FinVERIFY: Multi-Aspect Retrieval-Augmented Financial Fact-Checking

Shashank Dugad (sd5957), Utkarsh Arora (ua2152),
Shivam Balikondwar (ssb10002), Surbhi (xs2682)
NLP Final Project Proposal - Fall 2025

1 Paper Summary

We focus on **"Multi-Aspect Integration for Enhanced Retrieval-Augmented Generation"** (Wang et al., 2025, arXiv:2503.15191). Standard RAG systems use only dense semantic retrieval, missing documents that match lexically or contain specific entities/dates. MAINRAG addresses this by combining semantic, lexical (BM25), entity-based, and temporal retrieval signals using Reciprocal Rank Fusion (RRF). **Why this paper:** Financial fact-checking requires precise entity matching (company names), temporal filtering (fiscal quarters), and numerical reasoning—impossible with semantic search alone. MAINRAG achieved 8-12% improvement over single-aspect baselines on FEVER and HoVer benchmarks. **Limitations:** Only tested on general fact-checking, lacks cross-encoder reranking, no evidence quality analysis. We extend MAINRAG to financial domain with reranking and citation accuracy metrics.

2 Project Description

2.1 Goal & Architecture

Build **FinVERIFY**, a financial fact-checking system that retrieves evidence from 200M+ tokens of financial documents using MAINRAG's multi-aspect retrieval to generate fact-checked answers with citations.

System Pipeline:

```
Query: "Did Apple's Q4 2023 revenue exceed $85B?"
   ↓
[MAINRAG Multi-Aspect Retrieval]
    Dense Semantic (BGE embeddings)
    Sparse Lexical (BM25)
    Entity Filtering (company names)
    Temporal Filtering (date ranges)
   ↓
Reciprocal Rank Fusion → Top 100 documents
   ↓
Cross-Encoder Reranking → Top 10 documents
   ↓
Evidence Extraction → Relevant snippets
   ↓
Flan-T5-XL Generation
   ↓
Output:
```

```
- Answer: "Yes, Apple's Q4 2023 revenue was $89.5B"
- Verdict: SUPPORTED / REFUTED / NOT_ENOUGH_INFO
- Citations: [doc_234], [doc_567]
```

2.2 Data Sources

Corpus (200M+ tokens): SEC EDGAR 10-K/10-Q filings (100M), HF edgar-corpus, financial news (50M)
Primary Evaluation: FinanceBench (10K questions) - real-world questions by non-experts, diverse financial topics
Secondary Evaluation: TATQA (16.5K questions) - arithmetic reasoning on tables, tests numerical accuracy
Training: Glaive RAG-v1 (10K examples) for fine-tuning
Rationale: FinanceBench tests real-world applicability, TATQA tests structured data reasoning. Together they cover fact-checking breadth (26.5K questions) within project timeline.

2.3 Evaluation & Baselines

Metrics: F1/EM (answer quality), verdict accuracy (fact-checking), Recall@10/Precision@10 (retrieval), citation F1 (evidence), latency
Baselines: (1) BM25+T5, (2) DPR+T5, (3) Standard RAG, (4) MAINRAG (reproduced), (5) FinVERIFY (ours)
Ablations: Remove semantic, lexical, entity, temporal aspects; remove cross-encoder

2.4 Milestones

Weeks 1-2: Download SEC EDGAR + FinanceBench + TATQA, chunk documents (512 tokens), generate BGE embeddings, build FAISS + BM25 indexes
Weeks 3-4: Implement BM25+T5 and DPR+T5 baselines, reproduce MAINRAG 4-aspect retrieval + RRF, establish baseline metrics
Weeks 5-6: Add cross-encoder reranking, implement evidence extraction, fine-tune Flan-T5-XL on Glaive RAG-v1
Weeks 7-9: Full evaluation on FinanceBench + TATQA, ablation studies (remove each MAINRAG aspect), error analysis
Weeks 10-11: Build Gradio demo, write final report, design poster, prepare presentation

3 Technologies & Innovation

Stack: Python 3.11, PyTorch 2.0, Transformers 4.35, FAISS, Rank-BM25
Models: BGE-large (embeddings), cross-encoder/MiniLM (reranking), Flan-T5-XL (generation)
Compute: NYU HPC (4× A100), Google Colab Pro

Our contributions beyond MAINRAG: (1) First financial domain application, (2) Cross-encoder precision boost, (3) Evidence quality metrics, (4) Comprehensive evaluation on real-world financial QA, (5) Ablation analysis quantifying each retrieval aspect's impact

4 Team Roles

Shashank (25%): Data download, demo, poster. **Utkarsh (25%):** SEC corpus, embeddings, BM25. **Shivam (25%):** FAISS index, reranking, metrics, ablations. **Surbhi (25%):** MAINRAG pipeline, RRF, T5 fine-tuning, report. **Shared:** Code review, experiments, presentation.