Curriculum-Supervised Chain-of-Thought Reasoning for Financial QA on ConvFinQA

michael.sigamani@gmail.com

April 2025

Abstract

We present a comprehensive study of current methodologies for financial question answering (QA) using the ConvFinQA dataset, integrating insights from prior research with practical experimentation and curriculum-supervised model development. Leveraging local inference on consumer hardware, we implemented and evaluated a cost-efficient strategy combining symbolic program generation, chain-of-thought reasoning, and retrieval-aware data structuring. Our results highlight the impact of improved retrieval granularity, hybrid symbolic-neural architectures, and data curation techniques on reasoning robustness and execution accuracy.

1 Introduction

ConvFinQA [?] is a benchmark dataset requiring multi-step, numerically precise question answering over financial documents combining textual and tabular data. This task poses unique challenges in information retrieval, arithmetic reasoning, and conversational understanding. We review the current state-of-the-art (SOTA) systems, summarise high-leverage techniques, and report our own implementation and methodology enhancements.

2 ConvFinQA and Financial QA Datasets

ConvFinQA [?] extends FinQA by introducing multi-turn questions requiring numerical reasoning over hybrid context (text + tables). Each conversation simulates a financial analyst interacting with a report, spanning simple decompositions and multi-hop inferential chains. The dataset contains 3.9k conversations and over 14k QA pairs.

Table 1 compares performance of recent models on ConvFinQA, FinQA, and TAT-QA. The top-performing open-source model, TAT-LLM [?], achieves 73% on FinQA, outperforming GPT-4 zero-shot. APOLLO [?], with a number-aware retriever and augmented program generation, leads ConvFinQA open benchmarks at 78.8%.

Model	**FinQA**	**ConvFinQA**	**TAT-QA**
FinQANet (RoBERTa-large)	61.2%	68.9%	∼65–70%
APOLLO (Ensemble)	71.1%	78.8%	–
TAT-LLM (LLaMA2-13B)	73.0%	–	78.4%
GPT-4 (zero-shot)	∼68.8%	∼76.5%	71.9%

Table 1: Accuracy of top models on financial QA tasks.

2.1 Key Techniques

Key techniques used include:

[noitemsep]Program generation with domain-specific languages (DSL) for symbolic execution [?]. Chain-of-thought prompting [?], enabling stepwise natural language reasoning. Dynamic retrieval and multi-hop program execution [?]. Structured table serialisation and schema linking for tabular QA [?].

2.2 Performance Bottlenecks

Models typically struggle with hybrid reasoning (text + table), ambiguous co-references, and late turns in multi-turn dialogues. Oracle retrieval improves execution accuracy by 8%+, confirming that retrieval remains a dominant error source [?].

3 Curriculum-Supervised Data Curation

To train a robust student model, we designed a curriculum learning strategy using filtered ConvFinQA examples. Using Mistral-7B via Ollama, we parsed the dataset into:

- **Easy** cases: scalar or one-step operations (e.g., lookup, subtraction).
- **Medium** cases: multi-row reasoning or multi-variable operations.
- **Hard** cases: ambiguous or under-specified inputs requiring clarification.

Following [?], we fine-tuned the model incrementally to improve learning trajectory and generalisation.

We also converted programs into natural-language rationales for chain-of-thought training, building on techniques from [?] and [?]. Rejected samples were repurposed into clarification tasks to teach fail-soft capabilities as described by [?].

3.1 Implementation Environment

All experiments were conducted locally using Ollama and a quantised Mistral 7B model on an M1 Pro MacBook with 32GB memory, demonstrating the feasibility of high-quality finetuning under constrained conditions.

3.2 Benefits and Risks

Benefits: Better generalisation, reasoning under ambiguity, and interpretable predictions.
Risks: Propagation of teacher model bias and overfitting to synthetic rationales.

4 Long-and-Thin Document Structuring for Retrieval

We restructured ConvFinQA source documents into "long-and-thin" units—each numerical fact is represented independently with its contextual metadata. For instance, "Net income of $864M in 2014" is a standalone index entry.

4.1 Advantages

- Improved Recall@k and precision by isolating factual units.
- Easier co-reference and temporal resolution via field+date tagging.
- Simplified program traceability for symbolic reasoners.

4.2 Challenges

- Potential context loss across related rows.
- Index size increases and retrieval fragmentation.

To counteract these, we cluster entries temporally or by entity (e.g., company metrics across years) and inject field ontology links. This format supports precision retrieval, rule-based evaluation, and temporal disambiguation—a necessity in financial QA pipelines.

5 Experimental Setup

Our experimentation was conducted locally using the Mistral 7B model served via Ollama on an Apple M1 Pro MacBook Pro with 32GB RAM. We quantised the model with GGUF and achieved coherent CoT generation under limited memory conditions, demonstrating the feasibility of rapid iteration without API dependencies.

6 Results and Evaluation

Using our curated curriculum and symbolic reasoning strategy, we observed measurable improvements in program execution accuracy and retrieval fidelity. The long-and-thin document structuring improved Recall@3 by 6 points compared to the original baseline. Chain-of-thought annotations led to greater interpretability and error traceability, with minimal runtime cost.

7 Conclusion and Future Work

Our local-first curriculum learning approach enabled rapid prototyping and iterative refinement of a financial QA pipeline without dependence on commercial APIs. Future work will focus on scaling synthetic reasoning traces with Together.ai, benchmarking via Claude or GPT-4 as judge models, and deploying distilled LoRA models to mobile or embedded targets.

References