FinForge: Semi-Synthetic Financial Benchmark Generation

Glenn Matlin[1 2 3], **Akhil Theerthala** [1 *], **Anant Gupta**[3 *], **Anirudh Jaidev Mahesh**[2], **Yi Mei Ng**[2], **Rayan Castilla**[3], **Sudheer Chava**[1 2 3]

[1]Financial Services Innovation Lab, Georgia Institute of Technology
[2]College of Business, Georgia Institute of Technology
[3]College of Computing, Georgia Institute of Technology
glenn@gatech.edu, akhiltvsn@gmail.com, agupta886@gatech.edu

Abstract

Evaluating Language Models (LMs) in specialized, high-stakes domains such as finance remains a significant challenge due to the scarcity of open, high-quality, and domain-specific datasets. Existing general-purpose benchmarks provide broad coverage but lack the depth and domain fidelity needed to assess LMs' capabilities for real-world financial reasoning, which requires both conceptual understanding and quantitative rigor. To address this gap, we introduce FinForge, a scalable, semi-synthetic pipeline for constructing finance-specific evaluation benchmarks through a hybrid of expert-guided data curation and controlled LM-based synthesis. FinForge combines manual and programmatic corpus construction from authoritative financial sources with structured question generation and validation using `Gemini 2.5 Flash`. The resulting benchmark comprises over 5,000 human-validated question–answer pairs across 11 finance subdomains, derived from a curated corpus of 100,000 verified documents totaling 143M tokens. Evaluation of state-of-the-art open-source and closed-source models on FinForge reveals significant differences in financial reasoning, with leading models achieving accuracy levels near 80%. These findings underscore FinForge's utility for diagnosing current model limitations and guiding future improvements in financial domain competence.

[*]Equal Contribution

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Introduction

Language Models (LMs) are increasingly adopted for decision support in complex, high-stakes domains such as finance, law, and public policy. While recent advances in LMs have demonstrated strong performance on general knowledge benchmarks and professional exams, reliably evaluating these systems in specialized, knowledge-intensive, and dynamic domains remains a significant challenge. Existing evaluation sets offer broad subject coverage and serve as useful indicators of general knowledge. However, static benchmarks are limited by potential data leakage into LM training corpora, which can artificially inflate performance

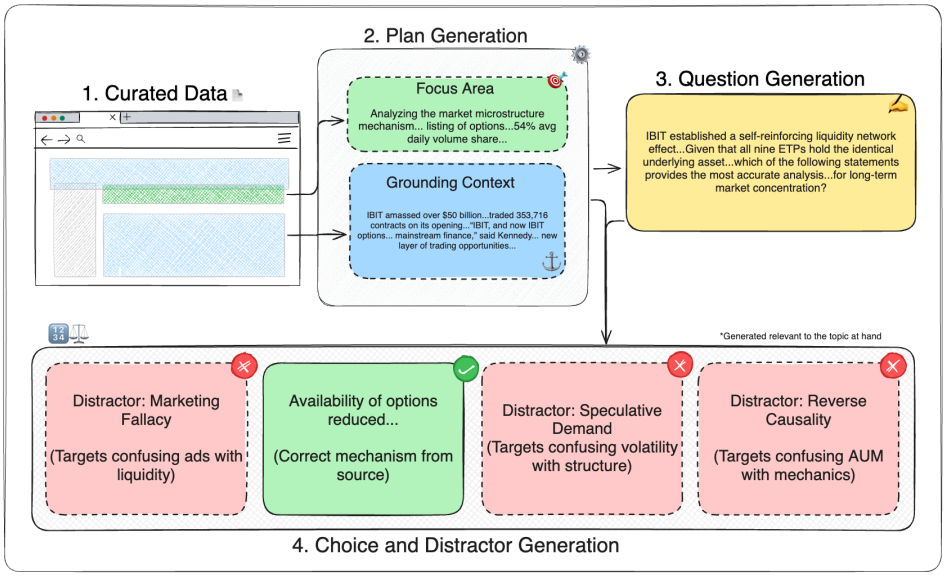

Figure 1: The FinForge Process - The framework ingests raw unstructured financial text (1) and identifies explicit causal triggers and outcomes. Crucially, the Reasoning Engine (2) applies domain-specific schemas to infer the implicit financial mechanism linking them. Finally, it synthesizes a complex, multiple-choice question and answer (3).

due to LM memorization. These challenges are exacerbated in dynamic domains like finance, where knowledge must be continually updated to reflect real-world developments. Empirical evidence suggests that top-performing general models do not necessarily excel at financial tasks requiring domain-specific nuance or quantitative reasoning. This underscores the necessity of dynamic benchmark generation to rigorously assess LM financial knowledge and reasoning robustness in realistic industry scenarios.

