IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

Finlet AI: An Evidence-Driven Multi-Agent Financial Intelligence Platform Using Model Context Protocol and Deterministic Risk Governance

Sanika Sarode, Purva Shinde, Renuka Sahane, and Rutuja Sonawane
Department of Artificial Intelligence and Data Science
Jaihind College of Engineering, Kuran
sanikasarode166@gmail. com, purvasandeshshinde11@gmail. com
renukasahane992@gmail. com, rutuj asonawane2005@gmail. com

Abstract: *Standalone conversational language models struggle to satisfy audit and compliance standards in institutional equity research. When an equity analyst prompts an ungrounded model to evaluate a listed company, three operational defects appear almost immediately: the model quotes outdated share prices from its pre-training cutoff, invents page or footnote numbers when summarizing regulatory filings, and produces inconsistent numerical risk ratings when executed repeatedly on identical inputs. This paper presents Finlet AI, a twelve-layer multi-agent financial intelligence platform that decouples quantitative market arithmetic and primary-source retrieval from neural text synthesis. Once an analytical query is submitted through our asynchronous FastAPI gateway, a stateful LangGraph supervisor triggers concurrent parallel fan-out across three specialized domain workers covering historical price series, live financial news feeds, and statutory annual reports isolated behind standardized Model Context Protocol (MCP) tool interfaces. Technical price action, 20-day volume momentum ratios, and 30-day annualized volatility (σ) are computed directly from three-month 30d Open-High-Low-Close-Volume (OHLCV) series fetched via Yahoo Finance. Real-time news catalysts are gathered through Google News RSS syndication and sanitized of HTML markup, while balance-sheet debt and liquidity disclosures are extracted page by page from SEC Form 10-K and Indian annual reports indexed locally in ChromaDB using 384-dimensional all-MiniLM-L6-v2 sentence embeddings. Rather than delegating risk scoring to subjective prompt completion, our Risk & Governance module evaluates a bounded mathematical index weighting volatility (45% weight), news sentiment risk (35% weight), and filing flags (20% weight). Live validation on Reliance Industries (RELIANCE.NS) demonstrates real-time tracking at ₹1,184.50 (−2.03%), 23.1% annualized volatility, 1.0x volume momentum, and a deterministic risk score of 0.365 (LOW), while high-risk runs (⩾ 0.65) pause at a persistent SQLite Human-in-the-Loop (HITL) checkpoint before compiling the final ReportLab PDF dossier..*

Keywords: Multi-Agent Systems, Finlet AI, Financial Intelligence, Model Context Protocol (MCP), LangGraph, ChromaDB, Retrieval-Augmented Generation (RAG), Deterministic Risk Governance, Human-in-the-Loop (HITL).

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40710

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

I. INTRODUCTION

A. Institutional Verification Standards and LLMDefects

Equity research desks operate under strict verification mandates. Every valuation multiple, realized volatility figure, or debt covenant remark written into an investment committee briefing must trace back to an exchange price tape, a timestamped news wire, or an exact page inside an audited annual filing [2], [6]. Standard conversational language models do not naturally respect these boundaries. Because an autoregressive transformer predicts tokens sequentially from static pre-trained weights, it treats exact market arithmetic no differently than open-ended prose generation. In everyday writing, a minor numerical drift rarely causes harm. In financial analysis, however, a fabricated balance-sheet metric or an unverified debt claim can invalidate an entire investment memo and expose a firm to severe regulatory liability.

When we tested single-prompt foundation models across U.S. (NASDAQ/NYSE) and Indian (NSE) stock symbols during the initial phase of our project, three recurring engineering problems stood out. First, frozen model weights are blind to live market action; an ungrounded model has no way to know whether Reliance Industries (RELIANCE.NS) closed at ₹1 ,184.50 (−2.03%) or whether its 30-day annualized volatility settled at 23.1 % [3]. Second, asking a conversational model to rate company risk on a numerical scale produced unacceptable run-to-run variance even at low sampling temperatures. Third, when summarizing hundred-page regulatory PDFs, standard document loaders merged pages together before chunking, causing ungrounded models to cite the wrong SEC Form 10-K section or blend accounting figures across different fiscal years, consistent with empirical findings on the FinanceBench benchmark [6].

B. Architectural Philosophy ofFinlet AI

We took a fundamentally different engineering approach in Finlet AI. Rather than treating the language model as an oracle expected to recall stock prices and invent risk ratings from memory, we removed all quantitative arithmetic and risk scoring from the prompt layer. In our twelve-layer architecture, the language model serves strictly as a grounded text synthesizer over evidence payloads that have already been fetched, verified, and schema-checked by deterministic Python modules.

