FIND-MA: A Retrieval-Augmented Multi-Agent Framework for Fundamental Company Analysis and Financial Insight

Vikranth Udandarao[1] and Akshat Parmar[1]

[1]IIIT Delhi, India
{vikranth22570, akshat22050}@iiitd.ac.in

Abstract

In this work, we introduce **FIND-MA** (**F**inancial **I**nsight via a **N**etwork of **D**istributed **M**ulti-**A**gents), a Retrieval-Augmented Generation (RAG)-based multi-agent framework for fundamental company analysis aimed at enhancing financial decision-making. FIND-MA leverages the reasoning capabilities of state-of-the-art large language models, including DeepSeek-R1 and Qwen3, which combine structured inference with deep contextual understanding. The framework orchestrates a network of specialized agents, each responsible for evaluating a distinct aspect of a company—such as financial health, market positioning, leadership quality, and strategic risk. These agents collaborate via a shared memory and inter-agent dialogue mechanism, enabling structured, multi-perspective analysis. By aggregating insights across roles, FIND-MA produces explainable, data-driven evaluations to support investors, analysts, and decision-makers. This work contributes toward building trustworthy, modular AI systems for financial due diligence and corporate valuation.

Keywords: retrieval-augmented generation, multi-agent systems, financial NLP, explainability

1 Introduction

In today's increasingly data-driven financial ecosystem, evaluating a company's intrinsic value requires analyzing a wide array of sources—ranging from financial statements and earnings call transcripts to news coverage and market sentiment. While large language models (LLMs) have recently shown promise in generating summaries and extracting information, their application to financial decision-making remains limited by issues of factual grounding, explainability, and modularity. Most existing tools either act as black-box summarizers or lack the structured reasoning required to support high-stakes evaluations.

Conventional single-agent systems and retrieval-based assistants—such as You.com and Perplexity AI—offer content aggregation but struggle with long-horizon consistency, interpretability, and role specialization. Even advanced frameworks like Grok Think and Grok DeepSearch provide improved retrieval and synthesis capabilities but fall short in coordinating distributed reasoning across multiple analytical dimensions, such as financial health, market dynamics, and managerial performance. These gaps restrict their utility in contexts where explainable, multi-perspective reasoning is essential.

To address these challenges, we introduce **FIND-MA** (**F**inancial **I**nsight via a **N**etwork of **D**istributed **M**ulti-**A**gents), a Retrieval-Augmented Generation (RAG)-based [6] multi-agent framework for fundamental company analysis. FIND-MA orchestrates a set of specialized agents, each responsible for a distinct analytical function—such as sentiment analysis, stock-price correlation, risk assessment, or news integration. These agents collaborate asynchronously via a

Project repository: github.com/Finance-LLMs/FIND-MA

shared memory and inter-agent dialogue protocol, enabling structured, modular, and explainable evaluations.

At the core of FIND-MA are state-of-the-art LLMs—**DeepSeek-R1** and **Qwen3** [16]—selected after empirical comparison for their superior performance in multi-hop reasoning, structured inference, and financial domain grounding. Each agent uses these models in role-specific contexts to extract insights from a curated retrieval base that includes five years of annual reports, earnings transcripts, and company news.

To support this pipeline, we manually analyzed and annotated data from 30+ companies across sectors including pharmaceuticals, telecommunications, technology, and defense. The dataset forms the retrieval backbone of the RAG system, enabling agents to ground their reasoning in verified, time-aligned, and sector-aware information.

FIND-MA produces structured outputs such as SWOT analyses, strategic risk profiles, and financial health assessments—each aligned with the source content and traceable to the contributing agents. The system has been evaluated through detailed case studies and human expert review, showing strong alignment with analyst expectations in terms of factual accuracy, interpretability, and coverage.

Contributions:

- We propose a novel RAG-based multi-agent framework that supports modular, explainable financial analysis through role-specialized agent collaboration.
- We develop and integrate agents powered by DeepSeek-R1 and Qwen3, each focused on distinct reasoning dimensions such as price trends, sentiment, and strategy.
- We construct a high-quality dataset of corporate reports and transcripts spanning five years and 30+ companies to ground agent reasoning in sector-specific evidence.
- We demonstrate, through case studies and expert evaluation, that FIND-MA delivers coherent, transparent, and decision-ready outputs for company evaluation.

