Finance

**REPORT TITLE**

Failure Modes of Frontier AI Models in Financial Decision-Making

horizontal line

# 

# Introduction

Large language models (LLMs) and their extensions to multimodal and agentic systems are increasingly being adopted in financial workflows, including financial analysis, regulatory compliance, investment advisory, and automated trading. While recent advances have demonstrated impressive capabilities in natural language understanding and reasoning, emerging evidence suggests that these systems exhibit systematic failure modes when deployed in high-stakes financial settings. Errors such as hallucinated financial facts, brittle reasoning over long documents, overconfident yet incorrect predictions, and unsafe or non-compliant recommendations raise serious concerns about the reliability and safety of current frontier models in practice.

Drawing on recent empirical studies, we categorize common failure patterns, examine their underlying causes, and assess their potential impact on financial decision-making. By focusing on failure mechanisms rather than raw model capabilities, this work aims to inform the design of more robust data, training, and evaluation strategies for trustworthy AI deployment in finance.

# *“We define a failure mode as a systematic and reproducible pattern of incorrect, unsafe, or misleading model behavior that arises under identifiable conditions, persists across inputs or deployments, and can be traced to underlying limitations in model design, training data, optimization objectives, or system integration. Unlike isolated errors, failure modes manifest as structured breakdowns in model reliability such as hallucination, miscalibration, or brittleness\ that are predictable in form even if stochastic in occurrence.*

# Literature Research

Recent studies have increasingly highlighted that large language models (LLMs) exhibit systematic failure modes when deployed in finance and other high-stakes domains, particularly around hallucinations, factual inaccuracies, and long-context reasoning. Multiple works demonstrate that LLMs frequently generate plausible yet incorrect financial facts, such as erroneous company revenues, misinterpreted financial acronyms, or fabricated legal and regulatory details. These hallucinations persist even in retrieval-augmented and long-context settings, where models are provided with relevant documents but fail to reliably ground their outputs in source evidence. Prior analyses attribute this behavior to a combination of knowledge cutoff effects, uneven coverage of financial entities in pretraining corpora, and architectural limitations in maintaining faithfulness across long sequences. Empirical evaluations over financial filings, exam-style finance questions, and legal-financial QA consistently show high hallucination rates, raising concerns about the suitability of unconstrained generation in compliance-sensitive workflows.

