion up to 27 billion parameters [18]. The models employ architecture enhancements such as Grouped Query Attention to boost efficiency. Gemma has been cited [19] as outperforming for its size with a 9 billion-parameter variant could outperform larger peers in certain settings. In the context of SLLMs, a distilled Gemma at approx. 10 billion parameters offers a viable local-deployment option without invoking the resource demands of 30+ billion-parameter models. In this study, we use Gemma3 [20], which is announced as the most capable model that runs on single GPU. Gemma3 comes in different versions, and we choose 12-billion parameter (second to the largest 27b).

3.2.2 Phi4:14b (Microsoft)

Developed by Microsoft Research, the Phi family explicitly targets SLLMs that are efficient, accessible, and deployable. The Phi version 4, announced in early 2025 was introduced as state-of-the-art open model built upon a blend of synthetic datasets, data from filtered public domain websites, and acquired academic books and Q&A datasets, it comes in only 1 version of 14 billion parameters. Phi-4 exemplifies Microsoft's "small but smart" philosophy, showing that curated, reasoning-dense datasets can yield state-of-the-art performance without massive parameter counts [21].

3.3.3. Llama3.1:8b (Meta AI)

The Llama line from Meta AI has become a staple in the open-source model ecosystem. While the original Llama and Llama 2 systems push into tens or hundreds of billions of parameters, smaller distilled versions in the less than 10 billion parameter range (for example, Llama 3.1 8B [22]) provide strong performance for local use. In benchmarking leaderboards, Llama 3.1 8B appears among top entries for MMLU, GSM8K, and HumanEval in the small-model class [23]. Thus, a distilled Llama3.1:8b is selected for this study for benchmarking with finance data.

3.3.4. DeepSeek-r1:14b (DeepSeek AI)

Based in China, DeepSeek gained attention for its efficient training and reasoning-optimized architecture. The foundational DeepSeek-r1 model (2025) and its distillations report strong performance in reasoning, Mixture of Expert (MoE) architecture and code generation tasks [24]. The distillation pipeline produces different versions from 1.5 to 32 billion parameters tailored for local deployment. The DeepSeek-r1 version with 14b parameter is qualified as SLLMs oriented toward high-performance on constrained hardware.

3.4.5. Qwen3:14b (Alibaba)

The Qwen family from Alibaba Cloud has evolved rapidly, with multiple sizes and modalities released. The Qwen 2.5 technical report (2024) presents a comprehensive series of LLMs trained on up to 18 trillion tokens [25]. While many Qwen models are very large (e.g., 72 billion parameters and beyond), smaller derivatives are designed for local or edge deployment scenarios. Especially within the SLLM paradigm, such Qwen variants support multilingual understanding, long-context reasoning, and fine-tuning flexibility, making them competitive for on-device or on-premises use. In this study, we use Qwen3 which is the latest generation of SLLMs in Qwen series with 14b parameter, offering a comprehensive suite of dense and MoE models [26].

3.4.6. Gpt-oss:20b (OpenAI)

OpenAI's Gpt-oss (Generative PreTrained Open-Source Series) represents a landmark release in the LLM ecosystem: it is the company's first set of open-weight models since

GPT-2, made available under the Apache 2.0 license for unrestricted use, modification and deployment [27]. The family consists of two principal variants: Gpt-oss:120b and Gpt-oss:20b, both utilizing a MoE architecture that activates a subset of parameters per token to balance reasoning capability with computational efficiency [28]. The larger model is designed to run on a single high-end GPU (e.g., 80 gb H100 class) while the smaller variant can operate on more modest hardware (e.g., with 16 gb Vram), thereby enabling on-device or on-premises deployment for enterprise and research applications. According to evaluations, Gpt-oss:20B as used in this study, demonstrates competitive performance in reasoning and code-generation tasks relative to other open-source models in its size class, though trade-offs remain multilingual and long-tail domain capability [28]. Its release marks a significant step toward democratizing advanced LLM access, enabling organizations to fine-tune, host and integrate powerful models under full data control and regulatory oversight.

Collectively, these six model families represent an emerging sweet-spot in the LLM design space with the parameter counts in the range of 10–20 billion range, competitive performance on reasoning and knowledge benchmarks, and practical deployability in local environments. By focusing on distilled or optimized versions of larger architectures, they enable research groups and enterprises to host capable LLMs without cloud-scale infrastructure.

The configuration of the above 6 SLLMs are summarized in Table 1.

Table 1 Configurations of selected SLLMs

Model	Params (Billions)	Context Length (thousand)	Attention Heads	Architecture Notes
Gemma3	12	128	Hybrid local–global + GQA (5 local : 1 global)	Uses RoPE with higher base frequency on global layers; multimodal via SigLIP encoder
Phi4	14	16	Decoder-only Transformer; full attention; heads not published	Data-centric training with curated + synthetic datasets; tiktoken vocab (~100k)
Llama3.1	8	128	GQA (8 KV heads; ~32 Q heads)	RMSNorm, RoPE; multilingual
Gpt-oss	20	132	GQA (64 Q / 8 KV heads); MoE (32 experts, top-4 active ~3.6B)	Open-weight mixture-of-experts model; optimized for long context
DeepSeek-r1	14	132	MLA (Multi-Head Latent Attention)	Distilled variant; reasoning-focused; often uses Qwen backbone
Qwen3	14	32	GQA (40 Q heads / 8 KV heads; 40 layers)	Supports long-context reasoning; expandable context with YaRN

3.3 Benchmarking score

When evaluating SLLMs applied to financial text understanding, summarization, or sentiment analysis, quantitative metrics are crucial to measure both linguistic fidelity and semantic correctness. Among the most widely adopted methods are BLEU,

ROUGE, and Accuracy, each offering unique perspectives on model quality. BLEU and ROUGE primarily assess the text generation quality of LLM outputs in RAG settings, while Accuracy is essential in classification tasks such as sentiment detection using structured datasets like the Financial PhraseBank and MMLU.

3.3.1 BLEU: BiLingual Evaluation Understanding

The BLEU score introduced by Papineni et al. (2002) [29] is a precision-oriented metric that measures how closely a model's generated text matches a set of human reference texts based on overlapping *n-grams* (contiguous word sequences). It is computed as the geometric mean of modified n-gram precisions, combined with a *brevity penalty* to penalize excessively short outputs.
For example, consider a LLM summarizing a quarterly financial report:

Generated: "Revenue increased by 10% and net profit reached $25 million."
Reference: "The company reported a 10% rise in revenue and a net profit of $25 million."