2 Related Work

Retrieval-Augmented Generation in Financial Systems. Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for grounding language model outputs in factual knowledge. Lewis *et al.* [6] introduced RAG by integrating dense retrieval with generative modeling for knowledge-intensive NLP tasks. Tools such as **Perplexity AI** and **You.com** have adopted similar approaches for real-time query answering [6]. While these systems are effective in open-domain contexts, they lack domain grounding, long-context memory, and agent-level role separation essential for financial document understanding. Systems like **Grok DeepSearch** and **Grok Think** enhance structured retrieval but struggle with temporal coherence, multi-hop reasoning, and structured output synthesis—key challenges that FIND-MA addresses through inter-agent collaboration and retrieval-specific attention [15].

Financial Language Models and Analysis Tools. Domain-specific models such as **FinBERT** [1] have been fine-tuned on earnings calls and financial filings for sentiment and intent classification. More recently, **BloombergGPT** [14] demonstrated the utility of training large models on mixed financial and general corpora. While effective, these models typically operate in a single-pass, monolithic fashion without decomposition into analytical roles or interpretability layers. **Manus AI** offers a commercial-grade pipeline for financial and legal domains, yet its black-box nature and absence of agentic reasoning limit transparency.

Multi-Agent Reasoning Frameworks. Multi-agent coordination frameworks such as **CAMEL** [7], **AutoGPT**, and **ReAct** [17] have explored task decomposition using agent roles and inter-agent communication. These architectures enable dynamic reasoning and memory sharing but have been largely demonstrated on general tasks like puzzles, coding, or open-ended Q&A. Their application to finance remains underdeveloped, particularly in integrating RAG and generating role-specific explainable outputs. FIND-MA extends this line of work by grounding each agent's inference in curated financial documents and facilitating collaborative decision flows.

Explainable AI in Finance. Explainability in financial AI systems is essential due to the high stakes of investment decisions and regulatory requirements. Prior work on explainable AI (XAI) [2] focuses primarily on model introspection, attention visualization, or post-hoc rule extraction. However, these methods often fall short in capturing reasoning provenance in generative models. In contrast, FIND-MA embeds explainability into the system's architecture by allowing each agent to contribute modular, traceable insights supported by retrieved evidence and shared memory logs, thus aligning with the principles of transparent decision support.

Datasets for Financial NLP. Datasets such as `EDGAR`, `FinancialPhraseBank`, and `FinStatements` have supported advances in financial NLP, particularly in relation to earnings call analysis and sentiment prediction. However, these datasets lack multi-document alignment, sector diversity, and annotations for inter-agent synthesis. To overcome this, we curated a proprietary retrieval corpus comprising five years of annual reports, investor transcripts, and press coverage from over 30 companies across multiple sectors, supplemented by open-source financial corpora for broader generalization.

Our Contribution. FIND-MA is the first system to unify retrieval-augmented generation, LLM-based specialization, and decentralized multi-agent collaboration for explainable financial analysis. Built atop **DeepSeek-R1** and **Qwen3**, the system orchestrates reasoning across agents focused on specific dimensions—financial signals, market dynamics, sentiment, and risk. Outputs such as SWOT analyses and executive summaries are traceable, modular, and grounded in curated evidence, offering a scalable and transparent solution for analysts, investors, and policy advisors.

3 System Overview

FIND-MA (*Financial Insight via a Network of Distributed Multi-Agents*) is a modular framework designed to produce explainable, high-resolution company analyses by combining retrieval-augmented generation (RAG) with a coordinated multi-agent reasoning pipeline. Unlike monolithic LLM systems, FIND-MA decomposes the analysis into multiple specialized agents that operate over a shared evidence base and asynchronously contribute to a collective financial report.

Architectural Components

The system is composed of five primary components:

- **Retrieval Engine:** A FAISS-based dense vector retriever [3] indexes a curated corpus of annual reports, earnings transcripts, and financial news spanning five years and 30+ companies. Retrieved chunks are filtered by relevance, metadata, and temporal alignment.