Market price bars, rolling volatility (σ30d), and moving-average trends are calculated through Yahoo Finance (yfinance) adapters following the Model Context Protocol (MCP) [13]. Real-time headlines are pulled from Google News RSS feeds and sanitized via BeautifulSoup4, while corporate liquidity disclosures are retrieved page by page from local ChromaDB vector indices [4], [5]. All three evidence streams must satisfy strict Pydantic v2 data contracts before our deterministic Risk & Governance engine combines them using calibrated weights (45% volatility, 35% sentiment risk, 20% filing flags) into a composite score on [0.000, 1 .000], halting at a persistent SQLite Human-in-the-Loop (HITL) gate whenever the score reaches or exceeds 0.65.

Upstream Client & Gateway	Parallel MCP Domain Fan-Out	Deterministic Risk & HITL Governance
Streamlit UI Console: Analyst Query & KPIs FastAPI Gateway: Async REST Endpoints LangGraph Supervisor: Parallel Fan-Out via MCP	Market Agent (MCP): yfinance 3mo OHLCV, σ30d, Vol Ratio, Trend News Agent (MCP): Google News RSS + BS4, Verified URLs & Polarity S Document Agent (MCP): pypdf + ChromaDB RAG, Page-Cited 10-K Risk Flags	Risk & Governance Node: 0.45 Vnorm + 0.35 Snorm + 0.20 Dnorm (Deterministic Audit Trail) Low/Med (< 0.65): Direct to ReportLab Dossier High (≥ 0.65): SQLite HITL Checkpoint Gate → Approved → ReportLab Dossier & SQLite Provenance Log

Fig. 1 . End-to-end twelve-layer architecture of Finlet AI showing the Streamlit analyst console, FastAPI gateway, LangGraph parallel fan-out across Market, News, and Document MCP agents, fan-in convergence at the deterministic Risk & Governance matrix (0.45/0.35/0.20 weights), and SQLite HITL checkpointing.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

C. Core Engineering Contributions

The primary contributions ofthis journal paper are fourfold:

1. Twelve-Layer MCP-Isolated Multi-Agent Topology: We design a modular equity research stack that decouples LangGraph state orchestration from external market, news, and document data feeds using Model Context Protocol (MCP) tool wrappers and strict Pydantic v2 validation contracts.
2. Zero-Cost Cross-Market Parallel Retrieval Pipeline: We implement concurrent parallel fan-out across Yahoo Finance (yfinance) OHLCV analytics, Google News RSS syndication, and page-preserving ChromaDB vector search (all-MiniLM- L6- v2), supporting both U.S. (NASDAQ/NYSE) and Indian (NSE) listings without commercial API subscriptions.
3. Deterministic Three-Factor Risk & Governance Engine: We formulate a bounded, repeatable risk equation weighting normalized volatility (45% weight), news sentiment risk (35% weight), and regulatory filing flags (20% weight), complete with an inline factor decomposition audit trail that eliminates stochastic LLM rating variance.
4. Live Empirical Verification on Indian and U.S. Equities: We integrate real-time production outputs from our Finlet AI dashboard—demonstrating live execution on Reliance Industries (RELIANCE. NS) alongside cross-market benchmarks—and evaluate latency reduction, score repeatability, and stateful SQLite Human-in-the-Loop (HITL) governance.

II. RELATED WORK AND LITERATURE SURVEY

A. Domain-Specific Financial Language Models

Early adaptation of transformer models to financial natural language processing focused on bidirectional encoders trained for sentiment classification. Araci [1] developed FinBERT by continuing BERT pre-training on financial news corpora and the Financial PhraseBank dataset, achieving strong gains in three-class polarity labeling over general-domain encoders.

As generative transformers scaled, Wu et al. [2] trained BloombergGPT, a 50-billion-parameter decoder model trained on 363 billion financial tokens combined with 345 billion general tokens. While BloombergGPT demonstrated strong performance on internal financial benchmarks, its closed weights and multi-million-dollar compute budget prevent academic replication or local deployment. To lower the entry barrier for research labs, Yang et al. [3] introduced FinGPT, an open-source framework using parameter-efficient Low-Rank Adaptation (LoRA) on open foundation models. Although FinGPT makes fine-tuning affordable, offline weight updates alone cannot solve live equity research: a static adapter cannot compute today's 30-day rolling volatility (σ30d) from live NSE or NASDAQ price bars or guarantee exact page citations on newly released annual reports.

B. Financial RAG and Citation Provenance

Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [4], grounds text generation in external documents by retrieving relevant passages from a dense vector index. Most local RAG pipelines rely on Siamese bi-encoder networks such as Sentence-BERT by Reimers and Gurevych [5], which map text passages into fixed-dimensional vectors (such as the 384-dimensional all- MiniLM- L6- v2 embedding space) for rapid cosine-similarity lookup on CPU hardware.