The BLEU score measures the overlap of terms such as "revenue", "10%", and "net profit," yielding a high score if the generated text accurately replicates the reference content. The BLEU score has a range from 0 to 1, with 0.5 considered good score. For the above sentence, the BLEU=0.04 or 4% which is quite low, mainly because BLEU heavily penalized short sentences with limited exact n-gram overlap, even though the generated and reference has very close semantic meaning. In financial contexts, BLEU is valuable when textual precision and terminological accuracy are critical, such as summarizing SEC filings or analyst statements.

3.3.2. ROUGE: Recall Oriented Understudy for Gisting Evaluation

The ROUGE metric introduced by Lin (2004) [30] complements BLEU by focusing on *recall*, that is, how much relevant information from the reference text appears in the model output. Common variants include ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), ROUGE-L (longest common subsequence), and ROUGE-S (skip-bigram overlap).

Similar to BLEU, the ROUGE score has the range from 0-1. The above sentence would yield a ROUGE-1's F1=0.66 and ROUGE2's F1=0.33. This is higher than BLEU. ROUGE would capture that most key ideas ("operating margin improved", "debt ratio decreased") are preserved, even if wording differs. In finance-specific RAG pipelines, ROUGE ensures the model does not omit critical information from retrieved SEC filings or internal financial memos, a key requirement for compliance and risk communication.

3.3.3. Accuracy:

Beyond generation tasks, Accuracy serves as the standard performance measure for classification problems such as *sentiment analysis* or *multiple-choice* option. In the financial domain, a well-known benchmark is the Financial PhraseBank [16], a dataset of financial news phrases annotated by domain experts into three sentiment categories: *positive*, *neutral*, and *negative*. Accuracy is computed as the ratio of correctly predicted labels to the total number of examples, as shown in Equation 1:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \quad (1)$$

For example, if a LLM predicts the sentiment for 1,000 phrases and correctly classifies 880 of them, the resulting accuracy is 0.88 (88%). Similarly, this is applied to MMLU data benchmarking with multiple choices option of A, B, C, D. This measure is particularly effective for assessing a model's domain adaptation whether fine-tuned or RAG-augmented on structured financial corpora. The Financial PhraseBank thus provides a foundational testbed for evaluating how effectively LLMs interpret and classify nuanced financial language, which is crucial when models are deployed on-premises for sensitive data analysis.

Together, BLEU and ROUGE quantify textual fidelity in financial summarization tasks, while Accuracy captures semantic correctness in classification tasks such as sentiment analysis and multiple choices. When applied to LLMs operating within secure, on-prem RAG frameworks, these metrics enable a multi-dimensional evaluation of both language quality and factual reliability for deploying LLMs responsibly in finance and regulatory domains.

3.4 Methodology

Ollama

In the context of deploying compact SLLMs on-premises or protective data environments, the framework Ollama has emerged as a practical enabler. Ollama is an open-source runtime and model management system that allows users to download, run, and manage various open-source LLMs wholly on local hardware (Linux, macOS, Windows, local HPC) without reliance on remote cloud services [31]. By abstracting away much of the infrastructure complexity, model file handling, tokenization pipelines, runtime orchestration, Ollama enables researchers and engineers to focus on model usage rather than setup [32].

Crucially for the SLLM paradigm, Ollama supports rapid prototyping and iteration in local or enterprise settings. It allows for downloading pre-packaged model files (e.g., "`ollama pull llama3.1:8b`"), running them via a CLI (e.g., "`ollama run llama3.1:8b`"), and integrating them with higher-level frameworks (e.g., via a Python API) [33]. As such, it serves as a bridge between selecting a SLLM (which might be distilled for local deployment) and embedding that model within secure, RAG workflows, especially in domains like finance where data sovereignty and auditability are paramount. Moreover, the design philosophy of Ollama aligns strongly with on-premises deployment scenarios where sensitive data (for example, internal corporate documents or regulatory filings) cannot be sent to external cloud APIs [34].

Ensemble and Ground truth model

We evaluated six SLLMs across the three benchmark datasets mentioned above. However, the benchmark scores alone do not indicate whether the models perform well or not. To establish a reference point, we plan to compare these results against a more advanced model, such as ChatGPT GPT5, which will serve as our ground truth for performance evaluation. Additionally, we implemented an ensemble approach that aggregates the outputs of the six SLLMs based on majority agreement. For example, if four models predict a positive outcome, one predicts neutral, and one predicts negative, the ensemble output is classified as positive. The overall experimental setup and workflow are illustrated in the Figure 1.

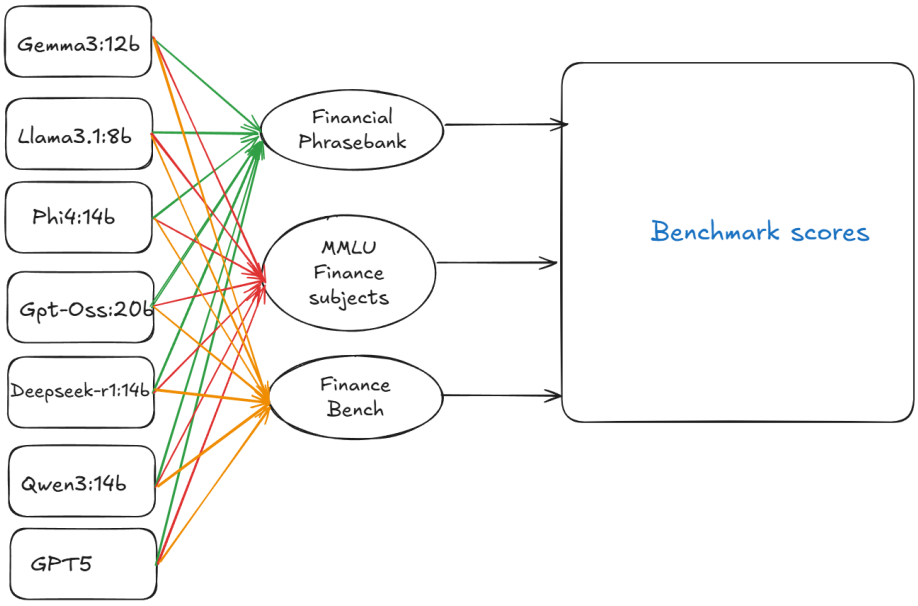

Fig. 1 Workflow of the benchmarking process

Retrieval Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) provides a robust framework designed to mitigate hallucinations and compensate for gaps in a model's internal knowledge by dynamically incorporating relevant external information during response generation, enabling the model to incorporate up-to-date, domain-relevant information at inference time rather than relying solely on its pre-trained parameters. This capability is particularly important for domains like finance, where timely and accurate information (for example, recent filings, regulatory changes, or internal corporate data) must inform model outputs. As Gao et al. (2023) note, RAG enhances generation by retrieving relevant documents from a knowledge base, thus reducing hallucinations and improving factual grounding [35].