Finance presents unique challenges for LM evaluation due to its multi-domain complexity, stochastic characteristics, and stringent regulatory requirements. Effective financial analysis requires both comprehensive domain knowledge—such as familiarity with financial instruments, regulations, and policies—and advanced quantitative problem-solving using real-world data, including asset valuations and risk projections. Additionally, the financial sector evolves rapidly, with new markets, regulations, and trends emerging continuously due to the dynamic nature of global economic systems. Consequently, maintaining up-to-date knowledge is essential for accurate financial reasoning.

This paper introduces FinForge, a methodology for generating semi-synthetic benchmarks tailored to the financial domain. FinForge addresses the gap in the dynamic evaluation of language models' financial knowledge by providing a novel approach to generate a large, diverse set of challenging finance questions grounded in real-world content. The methodology combines human-guided data curation with LM-driven question synthesis, thereby overcoming the limitations of prior static, easily memorizable methods. Initially, domain experts curate a high-quality corpus of up-to-date financial documents—comprising 143M tokens across more than 100,000 curated articles spanning 11 subdomains, including personal finance, corporate finance, macroeconomics, and securitized investments—to ensure comprehensive domain coverage.

The FinForge methodology curates relevant knowledge exclusively from authoritative sources, such as academic textbooks, institutional research, and domain experts, deliberately excluding user forums and trivial stock data. By leveraging verified knowledge, FinForge increases the difficulty of AI evaluations by generating challenging scenarios that test domain-specific problem-solving capabilities. The process employs a multi-stage language model workflow to analyze long-context documents, extract key information, and generate knowledge for creating question–answer pairs. Each question is planned by identifying a central concept or reasoning challenge within the source material, formulating a complex question with embedded background information, and providing a correct answer and distractors. A more advanced language model serves as a judge to validate the quality and financial relevance of each question. This controlled generation process produces difficult, self-contained questions that require expert-level economic insight and multi-step reasoning.

FinForge generates a benchmark comprising 5,000 expert-level finance question–answer pairs, enabling rigorous evaluation of language models' mastery of both conceptual and quantitative finance knowledge. Notably, the benchmark methodology is dynamic: the FinForge framework can be rerun and new documents incorporated to update the question set, allowing continual adaptation to emerging financial knowledge. This paper details the FinForge methodology and demonstrates its utility by benchmarking several state-of-the-art models, thereby highlighting current strengths and weaknesses in financial reasoning.

Related Works

Existing public datasets for evaluating Question Answering (QA) in the financial domain are limited in scope, focus, and recency. While several benchmarks have made significant contributions to specific subtasks, their coverage remains narrow. For example, FinQA (Chen et al. 2021) and its conversational extension ConvFinQA (Chen et al. 2022) provide 8,200 QA pairs centered on numerical reasoning with plaintext representations of tabular financial data, and TAT-QA (Zhu et al. 2021) similarly targets QA on annual reports with tables and text. Although valuable, these datasets primarily address numeric reasoning and do not encompass broader conceptual finance knowledge. The recent FinanceBench (Islam et al. 2023) initiative sought to expand the range of finance QA but offers only 150 question–answer pairs, with relatively simple questions that do not reflect real-world complexity. Similarly, FinTextQA, a long-form QA dataset compiled from textbooks and policy documents, encourages explanatory answers and highlights the limitations of prior benchmarks that focused predominantly on stock data or basic calculations. In the industry, the scarcity of robust evaluation sets has prompted efforts such as S&P Global Kensho's S&P AI Benchmarks, which assembled 600 expert-verified questions across categories, including domain knowledge, quantity extraction, and quantitative reasoning. The emergence of these initiatives underscores the growing demand for more realistic and challenging evaluations of LMs in finance.

A central challenge in constructing dynamic benchmarks is ensuring that questions remain both novel and sufficiently difficult. Since most language models are trained on extensive public internet data, static test suites risk being memorized during training, thereby compromising their evaluative value. To address this, researchers are increasingly exploring dynamic or semi-synthetic benchmarks that can be refreshed or generated as needed (Das et al. 2021; Guo et al. 2024). Previous studies have demonstrated that carefully controlled LM-based generation can yield high-quality, semi-synthetic data at scale, enabling contamination-free evaluation.