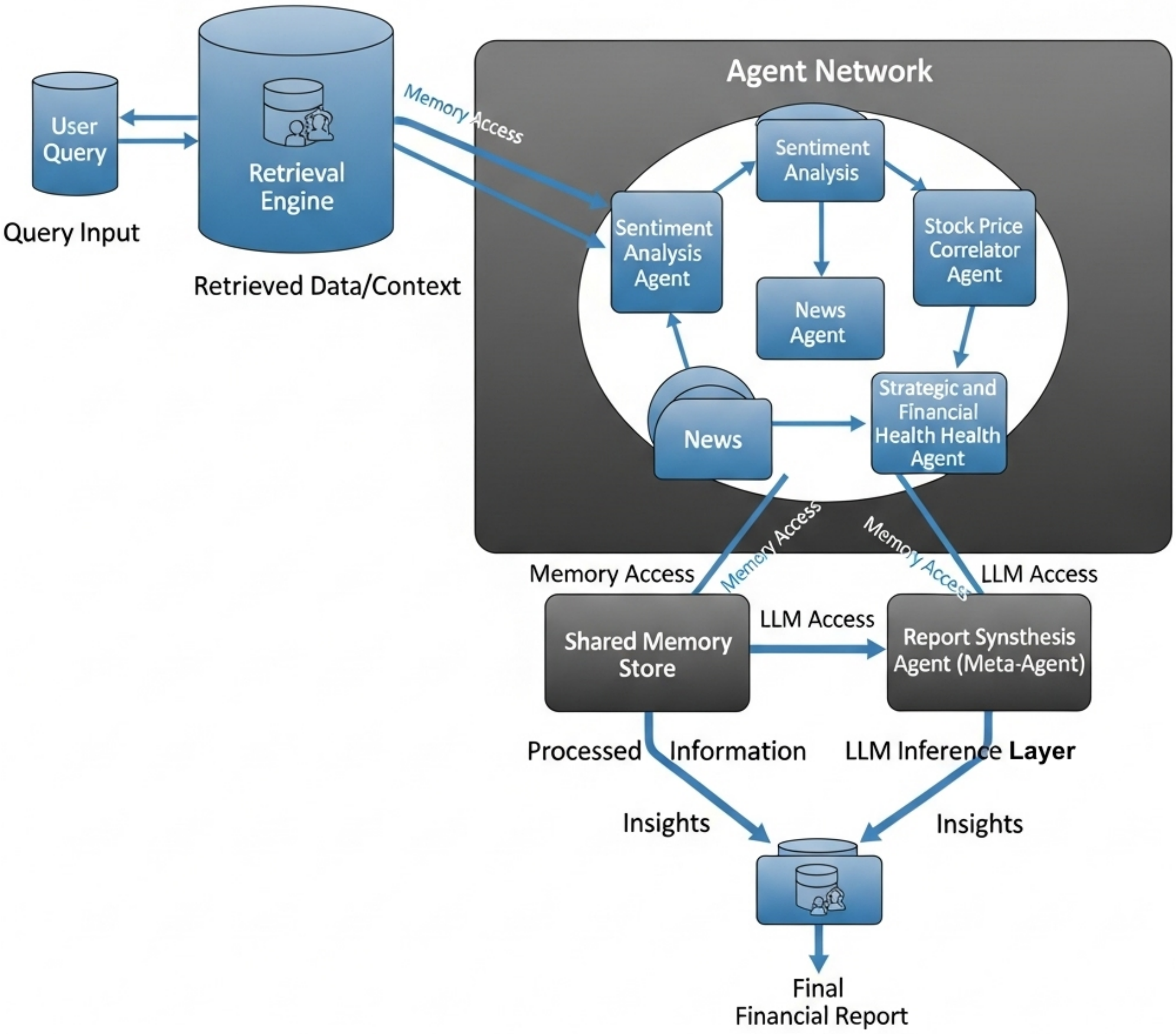

Figure 1: High-level architecture of FIND-MA. A Retrieval Engine feeds role-specialized agents (Sentiment, Stock Price Correlator, News, Strategic & Financial Health) operating over a Shared Memory Store. A Synthesis (meta-)agent compiles evidence-grounded insights into the final financial report.

- **Agent Network:** Each agent (e.g., Sentiment Agent, Stock Price Correlator, News Agent) is responsible for a specific analytical dimension. Agents use prompt templates and LLMs (DeepSeek-R1, Qwen3 [16]) to process retrieved context and generate role-specific insights.

- **Shared Memory Store:** A persistent memory module allows agents to write, read, and reference each other's outputs. This facilitates inter-agent communication, context sharing, and traceability.