A naïve RAG workflow typically proceeds through several stages [35]:

(1) ***Indexing***: Clean and extract raw data from diverse format like PDF, HTML, Word, Markdown, then split into smaller chunks for easier digest and preserve information. Chunks are then encoded into vector representation by embedding model and stored in vector database, which is crucial for retrieval phase in the subsequent state

(2) **Retrieval:** every input user query also transformed by the same embedding model. Retrieval steps find the similarity score between this input query and the collection of embedded chunks and return the top K-chunks which have similar semantic similarity to the query;

(3) **Augmentation**: where retrieved content in the top K-chunk is synthesized (via prompt engineering or context injection) into the input for the LLM;

(4) **Generation**, where the LLM produces a response conditioned on both the query and the retrieved context.

In this methodology section, the naïve RAG approach will be applied to a secure, on-premises deployment scenario in which a SLLM is used to respond to queries over internal financial or regulatory datasets. The retrieval component ensures that the model has access to the most relevant and recent domain-specific material, the augmentation step ensures that context is properly integrated into the prompt, and the generation step delivers the final output. By structuring our pipeline in this way, we aim to ensure both factual accuracy and domain relevance, while maintaining full data sovereignty and inference control.

Langchain framework

With the increasing adoption of SLLMs on-premises settings, the challenge shifts from merely selecting a capable model to integrating those models effectively into secure, domain-specific workflows such as RAG over proprietary data. The LangChain framework offers a strongly modular architecture for exactly this purpose. LangChain provides abstractions for model I/O, prompt templates, memory and state, external data retrieval (including vector stores and databases), tool and agent orchestration, and chaining of SLLM invocations [36]. By combining these components, organizations can build RAG pipelines that connect a SLLM to private repositories (e.g., internal documents, regulatory filings, internal databases) while retaining control of data flows and running fully on-premises - an especially important capability for handling sensitive data such as Securities and Exchange Commission (SEC) related private disclosures.

In such a deployment, a SLLM can serve as the generative component, while LangChain handles retrieval of relevant document chunks, context-window management, prompt engineering, and secure orchestration of generation and post-processing. For example, an on-premises RAG system may load internal SEC filings, embed them via a local vector store, have LangChain handle query ingestion and retrieval, invoke the SLLM for answer generation, and apply post-filters or guardrails before returning results, thus ensuring both domain alignment and data privacy. By enabling the reuse of components, swapping of underlying models, and fine-tuning of retrieval and memory strategies, LangChain makes it feasible for organizations to

deploy production-grade RAG systems without depending on external cloud models or losing control of sensitive information [37].

4 Results

4.1 Benchmarking with Financial Phrasebank data

The comparative analysis across three key performance dimensions reveals distinct trade-offs between accuracy, response latency, and computational efficiency among the evaluated models. Fig. 2 shows overall accuracy as well as average response time and throughput (average tokens per second). As illustrated in Fig.2, the benchmarked models are categorized into two architectural paradigms: general models (Gemma3:12b, Llama3.1:8b, Phi4:14b) that generate direct sentiment classifications, and reasoning-intensive models (Gpt-oss:20b, DeepSeek-r1:14b, Qwen3:14b) that employ chain-of-thought or similar mechanisms to process intermediate analytical steps. In terms of overall accuracy, Gemma3:12b and Qwen3:14b emerge as co-leaders at 0.93 and are equivalent to the "ground truth" GPT5, followed closely by Phi4:14b (0.92) and Gpt-oss:20b (0.91). In fact, the overall accuracy of all 6 SLLMs is very close. The Ensemble, which has the benefit of selecting the majority agreement between all 6 models, achieves the highest accuracy of 0.95.

The response timing metrics expose a critical inverse relationship: while Llama3.1:8b achieves the fastest response time at 0.24 seconds, reasoning models like DeepSeek-r1:14b require substantially longer (5.32s), though this slower response correlates with more comprehensive reasoning processes rather than purely computational constraints. Throughput presents a different performance hierarchy, with Gpt-oss:20b leading at average 91.60 tokens per second and Qwen3:14b at 85.33 tokens per second, while Gemma3:12b's mere 4.6 tokens per second reflects its deliberative processing style. These results suggest that practitioners must carefully balance accuracy requirements against real-time processing needs. In this dataset benchmarking, Gemma3:12b proves to be the best one with high accuracy and quick response time (with smallest throughput), but in general, all models' performance are close enough.

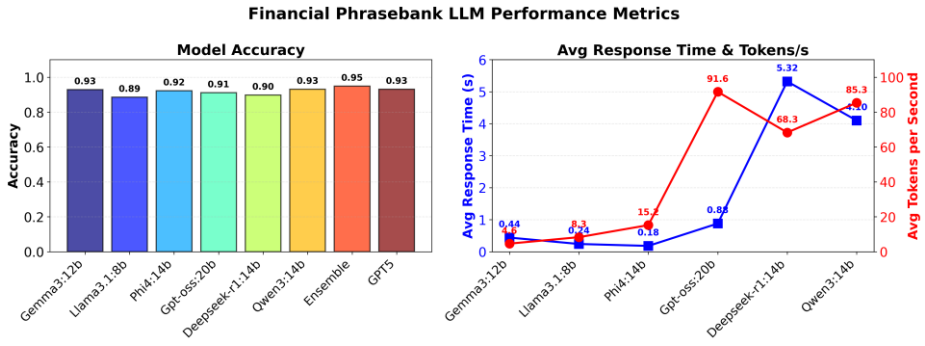

Fig. 2 Benchmarking with Financial Phrasebank data

4.2 Benchmarking with MMLU-Finance data

For the MMLU with finance focusing data, the 6 SLLMs were benchmarked with 1,571 test samples spanning eight business and economics subjects. MMLU dataset consisting of multiple choice with answers in A, B, C, D. The models demonstrated in Fig. 3 a substantial variation in macro accuracy scores ranging from 0.64 to 0.85, with Gpt-oss:20b achieving the highest accuracy at 0.85, followed by Qwen3:14b at 0.83, while Phi4:14b and Gemma3:12b both achieved 0.74. Llama3.1:8b has the lowest accuracy at 0.64. A critical finding reveals an inverse relationship between accuracy and computational efficiency: Llama3.1:8b has lowest accuracy, it also has near lowest response time to answer and throughput speed; in the meantime, reasoning intensive model like Qwen3:14b has substantially good accuracy of 0.83 but also has much longer time to response and high throughput. In term of accuracy, not a single SLLMs accuracy is close to ground truth GPT5 at 0.92. However, the Ensemble, which measure the majority vote among 6 SLLMs has very close accuracy to GPT5 at 0.9, it suggests the advantage of Ensemble method which is able to boost up the ranking of all models.