General Knowledge and QA Benchmarks

Several benchmarks have been developed to measure broad knowledge and reasoning ability in LMs. MMLU, which tests models on everything from elementary math to professional law and accounting, has become a standard evaluation suite. Other efforts include Big-Bench (Srivastava et al. 2023) and ARC (Chollet et al. 2025), which target complex reasoning or scientific questions. These static benchmarks have driven progress, but are increasingly prone to contamination from training data as models ingest questions and an-

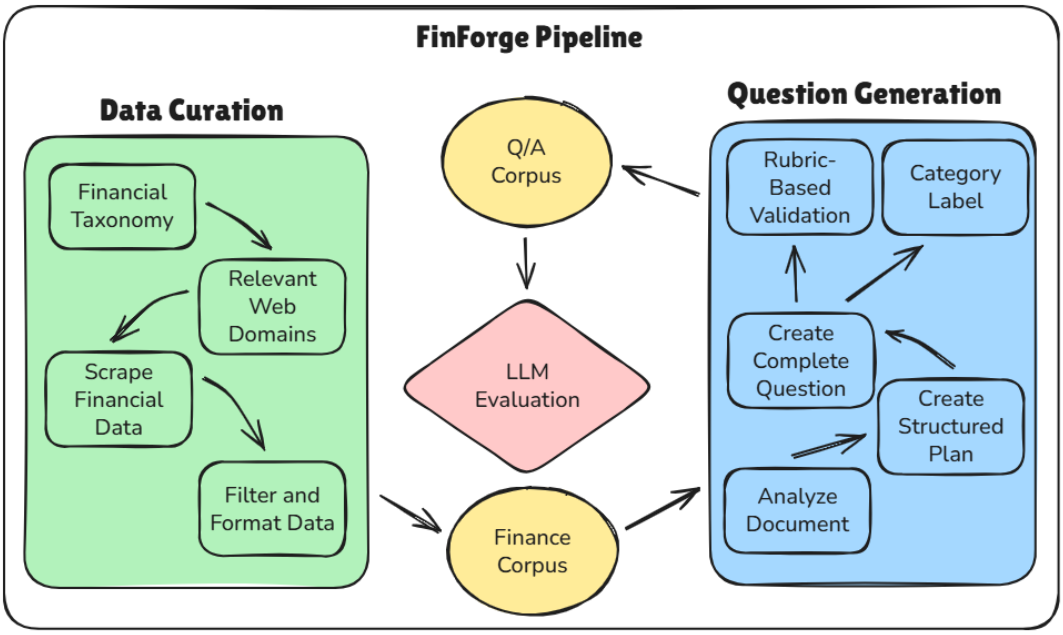

Figure 2: An overview of the FinForge pipeline.

swers from the web. This has sparked interest in more adaptive evaluation methods.

One line of work uses LM-based generation to create new test items. For instance, Auto-Dataset and StructEval prompt LMs to rewrite or expand existing benchmark questions into novel variants. More holistically, LatestEval constructs entirely new reading comprehension sets from real-time sources such as BBC News, using an LM to generate questions for up-to-date passages. There are also interactive evaluation schemes where one LM poses questions or follow-ups to another in a multi-turn dialogue to probe its understanding.

Recently, multi-agent systems have been proposed for automated benchmark creation: BenchAgents splits the task into planning, generation, and verification agents that collaborate (with humans in the loop) to produce high-quality evaluation data. Such approaches yield dynamically extendable benchmarks and help ensure test data novelty. In parallel, research on controllable question generation has introduced techniques to enforce difficulty and content constraints on generated questions. Notably, Li and Zhang (Li and Zhang 2024) propose a "Planning First, Question Second" (PFQS) method in which an LM first outlines a detailed answer plan (with target answer, relevant facts, and cognitive steps), and then another LM generates a question conforming to this plan. This leads to more faithful, expert-aligned questions, as the model must adhere to a blueprint for the desired reasoning. Our FinForge methodology draws inspiration from these advances in controllable question generation.

Financial QA Datasets and Benchmarks

Before our work, relatively few datasets existed for evaluating QA or reasoning in the financial domain, and each covered only a slice of the domain. FinQA was one of the first, featuring questions that require numerical reasoning over company financial reports. It introduced the challenge of performing multi-step arithmetic operations on statements and tables, and showed that models lag far behind human experts on such tasks. TAT-QA (Table-and-Text QA) similarly targets reasoning with hybrid data (earnings tables + text) in financial reports. These datasets primarily evaluate the ability to do structured data reasoning (e.g., reading an annual report) and include annotated programs or formulas for interpretability. A later extension, ConvFinQA, turned FinQA into a multi-turn dialogue challenge. Beyond corporate reports, other benchmarks have been even more limited: the FiQA challenge (Macedo Maia et al. 2018) (2018) released a small set of user-submitted questions and answers on personal finance topics. Most of these lack the complexity and diversity of knowledge needed to test an advanced AI's full financial acumen. For instance, FinQA and TAT-QA do not include conceptual questions on economics or open-ended advisory questions, focusing instead on factoid numeric problems.