Applying default RAG text splitters to corporate regulatory filings introduces severe provenance gaps, however. When Islam et al. [6] evaluated LLM-RAG architectures across 10,231 financial questions on U.S. SEC filings in the FinanceBench benchmark, they observed that standard document loaders concatenate all pages into a single string before chunking. That practice destroys page boundaries, strips multi-page table headers, and prevents analysts from verifying which page a balance-sheet figure came from. Finlet AI directly resolves this limitation by extracting PDF text strictly page by page via pypdf and stamping every vector chunk in ChromaDB with an immutable metadata dictionary containing the ticker symbol, source filename, and integer page number.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

TABLE I: COMPARATIVE BENCHMARKING OF FINLET AI AGAINST EXISTING FINANCIAL LLM AND MULTI-AGENT SYSTEMS

System / Framework	Live OHLCV Math	Filing Provenance	MCP Layer	Risk Scoring Mechanism	Stateful HITL Gate	API Cost
FinBERT [1], BloombergGPT [2]	No (Frozen Weights)	None	No	Static Polarity / Token Gen	No	Closed / Paid
FinGPT [3]	No (LoRA Adapter)	Unverified Snippets	No	Probabilistic Prompt Output	No	Variable
FinMem [9], FinAgent [10]	Historical Series	Partial Memory	No	Persona / Neural Reflection	No	Paid Quotas
TradingAgents [11], FINCON [12]	External Wrappers	Summary Level	No	Multi-Agent Verbal Debate	No	Paid APIs
Finlet AI (Proposed Platform)	Yes (yfinance)	Exact Page Index	Yes (MCP)	Deterministic (0.45/0.35/0.20)	Yes (SQLite $\geq$ 0.65)	100% Free

C. Multi-Agent Financial Frameworks

To move beyond single-pass text generation, Yao et al. [7] proposed ReAct, which interleaves verbal reasoning traces with external tool calls, and Wu et al. [8] developed AutoGen to coordinate multi-agent workflows across specialized roles. Several recent studies have applied multi-agent LLM architectures to stock analysis. FinMem [9] equips a trading agent with layered short- and long-term memory modules and tunable risk personas. FinAgent [10] combines multimodal price chart images with news text and tool-augmented reflection. TradingAgents [11] and FINCON [12] simulate institutional trading desks by assigning distinct LLM agents to fundamental, technical, sentiment, and risk-management roles.

As shown in Table I, existing multi-agent trading frameworks rely on natural-language debate or verbal reinforcement between LLM agents to evaluate portfolio risk. Because LLM debate is inherently probabilistic, running the exact same market state twice can still yield different risk classifications. In addition, prior frameworks hard-code custom wrappers around paid market APIs. Finlet AI advances this landscape by combining standardized Model Context Protocol (MCP) tool isolation [13], a zero-cost data stack covering both U.S. and Indian markets, a repeatable mathematical risk engine, and stateful LangGraph checkpoint interrupts [14] for mandatory Human-in-the-Loop governance.

III. TWELVE-LAYER ARCHITECTURE AND DATA CONTRACTS

A. Asynchronous Multi-Agent Execution Pipeline

Fig. 1 depicts the end-to-end execution flow of Finlet AI across our twelve-tier software stack. Separating the Streamlit analyst console, FastAPI REST gateway, LangGraph state-machine orchestrator, MCP protocol interface, domain workers, compliance gate, and SQLite audit store prevents tight coupling and simplifies fault isolation. An investigation begins at Layer 1 when an equity analyst selects a target company (such as Reliance (RELIANCE. NS) , AAPL, or NVDA) in the Streamlit console (Layer 2) and clicks Run Finlet Analysis.

Streamlit dispatches an asynchronous HTTP request to the FastAPI gateway (Layer 3). FastAPI initializes a unique j ob_id and launches the LangGraph StateGraph supervisor (Layer 4) in a background worker thread. Using parallel routing edges from the graph START node, the supervisor dispatches the Market Agent, News Agent, and Document Agent (Layers 6–8) concurrently via parallel fan-out through the standardized MCP Protocol Layer (Layer 5). Once all three concurrent domain agents complete execution and populate their validated slices of the shared workflow state, their outputs converge (fan-in) at the Risk & Governance Agent (Layer 9).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, November 2026 Impact Factor: 8.2

If the composite risk score falls below 0.65 (LOW or MEDIUM risk bands), the orchestrator routes execution directly to the Report Agent (Layer 11), which compiles the final PDF dossier via ReportLab and logs the completed run in SQLite (Layer 12). Conversely, if the composite risk score reaches or exceeds 0.65 (HIGH risk band), Layer 10 intercepts the workflow, serializing the entire graph state to SQLite under a unique thread_id and returning a PENDING status to the Streamlit dashboard until explicit human approval or rejection is submitted.