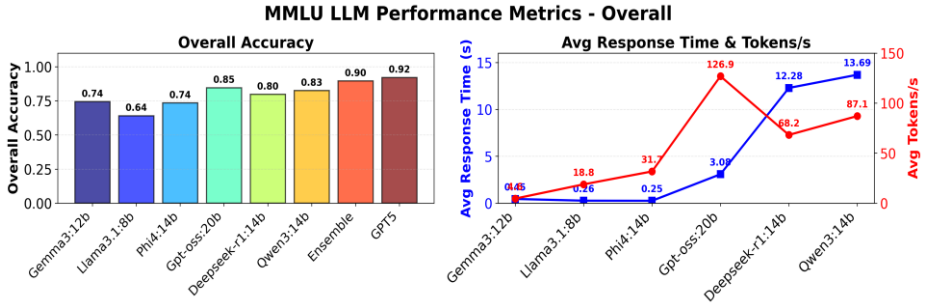

Fig. 3 Benchmarking with MMLU-Finance focused data

Fig. 4 demonstrates subject-specific performance analysis with significant variation across domains. Marketing emerging as the most accessible subject where five of six models achieved accuracy above 0.9, likely due to marketing content prevalence in pre-

training corpora. High School Microeconomics and Management also demonstrated strong performance, with multiple models achieving above 0.85 and 0.75 accuracy respectively, suggesting these qualitative business subjects align well with LLM capabilities. In contrast, Econometrics proved most challenging with accuracy scores ranging from 0.44 (Llama3.1:8b) to 0.81 (Gpt-oss:20b), as only Gpt-oss:20b exceeded 0.80 accuracy, reflecting the difficulty language models face with statistical reasoning and mathematical formulations. Professional Accounting presented similar challenges, with 3 general models achieving below 0.55 and 3 reasoning intensive models reaching more than 0.62, indicating that technical precision and numerical reasoning exceed the capabilities of smaller models. High School Macroeconomics showed moderate difficulty with clear performance stratification: Gpt-oss & Qwen3 (0.87) performed well. Business Ethics and Public Relations demonstrated moderate performance across models (0.66-0.81 and 0.56-0.74 respectively).

Overall, Gpt-oss exhibited the most consistent performance with narrow variance across subjects (0.72-0.91), while Phi4 excelled in qualitative subjects (Marketing: 0.92, Management: 0.85) but struggled with quantitative domains (Professional Accounting: 0.47, Econometrics: 0.48), suggesting optimization for conceptual understanding over numerical problem-solving. These findings underscore that no single model dominates across all dimensions, necessitating application-specific selection considering accuracy requirements, computational constraints, and subject-domain characteristics.

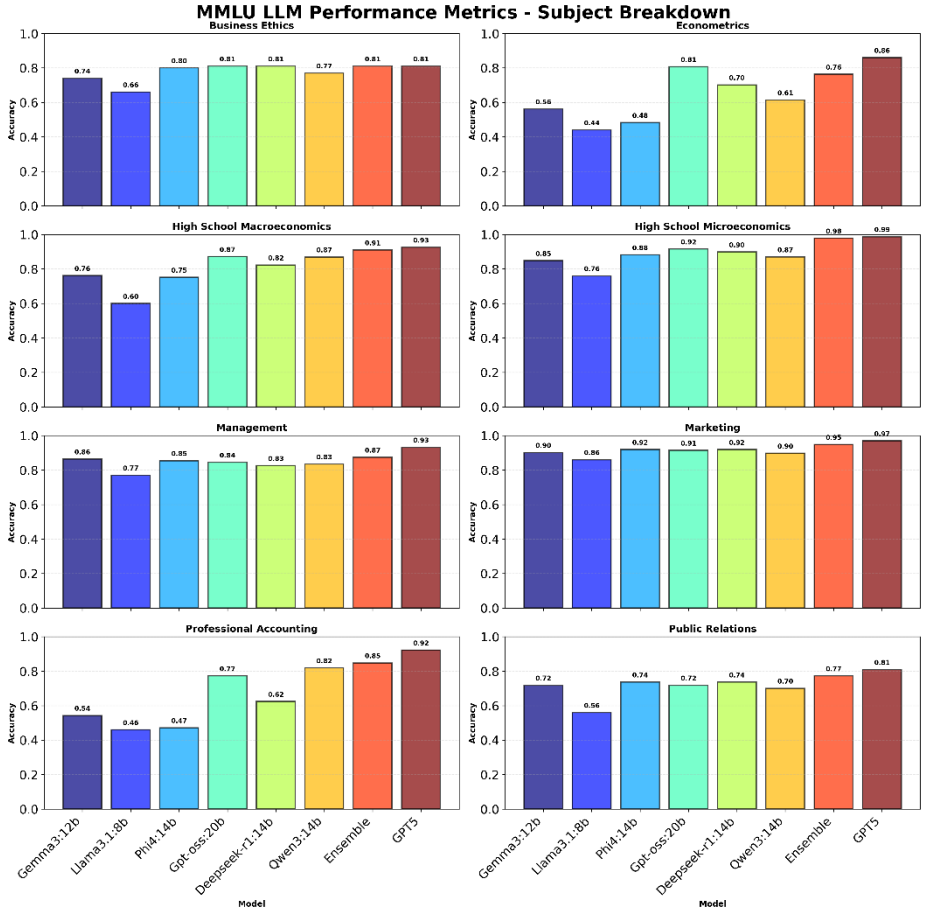

Fig. 4 Benchmarking with MMLU-Finance focused data for selected subjects

4.3 Benchmarking with Finance-Bench data

The evaluation of LLM performance on the FinanceBench dataset using traditional n-gram based metrics (BLEU and ROUGE) revealed consistently low scores across all models, with BLEU scores ranging from 0.04 to 0.07 and ROUGE-1 scores between 0.19 and 0.24 (Table 2). GPT5's BLEU and ROUGE-l scores are also smaller than 0.5. These low values, however, do not necessarily indicate poor model performance but rather highlight a fundamental limitation of token-overlap metrics for evaluating reasoning-intensive tasks. FinanceBench is designed as a RAG scenario where models must extract relevant information from lengthy financial documents, perform multi-step reasoning, and synthesize answers in natural language. Consequently, models often generate verbose, explanatory responses that include intermediate reasoning steps and contextual elaboration (e.g., "The FY2018 capital expenditure for 3M is $1,577 million, calculated from the cash flow statement under investing activities"), while the expected

outputs are typically concise numerical or categorical answers (e.g., "$1577.00"). Despite semantic equivalence, such differences in verbosity and phrasing result in minimal n-gram overlap, yielding artificially low BLEU and ROUGE scores. This discrepancy between metric values and actual answer correctness necessitated the development of a more sophisticated evaluation approach using LLM-as-judge methodology, which can assess semantic similarity and correctness independent of surface-level text matching. This observation aligns with recent literature suggesting that traditional metrics designed for machine translation and summarization tasks are inadequate for evaluating open-ended, reasoning-heavy natural language generation tasks [38], [39].