- **LLM Inference Layer:** DeepSeek-R1 and Qwen3 serve as the generative engines behind each agent. These models are selected dynamically based on reasoning complexity, with DeepSeek used for multi-hop synthesis and Qwen for lightweight tasks, served efficiently via vLLM [5].

- **Report Synthesizer:** A final module compiles outputs from all agents into structured, human-readable reports. The reports include SWOT analyses, risk flags, sentiment summaries, and financial commentary—each linked to its generating agent and source evidence.

Design Philosophy

FIND-MA is built with three key principles:

1. **Explainability:** Each agent provides transparent outputs grounded in retrieved context, with reasoning steps accessible through shared memory logs.
2. **Modularity:** New agents can be added without system redesign, supporting extensibility to ESG, macroeconomic, or policy domains.
3. **Factual Grounding:** All generation is conditioned on a curated retrieval base, reducing hallucination and increasing decision traceability.

Pipeline Overview

Upon receiving a query (e.g., a company name or risk theme), FIND-MA performs evidence retrieval, launches agents in parallel or sequence, enables cross-agent updates via memory, and finally compiles the results. This mimics the behavior of a team of financial analysts operating collaboratively, with each agent contributing a distinct evaluative lens.

4 Methodology

FIND-MA decomposes the task of fundamental company analysis into specialized subtasks handled by independent agents. Each agent operates over a common retrieval base using its own reasoning strategy and language model configuration, coordinated via a structured memory and dialogue mechanism. This section outlines the core components of the system pipeline.

Input and Preprocessing

Given a company name or user query, FIND-MA first resolves the entity against its indexed metadata (e.g., sector, ticker, report availability). Document chunks relevant to the query are retrieved using dense semantic search (via SentenceTransformers/Sentence-BERT [11] and FAISS [3]). Each document is preprocessed through:

- **Chunking:** Long documents (e.g., annual reports) are split into overlapping windows of 300–500 words to preserve context.
- **Metadata Tagging:** Each chunk is labeled with company, year, source type (e.g., concall, filing), and document confidence.
- **Prompt Injection:** Retrieved text is embedded within prompt templates customized for each agent's task.

Retrieval-Augmented Generation (RAG) Layer

Each agent queries the FAISS index independently using a role-specific embedding. Retrieved chunks are incorporated into its generation prompt, enabling the agent to ground outputs in domain-specific evidence [6]. The RAG layer supports:

- Top-k semantic retrieval using cosine similarity.
- Optional re-ranking using keyword overlap or recency.
- Temporal filtering to avoid hindsight leakage.

Agent Reasoning Cycle

Each agent executes the following loop:

1. **Role Initialization:** Instantiated with its task (e.g., "analyze sentiment from concalls") and agent-specific prompt structure.
2. **Context Retrieval:** Chunks are selected based on query embedding and role filters.
3. **Inference:** The LLM (DeepSeek-R1 or Qwen3) processes the prompt and generates structured output, optionally using Chain-of-Thought/self-consistency/ToT [12, 13, 18].
4. **Memory Write-Back:** The output is logged in the shared memory store with metadata for later use by other agents or the report compiler.

Inter-Agent Dialogue and Shared Memory

Agents interact asynchronously via a centralized memory that stores all intermediate outputs. This enables:

- **Cross-Agent Referencing:** Agents may use prior outputs as input context, e.g., sentiment agent citing news events from the News Agent.
- **Conflict Resolution:** Discrepancies (e.g., bullish sentiment vs. declining stock) can be flagged for report synthesis logic.
- **Traceability:** Each output includes references to its source chunks and contributing agents, enabling full auditability.

[7, 15, 17]

Report Synthesis

The system compiles a structured report by aggregating the outputs of all agents. The report includes:

- **Executive Summary**
- **SWOT Analysis**
- **Financial Health Assessment**
- **Strategic Risk and Sentiment Trends**

Each section links to its underlying evidence and agentic source, providing users with both high-level insight and transparent provenance. The report can be rendered in markdown, JSON, or interactive web formats.

5 Agent Roles and Specialization

FIND-MA delegates distinct analytical responsibilities to a set of specialized agents, each equipped with its own prompt template, reasoning strategy, and retrieval filters. This modular architecture mirrors the functional decomposition of a human financial analyst team.

Sentiment Analysis Agent

This agent evaluates tone, optimism, and caution embedded in company communications, particularly earnings call transcripts and executive statements.