A recent effort, FinanceBench (Islam et al. 2023), sought to compile a wider range of financial QA pairs (covering banking, markets, accounting, etc.). Still, it contains only 150 questions in total – too small to capture the breadth of finance or to reliably benchmark modern LMs. Moreover, FinanceBench's question complexity remains limited, falling short of the complexity of real-world expert queries. Recognizing these gaps, Chen et al. introduced FinTextQA (Chen et al. 2024), a long-form QA dataset drawn from finance textbooks and government agency documents. FinTextQA's 1,262 questions are designed to elicit paragraph-length answers, emphasizing explanatory responses over simple cal-

culations. This provides a valuable test of explanation and retrieval capabilities. However, the dataset is relatively small and requires generative answers rather than the multiple-choice format often used in benchmarking. Complementary to academic datasets, industry researchers have also developed proprietary benchmarks. Notably, Kensho's S&P "BizBench" (developed by Koncel-Kedziorski et al.) evaluates models on finance and business tasks across three main categories: domain knowledge (e.g., concept definitions or CFA exam questions), quantitative reasoning (multi-step problems requiring math and finance formulas), and quantity extraction from financial documents. The benchmark consists of 600 expert-curated questions and includes eight tasks for financial reasoning, ranging from code generation for math problems to the FinKnow QA task for conceptual questions. A public leaderboard shows that even top-tier models (e.g., Claude 3.5, GPT-4) struggle with the hardest numerical reasoning questions. These efforts underscore the growing importance of domain-specific evaluation. Our work differs in that we propose an open-source, automated pipeline to generate a much larger set of finance QA pairs (5,000), blending the breadth of coverage of FinanceBench/FinTextQA with the realism and difficulty seen in expert-written exams. FinForge's use of real financial texts as the grounding for each question ensures that the content is current and verified, addressing both the dynamic knowledge aspect and the quality-control issue by providing source evidence for each answer. We view FinForge as a complement to prior benchmarks – pushing the envelope on scale and difficulty – and hope it will enable more robust assessment and improvement of LMs for financial applications.

Methodology

To enable scalable and controlled evaluation of financial reasoning, we required a robust corpus of finance-relevant documents encompassing both numerical and conceptual knowledge. Specifically, the corpus needed to capture (i) quantitative material such as financial calculations and analytical exercises, and (ii) qualitative content reflecting economic principles, market behavior, and institutional context. However, the lack of open-source datasets with verified, diverse, and high-quality financial text posed a significant bottleneck.

To address this gap, we constructed a semi-synthetic dataset using a two-stage pipeline that integrates expert-guided data curation and LM-based question generation. In the first stage, we curated a corpus of high-quality financial documents drawn from authoritative web sources, leveraging a hybrid manual–programmatic pipeline that combines domain expertise with automated filtering and extraction. In the second stage, we employed frontier LMs (specifically, `Gemini 2.5 Flash`)(Team et al. 2025a) to generate diverse, high-quality question–answer pairs from a representative subset of this corpus. Together, these stages yield a scalable, high-fidelity foundation for benchmarking and training models on financial reasoning tasks.

Data Curation

We structured our data curation methodology as a hybrid manual–programmatic pipeline that balances domain expertise with scalability. To minimize effort while maximizing domain coverage and quality, we first decomposed finance into a structured hierarchy of 11 subdomains, including personal finance, corporate finance, investment theory, and macroeconomics. This taxonomy was guided in part by authoritative educational frameworks (i.e., Chartered Financial Accountancy exam) to ensure conceptual completeness and internal consistency.

For each subdomain, humans identified authoritative web domains based on content rigor, institutional credibility, and topical relevance. Sources without clear editorial oversight or academic grounding—such as discussion forums or informal opinion sites—were systematically excluded. This filtering ensured that only high-quality, verifiable content was included in the corpus.

The pipeline then applies a suite of open-source tools and heuristics—including domain whitelisting, keyword co-occurrence, sitemap traversal, and link structure analysis—to automatically identify, filter, and rank candidate sites for extraction. Once candidate sites were finalized, we leveraged each site's structure to efficiently extract relevant financial text, avoiding the need for exhaustive manual traversal. For text extraction and parsing, we employed `Trafilatura` and `BeautifulSoup` for HTML-based content, and `PyMuPDF4LLM` for PDF documents, ensuring consistent text normalization and formatting across source types.