Table 2. BLEU and ROUGE-score for SLLMs in FinanceBench dataset benchmarking

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
DeepSeek-r1_14b	0.07	0.24	0.12	0.20
Gemma3_12b	0.05	0.19	0.11	0.18
Gpt-oss_20b	0.04	0.23	0.1	0.20
Llama3_8b	0.05	0.20	0.1	0.18
Phi4_14b	0.05	0.22	0.11	0.19
Qwen3_14b	0.06	0.21	0.11	0.19
GPT5	0.08	0.40	0.21	0.38

To demonstrate for the difficulty as mentioned above, let's take a look at the first question (index 0) return from DeepSeek model in json format:


```
{
    "index": 0,
      "financebench_id": "financebench_id_03029",
      "company": "3M",
      "question": "What is the FY2018 capital expenditure amount (in USD
millions) for 3M? Give a response to the question by relying on the
details shown in the cash flow statement.",
         "context": "Table of Contents 3M Company and Subsidiaries
Consolidated Statement of Cash Flow s Years ended December 31 (Millions)
2018 2017 2016 Cash Flows from Operating Activities Net income including
noncontrolling interest $ 5,363 $ 4,869 $ 5,058 Adjustments to reconcile
net income including noncontrolling interest to net cash provided by
operating activities Depreciation and amortization 1,488 1,544 1,474
Company pension and postretirement contributions (370) (967) (383) Cash
Flows from Investing Activities Purchases of property, plant and equipment
(PP&E) (1,577) (1,373) (1,420) Proceeds from sale of PP&E and other assets
262 49 58 Acquisitions, net of cash acquired 13 (2,023) (16) Purchases of
marketable securities and investments (1,828) (2,152) (1,410) Effect of
exchange rate changes on cash and cash equivalents (160) 156 (33) Net
increase (decrease) in cash and cash equivalents (200) 655 600 Cash and
cash equivalents at beginning of year 3,053 2,398 1,798 Cash and cash
equivalents at end of period $ 2,853 $ 3,053 $ 2,398 The accompanying
Notes to Consolidated Financial Statements are an integral part of this
statement. 60",
      "expected_output": "$1577.00",
```

```
      "generated_output_raw": "The FY2018 capital expenditure for 3M is
$1,577 million.\n\nAnswer: $1,577 million",
      "generated_output": "The FY2018 capital expenditure for 3M is $1,577
million. Answer: $1,577 million",
      "metrics": {
        "bleu": 0,
        "rouge1": 0.0,
        "rouge2": 0.0,
        "rougeL": 0,
        "semantic_similarity": 0.0,
        "length_ratio": Infinity,
        "response_length": 12
      },
      "performance": {
        "response_time": 10.71139669418335,
        "tokens_generated": 241,
        "tokens_per_second": 22.499400113793858,
        "prompt_eval_count": 1174,
        "eval_duration": 2.973715188,
        "total_duration": 10.704794484
    }
```


From the box above for question index 0, we can see that the question specifically asked for "FY2018 capital expenditure amount (in USD millions) for 3M" and the expected output generated $1577.00. However, the generated_output has the very long answer with reasoning "The FY2018 capital expenditure for 3M is $1,577 million. Answer: $1,577 million" and led to mismatch in BLEU and ROUGE calculation (0 in this case). Therefore, an additional step is needed, which require the detailed judgement with the support from an SLLM, to read this output and return 1 if the generated_output match with expected_output and 0 otherwise. In this case, the SLLM would return 1 because of the match in final output, following is the prompt we used in extra step:

```
{
Compare expected and generated outputs intelligently based on question.
Returns 1 if they match (considering format differences), 0 otherwise
}
```

The accuracy and average response time with throughput for FinanceBench benchmarking is plotted in Fig. 5. We can see the significant difference between 2 groups general and reasoning intensive in Fig. 5 with the obvious higher accuracy for the latter group. In particular, DeepSeek-r1:14b achieved the highest accuracy at 0.96, even higher than the ground truth GPT5 (0.95) demonstrating superior breaking the tough question into chain of thought capabilities on Financial QA tasks, though at the cost of an average response time of 11.89 seconds per question. Qwen3:14b exhibited comparable accuracy (0.94%) but with notably longer response times (17.43 seconds), compensated by a high token generation throughput of 80.58 tokens per second, suggesting extensive reasoning chains or longer outputs. In contrast, Gpt-oss:20b, despite being the largest model evaluated (20 billion parameters), achieved 0.93 accuracy with substantially faster response times (2.45 seconds) and the highest throughput (108.54 tokens/second), indicating superior computational optimization. The mid-tier models showed moderate performance, with Phi4:14b reaching 0.71 accuracy at 1.35 seconds per response. Both Gemma3:12b and Llama3.1:8b stood last

in accuracy. These results highlight a clear pattern: larger, more sophisticated models achieve higher accuracy at the expense of inference speed, with accuracy improvements of approximately 30-34 percentage points from the smallest to largest models, while response times increased by a factor of 12-25 times. This performance-accuracy trade-off presents important considerations for practical deployment scenarios, where real-time financial analysis applications may require balancing answer quality against latency constraints. Ensemble is also recommended in this step, although it has a little lower accuracy compared to Deepseek but it is more trustworthy than this single model alone.