- **Sources:** Concall transcripts, press releases.
- **Capabilities:**
  - Speaker-aware analysis (CEO vs. analyst).
  - Temporal tracking of sentiment shifts across quarters.
  - Classification of language into bullish, neutral, or bearish tone.
- **Model:** DeepSeek-R1 for nuanced language inference.

Stock Price Correlator Agent

This agent connects textual events to movements in stock prices, aiming to generate plausible causal hypotheses.

- **Sources:** Earnings calls, filings, news; aligned with historical price data.
- **Capabilities:**
  - Aligns stock movement with event windows.
  - Hypothesis generation on price triggers.
  - Visual correlation summaries (planned).
- **Model:** DeepSeek-R1 with multi-hop reasoning prompts.

News Agent

This agent surfaces recent events and evaluates their potential impact on internal metrics and investor perception.

- **Sources:** Financial news, media coverage, press releases.
- **Capabilities:**
  - Relevance scoring of news items.
  - Linking public events to internal risks or opportunities.
  - Tone and credibility evaluation.
- **Model:** Qwen3 for fast summarization and entity matching.

Strategic and Financial Health Agent

This agent evaluates a company's capital structure, revenue trajectory, margins, debt, and strategic goals. It uses information from filings to assess overall business resilience.

- **Sources:** Annual reports, investor presentations.
- **Capabilities:**

  - Extracts financial metrics from MD&A sections.
  - Performs ratio-based assessments (planned).
  - Generates commentary on business outlook and strategic levers.
- **Model:** DeepSeek-R1 with tabular context-aware prompts (optional future extension for table parsing).

Report Synthesis Agent (Meta-Agent)

This agent collects and compiles the outputs from other agents into a cohesive final report. It ensures cross-agent coherence and traceability.

- **Sources:** Outputs from all agents.
- **Capabilities:**
  - Assembles summaries, SWOT tables, and strategic flags.
  - Flags conflicts (e.g., positive sentiment + declining price).
  - Renders human-readable and JSON-exportable outputs.
- **Model:** Qwen3 or DeepSeek-R1 depending on generation complexity.

6 Dataset

To ensure high factual grounding and domain alignment, FIND-MA operates over a curated, company-specific retrieval corpus comprising five years of real-world financial documents. This dataset enables retrieval-augmented agents to reason over structured evidence rather than relying solely on pre-trained model weights.

Motivation and Design Criteria

Existing financial NLP datasets such as `FinancialPhraseBank` [10] and finance lexicons like Loughran–McDonald [9] and `FinStatements` primarily support classification and extraction tasks but lack longitudinal structure, document diversity, and sector-specific metadata required for multi-agent reasoning. FIND-MA's dataset is designed to support:

- Multi-source grounding (annual reports, news, transcripts).
- Sector and company-level filtering for agent specialization.
- Time-aligned document indexing (to avoid hindsight leakage).
- Inter-agent retrieval relevance across overlapping contexts.

Corpus Composition

The dataset includes over 500 documents covering 30+ companies across diverse sectors such as pharmaceuticals, semiconductors, telecommunications, technology, and defence. Each company's profile includes:

- **Annual Reports (2019–2024):** MD&A sections, risk disclosures, strategy discussions.

- **Earnings Call Transcripts:** Annotated by quarter and speaker role (e.g., CEO, analyst).
- **Financial News Articles:** Timestamped and filtered for named entities and events.
- **Manually Authored Summaries:** Created by the research team to bootstrap grounding and fine-tune prompt design.

Metadata and Storage Format

Each document chunk is embedded with metadata to support targeted retrieval and filtering:

- `company_name`, `year`, `sector`, `source_type`
- `chunk_id`, `summary`, `source_url` (if public)

The corpus is stored as a collection of JSONL files and indexed using FAISS [3] for fast, dense retrieval. Embeddings are generated using Sentence-BERT/SentenceTransformers [11] fine-tuned on financial corpora to maximize semantic alignment.

Annotation and Quality Control

Document summaries and tag validation were performed manually to ensure retrieval precision. Each company profile was reviewed to confirm sector relevance, metadata correctness, and event coverage. Feedback from early agent runs was used to iteratively refine document chunking and prompt injection strategies.