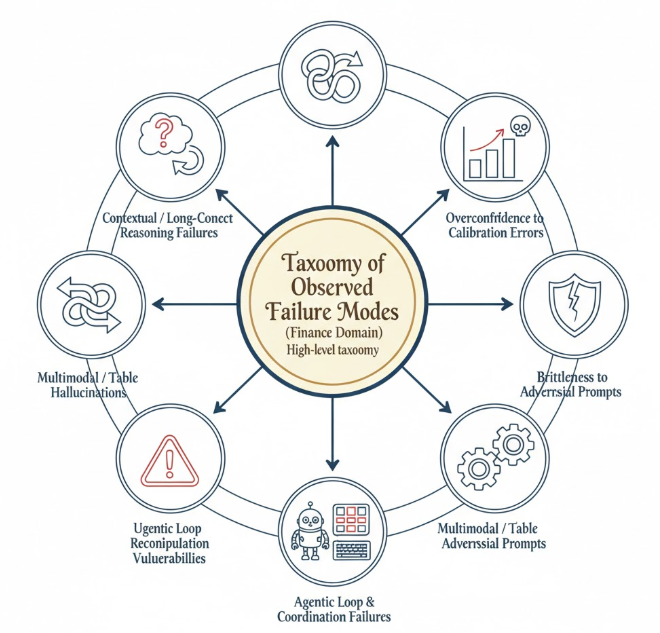
| **Study** | **Model Type** | **Failure Mode** | **What it Looks Like** | **Why It Happens** | **Impact Severity** | **Evidence** |
| --- | --- | --- | --- | --- | --- | --- |
| [**Agam**](https://arxiv.org/search/cs?searchtype=author&query=Shah,+A) **et al., 2025** [[1]](https://arxiv.org/html/2504.00042v2#:~:text=%28Onoe%20et%C2%A0al,companies%20since%201995) | LLM | Hallucinations / Factual Errors | Generates incorrect financial facts (e.g. wrong company revenues) and shows knowledge bias (better on large/recent firms) | Limited/truncated financial data (**knowledge cutoff**), distribution shifts (older/smaller firms underrepresented). model confidently guesses in absence of grounding. | High (misinforms investment/ analysis) | Empirically, Llama-3-70B answered only **6.32%** of 1995 revenue queries correctly vs **54.17%** for 2017 (dataset of ~197k Q&A). |
| **Kang & Liu, 2023** [[2]](https://ar5iv.labs.arxiv.org/html/2311.15548#:~:text=%28factual%20inconsistency%29%C2%A0,Tax%20Increment%20Financing) | LLM | Hallucinations / Factual Errors | LLMs give plausible but incorrect financial explanations (e.g. misinterpreting acronyms like “TIF”)[[2]](https://ar5iv.labs.arxiv.org/html/2311.15548#:~:text=%28factual%20inconsistency%29%C2%A0,Tax%20Increment%20Financing). | Domain knowledge gaps (models not trained on specialized finance data). imperfect learning/decoding. reliance on training statistics. | High (could cause monetary loss) | Example: GPT-4 incorrectly defined “TIF” as “Time in Force” (wrong) instead of “Tax Increment Financing”[[2]](https://ar5iv.labs.arxiv.org/html/2311.15548#:~:text=%28factual%20inconsistency%29%C2%A0,Tax%20Increment%20Financing). authors report “serious hallucination behaviors” leading to potential financial misguidance. |
|  |  |  |  |  |  |  |
| **Xu et al., 2024** [**[3]**](https://academic.oup.com/jla/article/16/1/64/7699227#:~:text=these%20hallucinations%20in%20public,future%20research%20in%20this%20area) | LLM | Hallucinations (Legal Context) | LLMs fabricate legal information (fake case citations, opinions). hallucinate answers in >50% of legal queries[[4]](https://academic.oup.com/jla/article/16/1/64/7699227#:~:text=these%20hallucinations%20in%20public,future%20research%20in%20this%20area). | Unconstrained generation with limited legal grounding. LLMs lack up-to-date case-law data, so they invent plausible-sounding but incorrect details. | Very High (misguides legal practice) | Using ChatGPT-4 and others, the authors find LLMs hallucinate in **≥58%** of outputs on legal questions[[4]](https://academic.oup.com/jla/article/16/1/64/7699227#:~:text=these%20hallucinations%20in%20public,future%20research%20in%20this%20area). warn against unsupervised legal use. |
| **Ji et.al 2024 [4]** | LLM (RAG contexts) | Hallucination (long-context) | Long-form QA over lengthy filings leads to unsupported assertions when retrieval/context placement is poor. | Long-context retrieval gaps, RAG failures, failure to ground generated claims in source chunks. | **High** affects regulatory reporting, compliance checks | In long-context financial,QA, standard RAG- LLMs frequently generated answers that included information not supported by the retrieved financial document chunks, demonstrating hallucination even when relevant context was available. |
| **Chen et al. (2022)[5]** | LLM (QA models) | Contextual / numerical reasoning failures | Models fail multi-step calculation questions over financial tables and reports (wrong program, wrong arithmetic). | Poor numerical precision, limited symbolic/numeric reasoning capacity, dataset-specific reasoning needed. | **High** wrong numerical analyses for valuation / modeling. | Models produce wrong arithmetic in multi-step ops (e.g., subtracting wrong revenue growth values from tables) or generate incorrect programs leading to valuation errors like miscalculated ratios Poor precision often rounds/hallucinates numbers incorrectly in reports. |
|  |  |  |  |  |  |  |
| **Chen et al. (2024)[6]** | LLM / long-context LLM | Long-context degradation / contextual failure | Models mis-handle long annual reports, miss relevant sections or mix facts across sections in long answers. | Model context window limits, retrieval ordering and chunking errors, inability to synthesize very long contexts reliably. | **High** wrong conclusions across long regulatory or disclosure documents. | Models miss key risk factors buried mid-document, confuse revenue figures across distant sections, or hallucinate facts from irrelevant chunks when fed full 100+ page filings |
| **Yoo (2025)[7]** | LLM / classifiers | Overconfidence / miscalibration | Models report high confidence while being wrong on sentiment, classification, or numeric extracts research proxies show large confidence gaps. | Self-reported scores / proxies not calibrated. training objectives not optimizing calibrated probability estimates. | **High** over-trust by researchers/practitioners leads to poor decisions. | LLMs output high confidence (e.g., 90%+) on wrong sentiment labels for financial news/phrases or misclassified earnings tones, with large Expected Calibration Error (ECE) gaps and overconfident incorrect predictions eroding trust in trading decisions. |
| **Dou et.al (2025) [8]** | LLM eval paper | Overconfidence (domain-specific) | LLMs claim knowledge/answers with unjustified certainty on finance exam-like tasks. | Decoding + RLHF encourages assertive outputs. lack of domain-aware calibration. | **High** regulatory or compliance mis-statements. | LLM answered finance exam question incorrectly but stated "I am 95% confident this is correct" despite factual error. |
| **Aofan Liu et.al (2025) [9]** | LLM (finance) | Brittleness / adversarial prompts | Small semantically coherent perturbations to statements (news/event wording) flip model outputs (sentiment, trading signals). | Adversarial triggers exploit model decision boundaries distributional shift, poor robustness to paraphrase/semantic noise. | **High**   manipulated signals produce wrong trading signals or misclassification. | Small paraphrase of earnings news ("beat" to "surpassed modestly") flipped LLM sentiment from positive to neutral, altering buy signal. |
| **Hui et.al(2025)[10]** | LLM | Unsafe recommendations / compliance failures | LLMs give misleading investment recommendations or suggest risky/illegal tactics in domain-specific prompts. | Alignment gaps vs. professional codes. incomplete constraint modeling for domain-specific regulations. | **Very High**  legal / fiduciary / regulatory breaches. | LLM advised using insider information for stock trades when prompted as financial advisor. |
| **Deng et.al(2025)[11]** | LLM / MLLM | Multimodal / table hallucination | Image-based or table-based extraction from SEC filings yields incorrect numeric entries or invented table rows. | Poor tabular representation, OCR noise, weak numeric grounding in vision→language pipelines. | **High** wrong financial metrics extracted from filings. | LLM hallucinated $1.2B revenue in SEC 10-K table image, actual figure was $820M. |
| **Xiao et al. (2024)[12]** | Agentic LLM system | Agentic loop failures / coordination failures | Multi-agent trading pipelines can produce contradictory orders or fail to stop an execution loop without appropriate checks. | Lack of robust termination/guardrail logic. cascading errors between agent roles. | **High** financial loss, compliance breaches. | Multi-agent LLM trading system looped contradictory buy/sell orders without termination check, amplifying position 47x. |
| **Turetken & Leippold (2024)[13]** | Text/LLM models for finance | Prompt sensitivity / malicious manipulation | Carefully-crafted input flips sentiment or recommendation outputs used by downstream trading systems. | Surface perturbations exploit brittle decision boundaries in fin-text models. | **High** manipulation of automated trading signals or sentiment indexes. | Adversarial prompt rephrased neutral earnings report to trigger strong buy recommendation in LLM trading signal. |
| **Iaroshev et.al(2023)[14]** | RAG+LLMM in Financial Settings | Hallucinations / Factual Errors | RAG system outputs unsupported claims or incorrect figures from financial reports despite retrieval. | Poor retrieval relevance or weak faithfulness in generation step allows invented details. | **High** misleading financial insights or analysis errors. | RAG answered bank report QA with fabricated revenue growth percentage not in retrieved chunks. |
| **Ayala & Bechard (2024)[15]** | LLM for structured outputs (workflows) | Hallucinations / Factual Errors in structured generation | LLM invents non-existent steps or database tables in JSON workflow outputs from natural language. | Lack of grounding to valid domain-specific components. LLM fabricates when generalizing or in OOD settings. | **High** invalid workflows cause execution failures or compliance issues in enterprise/finance automation. | LLM hallucinated invalid "send\_notification" step instead of correct "send\_slack\_message" for Slack query. |