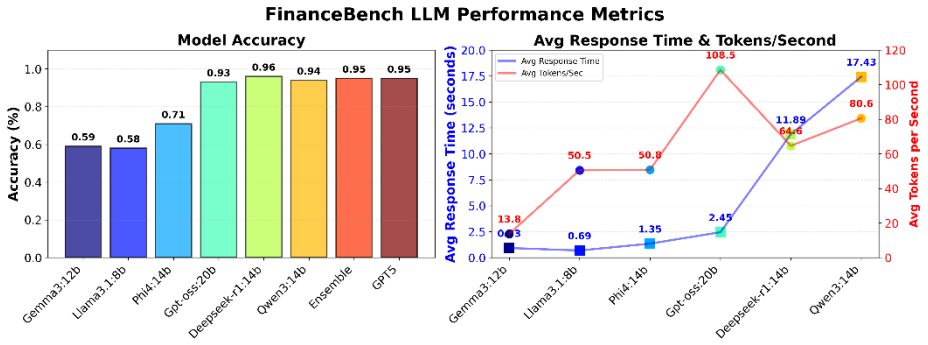

Fig. 5 Benchmarking with FinanceBench data

4.4 Ensemble of SLLMs for Financial data

The evaluation across three benchmarks highlights a clear task-dependent gap between reasoning and general LLMs, informing optimal RAG deployment strategies. For classification and retrieval tasks as in Finance PhraseBank and MMLU, general models like Gemma3:12B and Phi4:14B achieved high accuracy (~0.93) with sub-second latency, outperforming reasoning intensive models such as DeepSeek-R1:14b (0.9 accuracy, 5.3s latency). However, on FinanceBench, which demands multi-step reasoning, reasoning models excelled: DeepSeek-r1:14b and Qwen3:14b reached 0.95 accuracy, much higher than general LLM counterparts despite slower speeds. These results support a hybrid RAG architecture: (1) a fast, lightweight tier using partial-reasoning models for sentiment, retrieval, and factual queries, and (2) a reasoning tier (e.g., DeepSeek-r1 or Gpt-oss) for complex analytical tasks. A query classifier could route requests by complexity, reducing average response time 15–20% while maintaining accuracy within 2–3% of a full reasoning setup—ideal for enterprise-grade, latency-sensitive financial systems.

4.5 RAG with SLLMs for SEC report analysis

Based on the benchmarking results for the 6 SLLMs, an ensemble RAG approach is proposed for querying SEC filings. The dataset includes approximately 1,700 company reports from 2019, covering various document types such as 8-K (Current Reports),

MD&A (Management's Discussion and Analysis), Earnings or Investor Conference Calls, and Risk Disclosures - each containing dense, unstructured financial narratives. For each report, a standardized 50-question prompt template was used to query the RAG framework and generate model responses. A batch processing system was implemented using the SLURM scheduler on an HPC cluster, with each batch script assigned to a specific SLLM. Each script iteratively processed its designated reports, saving outputs in CSV format (for example, company1_MDA_SLLM1.csv). In total, six batch jobs were executed concurrently across all reports, producing over 10,000 CSV files containing RAG-generated responses from the six evaluated models.

The template of input prompt for the first and last 4 questions is as below:

```
{
1. Are they planning to raise product prices in the future? (1 = yes, 0
= no)
2. (Only if Q1 = 1) How confident are you in your analysis of Q1 (1-10)?
3. (Only if Q1 = 1) How certain is the speaker about their future price
increase (1-10)?
4. (Only if Q1 = 1) Was inflation or inflationary pressure cited as a
reason for future price increase (1 = yes, 0 = no)?
…
48. Based on this disclosure, what do you think about this firm's
competitive situation. Is the firm a monopolist (use "1"), in an oligopoly
with less than 5 competitors (use "2"), or an industry that represent
(close to) perfect competition (use "3").
49. How confident are you in your analysis of Q48 (1-10)?
50. Provide the sentence(s) that led you to your conclusion for Q47.
Please provide the sentence or up to three sentences that most strongly
support your conclusion.
}
```

The questionnaire contains sentiment analysis (Q1), rating capability (Q2), reasoning (Q48) and supporting question (Q50) that in line with all the SLLMs that we have just evaluated.

The final response containing ensemble output for one SEC report is tabulated as in Table 3:

Table 3. Ensemble SLLMs in response to individual SEC filing

Model\Ques	**1**	**2**	**3**	**…**	**48**	**49**	**50**
Gemma3:12b	1	8	7		1	9	With many forecasting interest rates to fall, the economy to stay relatively healthy and conditions in the job market to remain favorable, there are certainly reasons to be optimistic about housing demand in the coming years
Llama3.1:8b	0	NA	NA		2	6	Based on the text, there is no indication of a monopolist or oligopoly with less than 5 competitors. The firm mentions 'increased demand' as a reason for increasing production, which suggests that the market is competitive and the firm needs to respond to changing customer needs
Phi4:14b	1	7	9		3	7	Based on the disclosure, it appears that this firm is in an oligopoly with less than 5 competitors. The company does not mention any plans to enter new geographical markets or produce new products, and there's no indication of a threat from new competitors entering their market. However, they

							do discuss the possibility of increased production due to increased demand
Gpt-oss:20b	1	8	8		1	10	By strategically increasing our spec production, we had more inventory available to meet the demand of first-time homebuyers and those consumers worried about mortgage rate volatility. We currently expect the average sales price of closings to remain in the range of $540,000 to $550,000 for the first quarter and the full year of 2024 which is consistent with our fourth quarter pricing
Deepseek-r1:14b	1	9	7		1	9	Given our constructive views on the outlook for housing demand, we are investing in our operations with the goal of growing unit volumes by 5% to 10% annually."" This suggests oligopoly as they're competing with others aiming for growth
Qwen3:14b	0	NA	NA		1	9	Our fourth quarter gross margin was 28.9%, which is down approximately 50 basis points from both the fourth quarter of last year and the third quarter of this year but likely remains the industry leader among the big builders.
Ensemble	1	8	7		1	9	{*Base on the previous 6 response for Question 50, summarize it here*}

5 Discussion

5.1 Interpretation and Implications

The models were grouped into two categories: reasoning-intensive models and general purpose models. The results confirm that all six SLLMs demonstrate a solid understanding of financial terminology and context, even without domain-specific fine-tuning. However, notable trade-offs emerged between accuracy, computational efficiency, and generation speed. General purpose models such as Gemma3:12b and Phi4:14b achieved high accuracy and fast response times on classification and factual tasks, outperforming reasoning models in latency and token generation rate. In contrast, reasoning-enabled models like DeepSeek-R1:14b and Qwen3:14b performed significantly better on tasks requiring multi-step inference and logical reasoning, albeit at higher computational costs. These findings suggest that while reasoning enhances analytical performance, it comes with a substantial time penalty that must be managed in large-scale deployments.

Prompt engineering also proved to be a critical factor influencing model behavior and efficiency. Reasoning models often require more detailed or structured prompts, particularly for seemingly simple questions to trigger appropriate reasoning pathways. Using generic or underspecified prompts led to longer inference times and sometimes redundant reasoning steps. Conversely, general purpose models performed best with concise, direct prompts that aligned with their instruction-following design. This difference highlights the importance of adapting prompt templates to each model's architecture to optimize performance within specific task categories.