7 System Implementation

FIND-MA is implemented as a modular, end-to-end system using open-source components and scalable orchestration layers. The architecture supports flexible deployment, rapid prototyping, and integration of multiple language models and agent roles.

Technology Stack

The core system is implemented in Python, with key components including:

- **LLM Inference:** DeepSeek-R1 and Qwen3 [16] are accessed via Hugging Face Transformers and vLLM [5] inference backends. Prompt templates are customized per agent.
- **Embedding and Retrieval:** SentenceTransformers is used to encode document chunks, which are stored and queried via a FAISS [3] vector index.
- **Memory Layer:** An in-memory Python dictionary is used to simulate shared memory during development; Redis is used for distributed memory in scaled settings.
- **Prompt Orchestration:** Agents communicate through structured prompt templates using Jinja2, allowing role-specific prompt injection and formatting.
- **Interface (Planned):** A Streamlit-based frontend enables user interaction for company selection, report generation, and visualization of agent outputs.

Agent Lifecycle

Each agent operates independently following a standardized lifecycle:

1. **Initialization:** The agent receives a task role and company context.
2. **Query Embedding and Retrieval:** The agent generates a query vector and retrieves top-k chunks from FAISS.
3. **Prompt Construction:** A role-specific prompt is assembled using retrieved content and memory from prior agents.
4. **Model Inference:** The selected LLM is invoked with the prompt and returns an answer structured in markdown or JSON.
5. **Memory Write-Back:** Output is added to shared memory with metadata tags for traceability.

Execution Model

FIND-MA supports both parallel and sequential agent execution. Parallel execution is used for agents with independent scopes (e.g., News and Sentiment Agents), while sequential chaining is used where inter-agent context is essential (e.g., Synthesis Agent depending on others' outputs). Logs and timestamps are recorded at each stage for latency analysis and debugging.

Scalability and Extensibility

The architecture is designed for extensibility:

- New agents (e.g., ESG or Macroeconomic Agents) can be added by defining their retrieval scope, prompt logic, and output format.
- Model routing can be optimized via lightweight benchmarking to dynamically select between DeepSeek-R1 and Qwen3.
- A RESTful API is under development to enable headless deployments and external integrations.

Deployment Plans

For user-facing interaction, a Streamlit-based dashboard is being developed. It will allow:

- Query input and company selection,
- Visualization of agent-wise outputs,
- Export of structured reports (JSON, PDF).

This implementation strategy supports a seamless transition from research prototype to production-grade tool for analysts and non-technical users alike.

8 Evaluation and Results

We evaluate FIND-MA along two key axes: (1) the quality and interpretability of agent-generated insights, and (2) the coherence and completeness of the synthesized financial reports. The evaluation combines structured human feedback with case study-based benchmarking across multiple sectors.

Evaluation Objectives

Our goals are to assess whether:

- Each agent produces relevant, well-grounded outputs from retrieved evidence.
- Inter-agent coordination yields consistent and multi-perspective evaluations.
- Final reports are actionable, interpretable, and aligned with human analyst expectations.

Case Study Setup

We ran FIND-MA on 20 companies across six sectors—technology, pharmaceuticals, telecom, defense, healthcare, and semiconductors. For each case, the system generated a full report based on the past five years of filings and related news. Each report was reviewed by human evaluators with backgrounds in finance or business analytics.

Metrics

We adopted both quantitative and qualitative measures:

- **Factual Accuracy (1–5):** Alignment of agent insights with ground-truth filings and events.
- **Interpretability (1–5):** Clarity and traceability of reasoning paths and outputs.
- **Relevance (1–5):** Applicability of outputs to company context and time frame.
- **Cross-Agent Consistency:** Agreement between agents on overlapping concepts.
- **Latency (s):** End-to-end report generation time on an A100 GPU instance.

Results Summary

- **Factual Accuracy:** 4.3 ± 0.4 — Agent outputs were strongly grounded in retrieval context.
- **Interpretability:** 4.5 ± 0.3 — Modular outputs and role-specific prompts contributed to clarity.
- **Relevance:** 4.2 ± 0.5 — Most outputs directly addressed company-specific conditions or strategic inflection points.
- **Cross-Agent Consistency:** 86% agreement — Minimal contradictions, with some trade-off between sentiment tone and stock movement patterns.
- **Latency:** 42.7 seconds average — Full report generation took under a minute for most runs, including all agent passes.