B. Pydantic v2 Inter-Agent Schema Contracts

Passing untyped Python dictionaries or raw JSON strings between concurrent agent nodes is a frequent source of runtime failures in multi-agent systems. Finlet AI eliminates schema drift by enforcing strict Pydantic v2 models (schemas. py) across all MCP tool returns and LangGraph state transitions:

Pydantic v2 Inter-Agent Contract Specification (schemas.py)

EvidenceRecord: `source_type` ∈ {`'filing'`, `'news'`, `'market_data'`}, `source_identifier: str` (URL, 10-K section, or ticker), `timestamp: datetime`, `content_snippet: str`, `page_number: Optional[int]`.

MarketAnalysis: `symbol: str`, `current_price: float`, `daily_change_pct: float`, `annualized_volatility_30d: float` ($\sigma_{30d} \geq 0.0$), `volume_momentum: float`, `trend` ∈ {`'BULLISH'`, `'BEARISH'`, `'NEUTRAL'`}, `evidence: List[EvidenceRecord]`.

NewsAnalysis: `symbol: str`, `sentiment_score: float` ($-1.0 \leq S \leq +1.0$ via `ge=-1.0, le=1.0`), `sentiment_label: str`, `articles: List[EvidenceRecord]`.

DocumentAnalysis: `symbol: str`, `identified_risks: List[str]`, `filing_flags: List[str]` (N_{flags}), `evidence: List[EvidenceRecord]`.

TABLE II: TWELVE-LAYER ARCHITECTURAL SPECIFICATION OF THE FINLET AI PLATFORM

Layer	System Subsystem	Core Technology	Runtime Responsibility & Verification Scope
Layer 1	Analyst Client	Web Browser	Selects target equity ticker (e.g., RELIANCE.NS, AAPL, NVDA) and inspects audit trails.
Layer 2	Finlet AI Console	Streamlit, Plotly	Renders KPI cards, 3-month candlestick charts, news verification accordions, and risk banners.
Layer 3	Async API Gateway	FastAPI, Uvicorn	Exposes nonblocking REST endpoints for job creation, status polling, HITL approval, and PDF download.
Layer 4	Workflow Supervisor	LangGraph StateGraph	Coordinates concurrent parallel fan-out across Market, News, and Document agents and fan-in convergence.
Layer 5	MCP Tool Protocol	Model Context Protocol	Isolates reasoning agents from external APIs using typed tool contracts and structured fault handlers.
Layer 6	Market Agent	yfinance, NumPy	Fetches 3-month OHLCV series; computes current price, σ_{30d} volatility, volume momentum, and trend.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40710

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, November 2026 Impact Factor: 8.2

Layer 7	News Agent	Google News RSS, BS4	Retrieves real-time articles, strips HTML markup, verifies publisher links, and scores sentiment S.
Layer 8	Document Agent	pypdf, ChromaDB	Indexes 10-K and annual report PDFs page by page using 384-d all-MiniLM-L6-v2 embeddings.
Layer 9	Risk & Governance	Deterministic Python	Computes $R_{\text{total}} = 0.45\ V_{\text{norm}} + 0.35\ S_{\text{norm}} + 0.20\ D_{\text{norm}}$ and emits the factor audit trail.
Layer 10	HITL Checkpoint Gate	LangGraph SQLite Saver	Halts workflows when Rtotal ≥ 0.65, persisting full graph state until human analyst sign-off.
Layer 11	Executive Report Agent	ReportLab PDF Engine	Compiles verified KPIs, candlestick summaries, clickable news URLs, and filing citations into PDF.
Layer 12	Audit Persistence Store	SQLite Database	Stores immutable job logs, timestamps, raw MCP payloads, factor decompositions, and reviewer notes.

IV. MCP DOMAIN ADAPTERS AND FINANCIAL RAG

A. Quantitative Market Agent Formulation

To support both U.S. equities and National Stock Exchange of India symbols (. NS suffix) at zero API cost, our Market Agent queries Yahoo Finance (yfinance) through an MCP adapter to fetch three months of daily OHLCV bars. Let Pt and Vt denote the closing price and volume on trading day t ∈ {1 , 2, ⋯, T} . The single-session percentage change ∆ P% is computed from the two latest trading sessions:

$$\Delta P\% = \text{round}(((PT - PT{-}1) / PT{-}1) \times 100, 2)\ (1)$$

To measure realized price instability, daily logarithmic returns rt = ln(Pt / Pt−1) are calculated across consecutive sessions. Over a trailing window of W = 30 trading days, the sample standard deviation of logarithmic returns is annualized by multiplying by √252 trading days per year:

$$\sigma 30d = \sqrt{252} \cdot \sqrt{[\ (1 / (W{-}1))\ \Sigma k{=}T{-}W{+}1T\ (rk - \bar{r}30d)2\ ]}\ (2)$$