The process cleanly separates filtering and extraction stages, enabling parallelized domain filtering and asynchronous content extraction at scale.

Question Generation and Validation

Recent works in question generation demonstrate the scalability and controllability of using tailored LM agents for this task (Li and Zhang 2024; Noorbakhsh et al. 2025). Our approach synthesizes insights from these methods and adapts them for controlled generation from domain-specific documents.

FinForge uses an automated five-stage process, as illustrated in Figure 2. First, we perform a deep analysis of the input documents to extract salient information. This analysis informs the generation of a structured answer plan, which serves as a blueprint for guiding question formulation. Using this plan, we then generate complex questions that remain strictly grounded in the source material. In the final stage, an LM-as-a-judge framework validates these questions against a predefined rubric, filtering for relevance and quality.

Synthetic Question Generation The preliminary document analysis phase aims to discern deep financial thinking patterns within articles to uncover opportunities for probing questions. We specifically focus on identifying four essential characteristics in a document: causal relationships, prominent and competing hypotheses, necessary assumptions, and counterfactual possibilities.

This breakdown of the document is crucial during the planning stage. In this stage, we translate the unstructured information into a concrete blueprint for generation. This involves identifying a specific, testable conceptual nucleus within the text that serves as the "focus area." The agent also assesses the cognitive complexity of this concept, assigning a difficulty rating on a five-point scale. Finally, it extracts the minimal set of relevant passages required to construct a self-contained question. This plan ensures that the question is well-defined, appropriately challenging, and directly traceable to the source document.

In the third stage, we use this blueprint to formulate a complete question–answer pair. Adhering to the principle of self-containment, all necessary context and data from the relevant passages are embedded directly into the question's premise. The language model for this stage is prompted to generate a natural-language question, plausible distractors, and a concise explanation for the correct answer, all while adhering to the domain-specific requirements provided in the inputs.

In addition to the preceding processes, we implement a supplementary labeling phase to categorize the artifacts based on the financial issue, perceived difficulty, and the targeted model's finance-related capability. These labels help filter out irrelevant questions and provide valuable insights into a model's performance across diverse settings. Each label is linked to its own case-specific rules that help the model adapt to the nature of the topic it must address.

Validation and Filtering The primary issue of automated generation is that outputs frequently fail to meet benchmark standards. We frequently observed unclear inquiries, erroneous hypotheses, or entirely unrelated questions arising from ostensibly direct materials. Consequently, we implemented an extra question validation phase that uses a language model to assess the question's validity for the specific use case. Each question is evaluated across multiple dimensions, including financial relevance, self-sufficiency, logical consistency, clarity, and complexity, before being deemed suitable for a benchmark.

Each iteration of the pipeline development has undergone rigorous validation through a human-in-the-loop process, independent of the existing automated checks. Human validation requires sampling a small portion of the developed questions for thorough quality inspection. The selection of these questions is based on the response capabilities of the GPT-4o (team et al. 2024) and Qwen-2.5-72B (Team 2024) models. The sampled questions included responses from both models, including incorrect answers and instances where one model answered correctly while the other did not. This allowed the human expert to obtain a comprehensive understanding of the pipeline's generation quality compared to random sampling of generated questions. Based on this expert feedback, we iteratively refined the generation and filtering logic to produce a final set of higher-quality, more challenging questions.

Results

The complete corpus construction and question-generation pipeline was executed over seven days, encompassing taxonomy design, domain filtering, and content extraction. The data curation stage yielded over **100,000** high-quality and verified financial documents spanning eleven well-defined subdomains. Collectively, these documents contained **143M tokens** of domain-specific text, providing a diverse foundation that integrates both quantitative and conceptual financial knowledge.

From this corpus, we sampled 10,000 documents for question generation using `Gemini 2.5 Flash` as the synthesis model. This process produced **10,000** initial question–answer (Q/A) pairs. Subsequent automated filtering—combining rule-based quality control with LM-as-judge scoring for relevance, clarity, and factual accuracy—resulted in a refined, high-quality benchmark set of **5,000** Q/A pairs. To further validate reliability, a stratified random sample of **500** Q/A pairs was manually reviewed by domain experts to ensure alignment with real-world financial reasoning standards.