One limitation of this study is the use of BLEU and ROUGE metrics to evaluate the text generation performance of SLLMs on the FinanceBench dataset. While these metrics are standard for assessing surface-level similarity, they primarily measure lexical overlaps and do not fully capture semantic equivalence between model-

generated and reference responses. Consequently, the resulting BLEU and ROUGE scores were relatively low, reflecting differences in wording rather than true conceptual divergence. To address this issue, an additional screening step was implemented to refine model selection and improve the semantic alignment between generated and expected outputs. This adjustment proved effective, yielding responses that better preserved the intended meaning and contextual accuracy. Future evaluations may benefit from incorporating semantic-based metrics such as BERTScore or GPT-based evaluators to provide a more comprehensive assessment of text quality beyond surface-level n-gram similarity.

From a deployment standpoint, these findings have direct implications for RAG systems in finance. The results indicate that it is feasible to integrate SLLMs into a RAG framework for querying and analyzing complex documents such as SEC filings. Given their complementary strengths, an ensemble strategy is recommended: lightweight partial-reasoning SLLMs can handle straightforward queries such as sentiment analysis, keyword-based retrieval, and factual lookups, while reasoning-enabled models should be reserved for interpretive or analytical queries that require multi-step reasoning, cross-document inference, or causal explanation.

5.2 Pros and Cons

The use of SLLMs in financial analysis presents several advantages and challenges that influence their practicality in research and enterprise environments.

5.2.1 Advantageous of Open-source SLLMs

Open-source SLLMs are freely available and do not require API keys or paid subscriptions, making them highly accessible for academic institutions, researchers, and organizations with limited funding. Their open architecture allows full control over deployment and customization, enabling secure handling of sensitive or proprietary financial data without reliance on third-party platforms. Because these models can run locally, they support privacy-preserving workflows, which is crucial in finance and compliance-related applications. Additionally, open-source SLLMs integrate well with batch processing and automation frameworks, such as SLURM on HPC systems, allowing for large-scale, repeatable analyses across thousands of documents or queries. This flexibility facilitates efficient experimentation, scalability, and reproducibility in research pipelines. Ensemble also plays a significant role in this benchmarking, giving more confidence to model outputs and achieve significantly higher accuracy among all.

5.2.2 Limitations

Despite these advantages, open-source SLLMs require significant technical expertise and infrastructure to deploy effectively. Installation, model optimization, and environment configuration can be complex compared to proprietary models that provide ready-to-use APIs. Moreover, high computational demand, particularly GPU availability and memory capacity, can limit feasibility for smaller organizations or individual researchers lacking access to HPC resources. The cost and technical barriers of maintaining GPU clusters, managing dependencies, and optimizing inference speed can offset the initial benefit of zero licensing fees.

5.3 Future Research

This study employed a vanilla RAG approach to evaluate the performance of SLLMs in financial text analysis. Future research will extend this work by exploring more advanced RAG techniques, such as Adaptive Retrieval, Reranker integration, and multi-vector memory architectures, to improve both retrieval precision and contextual coherence of generated responses.

In addition, expanding additional benchmarking datasets will be essential to capture broader financial contexts and ensure more robust generalization. Future evaluations will also incorporate improved scoring systems, emphasizing semantic similarity metrics such as BERTScore or GPT-based evaluators to better reflect conceptual alignment between expected and generated outputs.

Another direction involves prompt optimization, where task-specific prompt templates will be designed to maximize model efficiency and accuracy across different financial tasks, such as sentiment analysis, reasoning, and question answering. Finally, the study will be expanded to include a wider range of SLLMs, including variations from the same model families with different parameter scales or distilled versions, to systematically assess performance trade-offs between model size, reasoning ability, and computational cost.

6 Conclusion

This study examined the potential of open-source SLLMs for financial applications, a domain known for its complex terminology, lengthy narratives, complicated structure, and diverse data formats. While recent advances in SLLMs have demonstrated strong reasoning and language understanding in general contexts, their performance on specialized financial tasks has remained largely unexplored. To bridge this gap, six open-source SLLMs from different developers (Google, Microsoft, Meta, OpenAI, Deepseek and Alibaba) were systematically evaluated using three benchmark datasets: Financial PhraseBank, MMLU (finance subset), and FinanceBench, representing sentiment analysis, factual knowledge retrieval, and financial reasoning tasks. GPT5 was also used as ground truth for model performance comparisons.

For the Financial PhraseBank benchmark on sentiment analysis, Gemma3:12b and Phi4:14b emerged as the most balanced and efficient model, achieving near-perfect accuracy: 0.93 overall, while maintaining a sub-second (less than 1s) average response time of and a generation rate of ~15.7 tokens per second, significantly faster than models in the reasoning intensive group. This combination of accuracy and speed makes these models an ideal choice for real-time financial analytics. The third-best performer is Gpt-oss:20b, achieved comparable accuracy and quick response time but having high token generated per second.

For the MMLU benchmark (finance subset), Gpt-oss:20b achieved the highest accuracy score of 0.85; however, general purpose models demonstrated equivalent accuracy with faster response times and more efficient token generation. In contrast, on the FinanceBench dataset, which requires complex chain of thought, the reasoning-enabled group clearly outperformed others. DeepSeek-r1:14b achieved the best accuracy of 0.96, followed by Qwen3:14b (0.94) and Gpt-oss:20B (0.93), with Gpt-oss demonstrating response speeds comparable to smaller models. Overall, the results

highlight a trade-off between reasoning capability and processing speed. Consequently, an ensemble approach that combines multiple SLLMs within the RAG workflow is recommended for SEC report analysis to balance accuracy, reasoning depth, and computational efficiency.

With ensemble of six SLLMs, the results demonstrate that integrating SLLMs with a RAG framework offers a practical, privacy-preserving solution for financial text analysis, particularly when working with sensitive documents such as SEC filings. The comparative evaluation revealed clear trade-offs between reasoning ability, accuracy, and computational efficiency.

Acknowledgments. The authors gratefully acknowledge the computational resources provided by the SMU HPC facilities, including the SuperPOD and ManeFrame III systems. Their generous allocation of computing time and technical support made it possible to conduct the large-scale model benchmarking and data processing required for this study.