Qualitative Feedback

Human reviewers praised:

- The transparency of agent-level outputs.
- The interpretability of SWOT and sentiment summaries.
- The factual grounding in real, time-aligned documents.

Limitations were noted in:

- Table-based metric extraction (e.g., parsing balance sheet figures).
- Sensitivity to ambiguous language in transcripts.
- Occasional verbosity in the synthesis agent's summarizations.

Ablation and Comparative Observations

To assess the effect of agent modularity, we compared FIND-MA with a single-prompt baseline using DeepSeek-R1 across all retrieved documents. FIND-MA's agent-based pipeline showed:

- Better traceability and explanation clarity.
- Higher inter-reviewer agreement on accuracy and usefulness.
- Lower hallucination due to role-specific prompt control.

9 Future Work and Conclusion

Future Work

While FIND-MA demonstrates the potential of multi-agent LLM frameworks in financial analysis, several directions remain for future enhancement:

- **Expanded Agent Network:** Introduce new agents such as an *ESG Agent*, a *Macroeconomic Impact Agent*, and a *Policy and Regulation Agent* to expand analytical coverage across sustainability, governance, and external policy factors.
- **Advanced Table Parsing and Visual Input Handling:** Current document parsing does not robustly handle numerical tables or embedded charts. Integration of tools like Donut and ChartQA [4, 8] could enhance the system's ability to interpret financial ratios and graphical trends.
- **Human-in-the-Loop Evaluation:** A structured feedback interface will allow analysts and retail investors to rate, revise, and interact with agent outputs—providing fine-tuning signals for future model improvements.
- **Interactive Deployment:** A fully deployed version of FIND-MA with a Streamlit frontend is under development. This interface will support company queries, agent-wise insight display, and export of structured reports for end users.
- **Scalability and Cross-Market Extension:** Evaluate FIND-MA on companies from non-English markets and domains beyond public equities—such as startups, IPOs, and ESG disclosures—to assess cross-linguistic and cross-sector generalizability.

Conclusion

We presented **FIND-MA**, a Retrieval-Augmented, Multi-Agent Framework for fundamental company analysis. By orchestrating role-specific agents—each grounded in domain-specific evidence and powered by state-of-the-art LLMs like DeepSeek-R1 and Qwen3—FIND-MA produces explainable, modular, and traceable evaluations. The system operates over a curated dataset of annual reports, earnings transcripts, and financial news spanning five years and 30+ companies, supporting grounded and high-fidelity reasoning.

Through human evaluation, case studies, and ablation analysis, we demonstrate that FIND-MA significantly improves the interpretability and relevance of financial insights compared to single-prompt baselines. Its architecture provides a scalable, extensible foundation for the next generation of trustworthy AI tools in financial services. As financial analysis becomes more complex and data-rich, systems like FIND-MA offer a promising path toward democratizing deep reasoning—bridging human expertise with the inferential power of large language models.

References

[1] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[2] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[3] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. In *IEEE Transactions on Big Data*, 2017. arXiv:1702.08734.

[4] Geewook Kim, Teakgyu Hong, et al. Donut: Document understanding transformer without ocr. In *ECCV*, 2022.

[5] Woosuk Kwon, Jinyoung Bae, et al. vllm: Easy, fast, and cheap llm serving with pagedattention. *Proceedings of SOSP*, 2023. arXiv:2309.06180.

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[7] Weiming Li, Yuxuan Cheng, Daochen Zha, Qi Zhang, Zhewei Ding, et al. Camel: Communicative agents for mind exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023.

[8] Yulei Liu, Zenglin Tang, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *AAAI*, 2022.

[9] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[10] Pekka Malo, Ankur Sinha, et al. Good debts or bad debts: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 2014. FinancialPhraseBank.

[11] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, 2019.

[12] Xuezhi Wang, Jason Wei, et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. arXiv:2203.11171.

[13] Jason Wei, Xuezhi Wang, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[14] Shijie Wu, Ramesh Menon, et al. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[15] Xuehai Wu, Yizhou Li, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[16] An Yang, Jinze Bai, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[17] Shinn Yao, Jeffrey Zhao, Kaixin Yu, Weijia Chen, et al. React: Synergizing reasoning and acting in language models. *Advances in Neural Information Processing Systems*, 2023.

[18] Shunyu Yao, Dian Yu, et al. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023. arXiv:2305.10601.