Model Benchmarking

The validated `FinForge` benchmark was used to evaluate a range of both open-source and closed-source language models (Table 1). Performance was measured using a consistent multiple-choice setup, with accuracy defined as the proportion of correctly predicted answers across 5,000 question–answer pairs.

The table summarizes performance across representative models. The models evaluated can be categorized into two distinct approaches: availability and scale. We initially categorized the evaluation models into proprietary and open-source models based on their availability. Our evaluations indicate that while closed-source models such as `GPT-4o` and `Claude Sonnet 4` exhibit superior generalizability compared to mid-range and large open-source models, they underperform on financial benchmarks, achieving comparable accuracies of **73.4**% and **72.6**%, respectively. Open-source models of the same generation, such as `Qwen-3-235B` and `DeepSeek v3.1`, demonstrate markedly superior performance on the benchmark.

Additionally, we have categorized the evaluated models into three key groups to assess the influence of model scale on their performance against the benchmark. Surprisingly, mid-range models (32-110B) demonstrate performance comparable to large-scale proprietary models, with `Qwen3-Next-80B` exhibiting only a 5% deficit relative to `Qwen-3-235B` and remaining within 1% of the DeepSeek, GPT-4o, and Sonnet models.

The performance scores reveal two crucial factors for financial evaluation:

- Even among generalist models, attaining state-of-the-art performance does not necessarily indicate superior financial reasoning capabilities.
- The foundational training distribution and the architectural designs for financial reasoning frequently overshadow the model's scale.

Table 1: Model accuracies (proportion correct) on the FinForge benchmark (5k samples).

Models	Overall	Alt/RE	Beh/Quant	Corp Fin & Val	FinTech	FAR	Ethics & Gov	Mkt & Deriv	Reg & Comp	Portf Mgmt	Wealth Mgmt	Pub/Intl Fin
Qwen 3 235B(2025)	**0.771**	**0.776**	**0.852**	**0.739**	**0.950**	**0.783**	**0.938**	**0.874**	**0.819**	**0.803**	0.610	0.815
DeepSeek V3.1(2024)	0.739	0.705	0.820	0.698	**0.950**	0.743	**0.938**	0.863	0.794	0.774	0.603	**0.818**
GPT-4o(2024)	0.734	0.717	0.762	0.704	0.875	0.743	0.875	0.822	0.767	0.746	**0.653**	**0.818**
Qwen3-Next-80B(2025)	0.732	0.720	0.803	0.700	0.925	0.739	**0.938**	0.855	0.782	0.744	0.590	0.793
Sonnet 4(2025)	0.726	0.756	0.713	0.693	0.875	0.770	0.812	0.798	0.787	0.750	0.613	0.771
Llama 3.3 70B Turbo(2024)	0.725	0.709	0.779	0.690	0.875	0.686	0.875	0.839	0.804	0.726	0.640	0.799
OLMo-2-7B(2024)	0.608	0.626	0.730	0.568	0.750	0.566	0.812	0.727	0.662	0.636	0.450	0.702
OLMo-2-32B(2024)	0.567	0.516	0.721	0.516	0.750	0.527	0.812	0.738	0.615	0.605	0.463	0.650
Llama 4 Scout(2025)	0.465	0.413	0.590	0.363	0.600	0.593	0.812	0.724	0.625	0.520	0.283	0.581

Notes:
(i) Alt/RE = Alternative Investments & Real Estate; Beh/Quant = Behavioral & Quant Finance; Corp Fin & Val = Corporate Finance & Valuation; FinTech = FinTech & Innovation; FAR = Financial Accounting & Reporting; Ethics & Gov = Financial Ethics & Governance; Mkt & Deriv = Markets & Derivatives; Reg & Comp = Regulation & Compliance; Portf Mgmt = Investment & Portfolio Management; Wealth Mgmt = Personal Finance & Wealth Management; Pub/Intl Fin = Public & International Finance.

Expert Evaluation

To enhance the validation of the quantitative results on the benchmarks, we conducted a qualitative assessment on 10% of the benchmark dataset. The qualitative assessment entailed verifying the clarity, self-containment, plausibility, and real-world relevance of the questions.

The expert review validated the high quality and complexity of the generated questions, with 70% of the 500 samples deemed clear, accurate, and relevant. Importantly, the remaining 30% were not necessarily factually incorrect, but were flagged by experts for ambiguity, missing contextual assumptions required for a definitive answer, or a lack of real-world plausibility, making them not ideal for a rigorous benchmark. This validation rate significantly contrasts with the 100% approval rate observed from the automated LM-as-a-judge (Stage 5) for the identical 500-sample set. The 30-point discrepancy offers quantitative support for our conclusion regarding the limitations of the 'LM-as-a-judge,' indicating that it currently lacks the sophistication required to assess complex financial reasoning, thereby underscoring the necessity of human-in-the-loop validation.