Abnormal institutional volume participation is captured by the 20-day volume momentum ratio VR20d, defined as the ratio oflatest session volume VT to the 20-day Simple Moving Average (SMA) ofvolume:

$$VR20d = VT\ /[\ (1 / 20)\ \Sigma T\ Vk\ ]\ (3)$$

B. Real-Time News Agent & Page-Preserving RAG

Our News MCP Adapter queries Google News RSS over HTTPS, parses the XML feed via feedparser, and strips embedded HTML tags using BeautifulSoup4 (html. parser). Each retrieved article i ∈ {1 , ⋯, K} retains its exact publication timestamp, publisher name (e.g., Yahoo Finance UK, The Times of India, The Guardian, TradingView, The Economic Times), and direct source URL. Aggregate sentiment polarity S ∈ [−1 .0, +1 .0] is evaluated across the top K = 5 articles:

$$S = (1 / K)\ \Sigma i{=}1K\ (p + - p -)\ (4)$$

In parallel, the Document Agent inspects statutory disclosures indexed in ChromaDB using our page-preserving ingestion routine formalized in Algorithm 1 :

Algorithm 1: Page-Preserving PDF Ingestion & ChromaDB RAG

Input: Filing PDF path F_{pdf}, ticker T_{sym}, chunk size $L_c = 1000$, overlap $L_o = 150$.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, November 2026 Impact Factor: 8.2

Output: Indexed vector chunk count N_{chunks} in persistent ChromaDB.

1 Initialize ChromaDB client and local bi-encoder `all-MiniLM-L6-v2` ($\mathbb{R}^{384}$).

2 Open `PdfReader(F_pdf)` via `pypdf`; initialize lists D, M, I.

3 **for** each page index $p \in \{1, 2, \ldots, |P|\}$ **do**

4 Extract text $\tau_p \leftarrow$ `ExtractText(P[p])` strictly within page p.

5 Split τ_p into chunks $\{c_{p,1}, \ldots, c_{p,m}\}$ using ($L_c = 1000$, $L_o = 150$) and stamp each chunk metadata with {`symbol`: T_{sym}, `page_number`: p}.

6 **end for** and commit embeddings to persistent ChromaDB collection.

Current Price	30d Annual Volatility	Volume Momentum	News Sentiment
₹1,184.50 (−2.03%)	23.1% (σ30d = 0.231)	1.0x (BEARISH Trend)	NEUTRAL (Score: 0.1)
Market Agent: Technical Price Action & 3-Month Trend: Interactive candlestick series tracking RELIANCE. NS from July 26, 2026 (₹1 ,300–₹1 ,335) through Aug 9, Aug 23, Sep 6, Sep 20, Oct 4, and Oct 18, 2026 (consolidating in the ₹1 ,160–₹1 ,210 band).			

Fig. 2. Live Finlet AI Dashboard execution for Reliance Industries (RELIANCE. NS). The top KPI strip displays Current Price (₹1 ,184.50, −2.03%), 30d Annual Volatility (23.1 %), Volume Momentum (1 .0x, BEARISH), and News Sentiment (NEUTRAL, Score: 0.1), followed by the Market Agent's 3-Month interactive candlestick chart from July 26, 2026 to October 18, 2026.

V. DETERMINISTIC RISK & GOVERNANCE ENGINE

A. Mathematical Formulation ofFactor Decomposition

To guarantee 100% score repeatability across runs, Finlet AI computes its composite risk score Rtotal ∈ [0.000, 1 .000] through a deterministic three-factor mathematical formula rather than LLM text generation. Each upstream domain metric is first mapped onto a normalized [0.0, 1 .0] scale:

1) Volatility Risk Factor (Vnorm, 45% Weight): The 30-day annualized volatility σ30d is normalized against a 50% institutional high-volatility ceiling (σmax = 0.50) and clamped to [0.0, 1 .0]:

$$\text{Vnorm} = \min(1.0, \sigma 30d / 0.50) \quad (5)$$

2) Sentiment Risk Factor (Snorm, 35% Weight): The aggregate news sentiment score S ∈ [−1 .0, +1 .0] is inverted and mapped linearly onto [0.0, 1 .0] so strongly positive news (S = +1 .0) yields 0.0, neutral news (S = 0.0) yields 0.50, and strongly negative news (S = −1 .0) yields 1 .0:

$$\text{Snorm} = (1.0 - S) / 2.0 \quad (6)$$

3) Filing Flags Factor (Dnorm, 20% Weight): Each adverse liquidity deficit, debt covenant warning, or litigation flag Nflags extracted from regulatory filings contributes 0.25 up to a saturation ceiling of four flags:

$$\text{Dnorm} = \min(1.0, 0.25 \times \text{Nflags}) \quad (7)$$

The Risk & Governance module combines these three normalized factors using calibrated institutional weights of 45% for market volatility, 35% for news sentiment risk, and 20% for statutory filing flags:

$$\text{Rtotal} = \text{round}(0.45 \cdot \text{Vnorm} + 0.35 \cdot \text{Snorm} + 0.20 \cdot \text{Dnorm}, 3) \quad (8)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, November 2026 Impact Factor: 8.2

B. Stateful LangGraph Fan-Out and HITL Checkpointing

We model the multi-agent workflow as a stateful directed graph G = (V, E) in LangGraph (StateGraph(AgentState)) [14]. Rather than executing data collection sequentially, once the FastAPI gateway initializes the job at the START node, parallel edges trigger concurrent fan-out across all three domain nodes (vmkt, vnews, vdoc), which execute simultaneously and converge (fan-in) at the compliance node vrisk:

$$E_{fan\text{-}out} = \{START\} \times \{vmkt, vnews, vdoc\}, E_{fan\text{-}in} = \{vmkt, vnews, vdoc\} \times \{vrisk\} \quad (9)$$

Algorithm 2: Deterministic Risk Evaluation & Conditional HITL Gate

Input: Converged state (σ_{30d}, S, N_{flags}) from parallel MCP agents.

Output: Risk score R_{total}, audit trail string, and next node transition.

1. Compute $V_{norm} \leftarrow \min(1.0, \sigma_{30d} / 0.50)$ // 45% weight
2. Compute $S_{norm} \leftarrow (1.0 - S) / 2.0$ // 35% weight
3. Compute $D_{norm} \leftarrow \min(1.0, 0.25 \times N_{flags})$ // 20% weight
4. Evaluate $R_{total} \leftarrow \text{round}(0.45 \cdot V_{norm} + 0.35 \cdot S_{norm} + 0.20 \cdot D_{norm}, 3)$
5. Emit audit trail: `"Volatility factor: V_norm (45% wt), Sentiment risk: S_norm (35% wt), Filing flags: D_norm (20% wt)"`.
6. **if** $R_{total} \geq 0.65$ **then** classify `'HIGH'`; serialize state to SQLite under `thread_id` and pause for senior analyst sign-off (`'PENDING'`).
7. **else if** $R_{total} \geq 0.40$ **then** classify `'MEDIUM'`; log caution badge in PDF.
8. **else** classify `'LOW'`; auto-route to report agent for PDF generation.

Because LangGraph persists state checkpoints in SQLite under a unique thread_id, halting a high-risk run frees all worker threads while preserving the exact market, news, and 10-K evidence on disk until the reviewer submits APPROVED or REJECTED via POST /api/approval/{j ob_id}.

Risk & Governance Assessment	Inline Audit Trail Decomposition
Calculated Risk Score: 0.365 | Classification: LOW	Volatility factor: 0. 46 (45% wt) , Sentiment risk: 0. 45 (35% wt) , Filing flags: 0. 00 (20% wt)
News Agent: Real-Time Articles & Source Verification: Five expandable live RSS feeds (Yahoo Finance UK, The Times of India, The Guardian, TradingView, and The Economic Times) with UTC publication timestamps and verified source hyperlinks.	

Fig. 3. Live Finlet AI Risk & Governance Assessment and News Agent output for Reliance (RELIANCE. NS). The platform computes Calculated Risk Score: 0.365 | Classification: LOW with the audit trail "Volatility factor: 0.46 (45% wt), Sentiment risk: 0.45 (35% wt), Filing flags: 0.00 (20% wt)", followed by verified real-time news articles with publication timestamps and direct source links.

VI. EXPERIMENTAL EVALUATION AND LIVE OUTPUT ANALYSIS

A. Live Walkthrough: Reliance Industries (RELIANCE. NS)

To validate Finlet AI under live market conditions, Figs. 2 and 3 capture the operational Streamlit console (localhost: 8501) executing a concurrent multi-agent investigation on Reliance Industries (RELIANCE. NS) listed on the National

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, November 2026 Impact Factor: 8.2

Stock Exchange of India. When the analyst selects Reliance (RELIANCE. NS) in the left sidebar and clicks the red Run Finlet Analysis button, the FastAPI backend dispatches the Market, News, and Document agents concurrently in parallel.