References

1. Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). *A survey of large language models for financial applications: Progress, prospects and challenges.* arXiv. https://doi.org/10.48550/arXiv.2406.11903 (arxiv.org)
2. Li, Y., Wang, S., Ding, H., & Chen, H. (2023). *Large language models in finance: A survey.* In *Proceedings of the 4th ACM International Conference on AI in Finance (ICAIF '23), November 27–29, 2023, Brooklyn, NY, USA* (pp. 374–382). ACM. https://doi.org/10.1145/3604237.3626869
3. Huang, A. H., Wang, H., & Yang, Y. (2023). *FinBERT: A large language model for extracting information from financial text.* Contemporary Accounting Research, 40(2), 806-841. https://doi.org/10.1111/1911-3846.12832
4. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). *BloombergGPT: A large language model for finance.* arXiv. https://doi.org/10.48550/arXiv.2303.17564
5. Mun, Y., & Kim, N. (2025). *Leveraging large language models for sentiment analysis and investment strategy development in financial markets.* Journal of Theoretical and Applied Electronic Commerce Research, 20(2), 77. https://doi.org/10.3390/jtaer20020077
6. Kirtac, K., & Germano, G. (2024, August). *Enhanced financial sentiment analysis and trading strategy development using large language models*. In O. De Clercq, V. Barrière, J. Barnes, R. Klinger, J. Sedoc, & S. Tafreshi (Eds.), *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 1-10). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wassa-1.1
7. Jeong, C. (2024). *Fine-tuning and utilization methods of domain-specific LLMs*. arXiv. https://doi.org/10.48550/arXiv.2401.02981
8. Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. Scientific Reports, 15(1), Article 13755. https://doi.org/10.1038/s41598-025-98483-1
9. Busch, K., & Leopold, H (2024). Towards a Benchmark for Large Language Models for Business Process management task. arXiv: https://doi.org/10.48550/arXiv.2410.03255

10. Lu, G., Guo, X., Zhang, R., Zhu, W., & Liu, J. (2025). BizFinBench: A business driven real world finance benchmark for evaluating LLMs. arXiv: https://doi.org/10.48550/arXiv.2505.19457
11. Wang, J., Ding, W., & Zhu, X. (2025). *Financial Analysis: Intelligent Financial Data Analysis System Based on LLM-RAG*. arXiv. https://doi.org/10.48550/arXiv.2504.06279
12. Soudani, H., Kanoulas, E., & Hasibi, F. (2024). *Fine tuning vs. retrieval augmented generation for less popular knowledge.* In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '24), December 9–12, 2024, Tokyo, Japan*. Association for Computing Machinery. https://doi.org/10.1145/3673791.3698415
13. Zhang, B., Yang, H., Zhou, T., Babar, A., & Liu, X-Y. (2023). *Enhancing financial sentiment analysis via retrieval-augmented large language models.* In *Proceedings of the 4th ACM International Conference on AI in Finance (ICAIF '23), November 27–29, 2023, Brooklyn, NY, USA* (pp. 349–356). ACM. https://doi.org/10.1145/3604237.3626866
14. Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). *FinanceBench: A new benchmark for financial question answering*. arXiv. https://doi.org/10.48550/arXiv.2311.11944
15. Patronus AI. (2023). FinanceBench: A new benchmark for financial question answering [Data set]. https://github.com/patronus-ai/financebench (Accessed date: 10/2025)
16. Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the American Society for Information Science and Technology, 65(4), 782–796. https://doi.org/10.1002/asi.23062
17. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. In Proceedings of the International Conference on Learning Representations (ICLR).
18. Rivière, M., et al. (2024). Gemma 2: Improving Open Language Models at a Practical Size arXiv. https://doi.org/10.48550/arXiv.2408.00118.
19. Sapling. (n.d.). The LLM Index: Gemma 2 (9B & 27B parameters) — Google. Sapling. Retrieved October 26, 2025, from https://sapling.ai/llm/gemma2
20. Gemma Team, & DeepMind. (2025). *Gemma 3 Technical Report*. arXiv. https://doi.org/10.48550/arXiv.2503.19786
21. Abdin, M. et al (2024). *Phi-4: A 14-Billion-Parameter Language Model with High Reasoning Capabilities*. arXiv. https://doi.org/10.48550/arXiv.2412.08905
22. Grattafiori, A., et al. (2024). The Llama 3 Herd of Models (arXiv:2407.21783). arXiv. https://doi.org/10.48550/arXiv.2407.21783
23. Gemma-3n.net. (2025). AI Model Leaderboard: The Best Open Source Models. Retrieved October 26, 2025, from https://www.gemma-3n.net/leaderboard
24. Guo, D. et al.. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://doi.org/10.48550/arXiv.2501.12948
25. Yang, A. et al. (2024). *Qwen2.5 Technical Report*. arXiv. https://doi.org/10.48550/arXiv.2412.15115
26. Yang, A., et al. (2025). *Qwen3 Technical Report*. arXiv. https://doi.org/10.48550/arXiv.2505.09388
27. OpenAI. (2025, August 5). Introducing Gpt-oss: Open-weight reasoning models from OpenAI. Retrieved October 26, 2025, from https://openai.com/index/introducing-Gpt-oss
28. Bi, Z., Chen, K., Tseng, C.-Y., Zhang, D., Wang, T., Luo, H., Chen, L., Huang, J., Guan, J., Hao, J., & Song, J. (2025). Is GPT-OSS Good? A comprehensive evaluation of OpenAI's latest open-source models. arXiv. https://doi.org/10.48550/arXiv.2508.12461.
29. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: A method for automatic evaluation of machine translation.* In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

30. Lin, C.-Y. (2004). *ROUGE: A package for automatic evaluation of summaries.* In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)* (pp. 74–81). Association for Computational Linguistics.
31. MiniToolAI Blog. (2025, April 28). *What is Ollama? How to run open-source LLMs on your own computer*. Retrieved October 26, 2025, from https://minitoolai.com/blog/what-is-ollama-how-to-run-open-source-llms-on-your-own-computer
32. DevTurtleBlog. (2024). *Ollama – Guide to running LLM models locally*. DevTurtle. Retrieved October 26, 2025, from https://www.devturtleblog.com/ollama-guide.
33. LangChain Docs. (2025). *OllamaLLM integration.* LangChain Documentation. Retrieved October 26, 2025, from https://python.langchain.com/docs/integrations/llms/ollama
34. Tavakkol, S. (2024, February 5). *Your guide to local LLMs: Ollama deployment, models, and use-cases.* DEV Community. Retrieved October 26, 2025, from https://dev.to/sina14/your-guide-to-local-llms-ollama-deployment-models-and-use-cases-2jng
35. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv. https://doi.org/10.48550/arXiv.2312.10997
36. Mavroudis, V. (2024). *LangChain v0.3* (Preprint, version 1). HAL Open Science. https://doi.org/10.20944/preprints202411.0566.v1
37. Narayan, S. R. B., & Agarwal, N. (2025). *Introduction to LangChain.* In *Mastering LangChain: A Comprehensive Guide to Building Generative AI Applications* (pp. 1–9). Apress. https://doi.org/10.1007/979-8-8688-1718-2_1
38. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). *G-Eval: NLG Evaluation using GPT-4 with better human alignment*. arXiv. https://doi.org/10.48550/arXiv.2303.16634.
39. Gao, M., Hu, X., Ruan, J., Pu, X., & Wan, X. J. (2024). *LLM-based NLG evaluation: Current status and challenges*. arXiv. https://doi.org/10.48550/arXiv.2402.01383