This expert review served two critical functions:

- It quantified the final benchmark's quality, with a 70% expert approval rate.
- It starkly demonstrated the critical 30-point gap between automated LM validation (100% approval) and human expert scrutiny, proving that human-in-the-loop oversight remains essential for this complex domain.

Addressing Data Contamination and Circularity

As a methodological sanity check, we evaluated the generator model (Gemini 2.5 Flash) and a model from the same family (Gemma 27B Team et al. (2025b)) on the resulting benchmark. They achieved scores of 79.3% and 74.0%, respectively. Due to the inherent risk of data contamination from this circular evaluation, we explicitly omit the Gemma and Gemini models from our primary benchmark (Table 1) to ensure a fair and rigorous assessment of other models.

Discussion

The purpose of a benchmark is to assess a model's performance in a particular domain or task, thereby helping researchers identify gaps that require attention. The sample benchmark created using the FinForge framework enables the identification of analogous deficiencies in the model's capabilities. To obtain a comprehensive assessment of model performance, we conduct a two-stage evaluation. In addition to the accuracy scores presented in **Table 1**, we perform a proportional error analysis to assess the model's performance on a given topic precisely.

Our analysis indicates that ***Personal Finance & Wealth Management*** and ***Corporate Finance & Valuation*** are notably challenging topics relative to others in the benchmark. This factor may arise from multiple reasons; however, fundamentally, we characterize these domains as essential personalization and true-reasoning challenges in contrast to traditional retrieval-based finance tasks. Recent studies indicate that LMs lack the personalization needed to address personal finance inquiries in their native state (Takayanagi et al. 2025; Hean, Saha, and Saha 2025). Another contributing factor may be the excessive availability of investment and stock-market data in contemporary finance datasets, as evidenced by the comparatively high performance of the models in the ***Markets and Derivatives*** and ***Portfolio Management*** questions. This excessive nature may negatively affect the model's reasoning process in other domains, leading to a decline in performance.

Conversely, across subjects, evidence indicates that the models struggle to address quantitative questions, followed by counterfactual inquiries. This aligns with the recent understanding of language model capabilities. It is noteworthy that the multi-hop reasoning questions within the framework are frequently derived from different sections of the same document, thereby considerably reducing the difficulty level.

Expert evaluation of incorrect quantitative answers identified two distinct failure modes. The initial issue was a failure in conceptual reasoning, characterized by the application

of incorrect financial methodologies, flawed assumptions, or inadequate multi-step logic construction (e.g., miscalculating depreciation or utilizing an incorrect tax rate). The second issue was arithmetic failure, in which the model recognized the appropriate financial steps but erred in the final calculation. This distinction is essential. Arithmetic failures, a recognized limitation of models as token predictors, can be effectively mitigated by integrating external calculators or tools. However, the conceptual failures reveal a more fundamental deficiency. Models are not merely failing in calculations; they are fundamentally misinterpreting the intricate financial reasoning demanded by the prompt. Future studies should concentrate on identifying and addressing these conceptual misinterpretations to facilitate meaningful advancement in financial reasoning tasks.

The findings possess substantial real-world implications. The significant shortcomings in *Personal Finance & Wealth Management* highlight that the average user cannot currently rely on these models for important financial decisions. This represents a significant gap that has previously been overlooked in benchmarks focused on conventional information retrieval.

This work contributes in two ways. We first identify and analyze specific, high-impact weaknesses in contemporary LMs concerning financial reasoning and personalization. We demonstrate that the FinForge framework is an effective method for generating nuanced, domain-specific benchmarks. This outlines a clear framework for guiding future advancements and focused enhancements to financial-sector models.

Conclusion

FinForge demonstrates that scalable, domain-grounded benchmark generation is achievable through a principled combination of expert oversight and controlled LM synthesis. By integrating verified financial sources with structured question generation and multi-stage validation, FinForge produces high-quality datasets that accurately reflect the depth of reasoning and quantitative rigor demanded in economic analysis. Our evaluations reveal not just variability but specific, high-impact weaknesses, particularly the models' tendency to fail at conceptual reasoning rather than simple arithmetic. Furthermore, our human-in-the-loop validation revealed a critical 30-point discrepancy between automated (100%) and expert (70%) approval, underscoring the necessity of human oversight in complex domains. Beyond its immediate utility as a financial reasoning benchmark, FinForge offers a generalizable methodology for creating transparent, extendable evaluation pipelines in other specialized domains. Future work will focus on expanding the corpus to additional subfields, integrating temporal and dynamic data to assess model recency, and establishing continual-learning benchmarks that evolve alongside financial markets. Ultimately, FinForge lays the groundwork for systematic, reproducible, and evolving evaluation of LMs in finance and other expert-driven disciplines.