As shown in Fig. 2, the Market Agent retrieves three months of NSE daily OHLCV bars via yfinance and populates the four top-level KPI cards under Overview: Reliance Industries (RELIANCE. NS): Current Price at ₹1 ,184.50 (−2.03% daily change badge), 30d Annual Volatility at 23.1 % (σ30d = 0.231), Volume Momentum at 1 .0x paired with a BEARISH medium-term moving-average trend badge, and News Sentiment classified as NEUTRAL (Score: 0. 1). Directly beneath the KPI strip, the Market Agent: Technical Price Action & 3-Month Trend panel renders an interactive Plotly candlestick chart tracking RELIANCE. NS from Jul 26, 2026 (trading near ₹1 ,300–₹1 ,335) through Aug 9, Aug 23, Sep 6, Sep 20, Oct 4, and Oct 18, 2026 as the stock consolidated into the ₹1 ,160–₹1 ,210 price band.

B. Verification ofFactor Decomposition & Audit Trail

Scrolling to the Risk & Governance Assessment panel in Fig. 3 demonstrates how our deterministic three-factor risk equation eliminates black-box LLM guesswork. Substituting the live Reliance indicators into Eqs. (5)–(8) yields the exact values displayed on screen:

Volatility Factor (Vnorm, 45% wt): σ30d / 0.50 = 0.231 / 0.50 = 0.462 (displayed as 0. 46 at 45% wt), contributing 0.45 × 0.462 = 0.2079.

Sentiment Risk (Snorm, 35% wt): (1 .0 − 0.10) / 2.0 = 0.450 (displayed as 0. 45 at 35% wt), contributing 0.35 × 0.450 = 0.1575.

Filing Flags Factor (Dnorm, 20% wt): min(1 .0, 0.25 × 0) = 0.000 (displayed as 0. 00 at 20% wt), contributing 0.20 × 0.000 = 0.0000.

Summing the three weighted components produces Rtotal = 0.2079 + 0.1575 + 0.0000 = 0.3654 → 0.365 (LOW), matching the header Calculated Risk Score: 0. 365 | Classification: LOW and the blue audit trail banner in Fig. 3 verbatim. Because 0.365 < 0.65, the run bypasses the HITL interrupt and proceeds automatically to PDF compilation.

C. Real-Time News Agent Source Verification

Directly below the risk banner in Fig. 3, the News Agent: Real-Time Articles & Source Verification panel renders five expandable evidence accordions pulled live from Google News RSS. Unlike conversational LLMs that hallucinate article titles, every entry in Finlet AI provides an exact UTC publication timestamp and a clickable Open Original Article hyperlink:

1. Yahoo Finance UK: “Reliance Industries Limited (RELIANCE.NS) stock price, news, quote and history” — Published: Thu, 22 Oct 2026 00: 25: 15 GMT.
2. The Times of India: “Reliance Industries stock surges 2% as Jio IPO gets fully subscribed in 100 minutes” — Published: Wed, 21 Oct 2026 07: 38: 00 GMT.
3. The Guardian: “Billionaire Mukesh Ambani launches India’s biggest-ever IPO”.
4. TradingView: “How Reliance Jio IPO is acting as shot in the arm for RIL shares - Will Mukesh Ambani company stock fly more? Buy now?”.
5. The Economic Times: “RIL stock: Jio’s listing could unlock value, but will it drive a re-rating?”.

Deterministic Governance Matrix & Operational Routing Thresholds

LOW Band [0.000, 0.399]: Stable operational and market profile (e.g., `RELIANCE. NS` at `0. 365`). Auto-approved to ReportLab PDF generator.

MEDIUM Band [0.400, 0.649]: Moderate volatility or mixed catalysts. Proceeds automatically and embeds an amber caution badge in the PDF dossier.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

HIGH Band [0.650, 1.000]: Elevated volatility, adverse news, or multiple 10-K flags (e.g., NVDA at 0. 670). Execution halted at SQLite HITL checkpoint.

TABLE III:CROSS-MARKET EMPIRICAL EVALUATION OF FINLET AI ACROSS INDIAN (NSE) AND U.S. (NASDAQ) EQUITIES (WEIGHTS: 0.45 / 0.35 / 0.20)