Limitations and Future Work

The primary limitation of the study is the reliance on Gemini 2.5 Flash for both question generation and evaluation. Manual verification indicated that, although the generated questions were often challenging, their contexts frequently lacked the necessary assumptions for a precise response. The ambiguity, likely arising from the generator's speed optimization, poses a risk of leading assessed LMs to make incorrect assumptions, thereby compromising the reliability of assessments.

Gemini 2.5 Flash's design, which prioritizes speed, results in a capabilities mismatch when evaluating complex reasoning models. This likely led to the evaluator lacking the sophistication necessary to accurately assess advanced outputs, which may have distorted performance metrics. The boolean validator operated as a "black box." The system effectively filtered out irrelevant questions; however, its lack of transparency obstructed the analysis of failed questions, thereby impeding the improvement of the generation pipeline.

It is important to consider the issues of data contamination and circular dependency arising from the evaluation of the same family of models on benchmark datasets. The present study omits the performance of Gemini and Gemma models from the discussion; however, future research should focus on addressing this issue to enhance the accuracy of performance evaluations for these models.

Future research should concentrate on creating more transparent, multi-dimensional validation strategies, such as clarity and contextual sufficiency, to address the "black box" issue and enhance question generation, moving beyond simplistic boolean evaluations.

References

Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4. Technical report, Anthropic.

Chen, J.; Zhou, P.; Hua, Y.; Xin, L.; Chen, K.; Li, Z.; Zhu, B.; and Liang, J. 2024. FinTextQA: A Dataset for Long-form Financial Question Answering. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics.

Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.-H.; Routledge, B. R.; et al. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3697–3711.

Chen, Z.; Li, S.; Smiley, C.; Ma, Z.; Shah, S.; and Wang, W. Y. 2022. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6279–6292. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Chollet, F.; Knoop, M.; Kamradt, G.; and Landers, B. 2025. ARC Prize 2024: Technical Report. arXiv:2412.04604.

Das, B.; Majumder, M.; Phadikar, S.; and Sekh, A. 2021. Automatic question generation and answer assessment: A survey. *Research and Practice in Technology Enhanced Learning*, 16(5): 1–15.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Guo, S.; Liao, L.; Li, C.; and Chua, T.-S. 2024. A Survey on Neural Question Generation: Methods, Applications, and Prospects. *arXiv preprint arXiv:2402.18267*.

Hean, O.; Saha, U.; and Saha, B. 2025. Can AI help with your personal finances? *Applied Economics*, 1–9.

Islam, P.; Kannappan, A.; Kiela, D.; Qian, R.; Scherrer, N.; and Vidgen, B. 2023. FinanceBench: A New Benchmark for Financial Question Answering. *arXiv preprint arXiv:2311.11944*.

Li, K.; and Zhang, Y. 2024. Planning First, Question Second: An LLM-Guided Method for Controllable Question Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 4715–4729. Bangkok, Thailand: Association for Computational Linguistics.

Macedo Maia, S.; Handschuh, S.; Freitas, A.; Davis, B.; McDermott, R.; Zarrouk, M.; and Balahur, A. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. In *WWW '18 Companion: The 2018 Web Conference Companion*, 1941–1942. Lyon, France: ACM. ISBN 9781450356404.

Meta AI. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Noorbakhsh, K.; Chandler, J.; Karimi, P.; Alizadeh, M.; and Balakrishnan, H. 2025. Savaal: Scalable Concept-Driven Question Generation to Enhance Human Learning. arXiv:2502.12477.

OLMo, T.; et al. 2024. 2 OLMo 2 Furious.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615.

Takayanagi, T.; Izumi, K.; Sanz-Cruzado, J.; McCreadie, R.; and Ounis, I. 2025. Are Generative AI Agents Effective Personalized Financial Advisors? arXiv:2504.05862.

Team, G.; et al. 2025a. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Team, G.; et al. 2025b. Gemma 3 Technical Report. arXiv:2503.19786.

team, O.; et al. 2024. GPT-4o System Card. arXiv:2410.21276.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; and Chua, T.-S. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3277–3287. Online: Association for Computational Linguistics.