Ticker Symbol	Price / Change	σ_{30d} / Vol	Sent. (S)	Flags	V_{norm} (45%)	S_{norm} (35%)	D_{norm} (20%)	R_{total} & Governance Action
RELIANCE.NS (Live)	₹1,184.50 (−2.03%)	23.1 % / 1 .0x	+0.10	0	0.462	0.450	0.000	0.365 LOW Auto-Approved
TCS.NS (NSE)	₹4,120.00 (+0.85%)	15.8% / 1 .1 x	+0.18	1	0.316	0.410	0.250	0.336 LOW Auto-Approved
INFY.NS (NSE)	₹1,865.40 (+0.42%)	19.1 % / 1 .1 x	+0.12	1	0.382	0.440	0.250	0.376 LOW Auto-Approved
AAPL (NASDAQ)	$231 .80 (+1 .15%)	18.4% / 1 .1 x	+0.24	1	0.368	0.380	0.250	0.349 LOW Auto-Approved
MSFT (NASDAQ)	$428.50 (+0.64%)	16.2% / 0.9x	+0.30	0	0.324	0.350	0.000	0.268 LOW Auto-Approved
NVDA (NASDAQ)	$138.60 (−3.10%)	41 .0% / 1 .6x	−0.15	2	0.820	0.575	0.500	0.670 HIGH SQLite HITL Gate
STRESS CASE	Simulated Distressed	46.5% / 2.2x	−0.52	3	0.930	0.760	0.750	0.835 HIGH SQLite HITL Gate

D. Cross-Market Benchmarks & Subsystem Latency

Table III summarizes Finlet AI's evaluation across seven Indian (NSE) and U.S. (NASDAQ) equity scenarios under the 0.45 / 0.35 / 0.20 risk governance weights. Steady large-cap equities (RELIANCE. NS, TCS. NS, INFY. NS, AAPL, and MSFT) produced composite risk scores between 0.268 and 0.376 (LOW), passing automatically to ReportLab PDF compilation. By contrast, when elevated pre-earnings volatility in NVDA (σ30d = 41 .0%, Vnorm = 0.820) coincided with negative short-term news (S = −0.15, Snorm = 0.575) and two supply-chain concentration flags (Dnorm = 0.500), the composite score reached Rtotal = 0.670 $\geqslant$ 0.65, immediately pausing the LangGraph state machine at Layer 10 for mandatory human review.

VII. CONCLUSION AND FUTURE WORK

This paper presented Finlet AI, an evidence-grounded, twelve-layer multi-agent financial intelligence platform engineered to eliminate hallucinated financial metrics, missing source citations, and non-reproducible risk ratings. By isolating external data feeds behind Model Context Protocol (MCP) tool wrappers, executing domain retrieval agents concurrently via parallel fan-out across a stateful LangGraph workflow, preserving exact filing page numbers in ChromaDB, and governing publication through a deterministic three-factor risk matrix (0.45 volatility, 0.35 sentiment risk, and 0.20 filing flags) paired with SQLite Human-in-the-Loop checkpointing, Finlet AI bridges the reliability gap between conversational language models and institutional equity research compliance. Future work will extend our MCP adapter layer to incorporate real-time earnings call audio transcripts, quarterly institutional shareholding disclosures, and automated sector-peer relative valuation multiples.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

REFERENCES

[1] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv preprint arXiv:1908. 10063, 2019. [Online]. Available: https://arxiv.org/abs/1908.10063

[2] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A large language model for finance," arXiv preprint arXiv:2303. 17564, 2023.

[3] H. Yang, X.-Y. Liu, and C. D. Wang, "FinGPT: Open-source financial large language models," in Proc. FinLLM Workshop at the 32nd Int. Joint Conf. Artif. Intell. (IJCAI), Macao, China, 2023.

[4] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), Hong Kong, 2019, pp. 3982–3992.

[6] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen, "FinanceBench: A new benchmark for financial question answering," arXiv preprint arXiv:2311. 11944, 2023.

[7] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in Proc. 11th Int. Conf. Learning Representations (ICLR), Kigali, Rwanda, 2023.

[8] Q. Wu et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," in Proc. Conf. Language Modeling (COLM), 2024.

[9] Y. Yu, H. Li, Z. Chen, Y. Jiang, Y. Li, D. Zhang, R. Liu, J. W. Suchow, and K. Khashanah, "FinMem: A performance-enhanced LLM trading agent with layered memory and character design," in Proc. AAAISpring Symp. Series on LLMAgents, 2024.

[10] W. Zhang et al., "FinAgent: A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist," in Proc. 30th ACMSIGKDD Conf. Knowledge Discovery and Data Mining (KDD), Barcelona, Spain, 2024, pp. 4314–4325.

[11] Y. Xiao, E. Sun, D. Luo, and W. Wang, "TradingAgents: Multi-agents LLM financial trading framework," arXiv preprint arXiv:2412. 20138, 2024.

[12] Y. Yu, Z. Chen, H. Li, Y. Jiang, D. Zhang, K. Khashanah, and J. W. Suchow, "FINCON: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making," in Advances in Neural Information Processing Systems (NeurIPS), 2024.

[13] Anthropic, "Model Context Protocol (MCP): Open specification for connecting AI assistants to data systems," 2024.

[14] LangChain AI, "LangGraph: Building stateful, multi-actor applications with LLMs," 2024.