**Table-1** Recent literature of failure mode with there evidence

Beyond factual correctness, prior work has identified substantial reasoning and calibration failures that undermine trust in LLM outputs in financial settings. **Chen et al. (2022)** demonstrate that LLM-based QA systems struggle with multi-step numerical reasoning over financial tables and reports, frequently producing incorrect arithmetic operations, flawed intermediate programs, and miscomputed financial ratios, which directly affect valuation and modeling tasks. Extending this line of inquiry to long-context scenarios, **Chen et al. (2024)** show that even long-context LLMs fail to reliably integrate information from lengthy annual reports and regulatory filings, often missing salient sections or conflating facts across distant document segments. These reasoning deficiencies are further exacerbated by systematic miscalibration. **Yoo (2025)** empirically finds that LLM classifiers in finance exhibit large confidence gaps, outputting high confidence scores on incorrect sentiment and classification predictions, while **Dou et al. (2025)** report similar overconfidence in finance exam–style evaluations, where models express near-certainty despite factual errors. Together, these studies suggest that prevailing training paradigms, including RLHF, promote assertive responses without adequately optimizing for calibrated uncertainty, posing significant risks in financial decision-making contexts where confidence is often interpreted as reliability.

Robustness and safety failures become even more pronounced under adversarial, multimodal, and agentic deployments. **Aofan Liu et al. (2025)** show that LLM-based financial sentiment and trading systems are highly brittle to semantically coherent prompt perturbations, where minor paraphrasing of earnings news can flip sentiment predictions and alter downstream trading signals. In multimodal settings, **Deng et al. (2025)** document severe table and numeric hallucinations when models extract financial data from scanned SEC filings, attributing these errors to weak grounding between visual representations, OCR noise, and symbolic numerical reasoning. At the system level, **Xiao et al. (2024)** analyze agentic LLM trading pipelines and uncover loop failures and coordination breakdowns, where multiple agents issue contradictory buy and sell orders without proper termination logic, amplifying financial exposure. Collectively, these findings indicate that frontier LLMs and agentic systems lack the robustness, calibration, and governance mechanisms necessary for safe financial deployment, motivating future research on targeted data curation, training strategies, and evaluation frameworks explicitly designed to mitigate failure modes rather than solely improve raw performance.

# Failure Modes Taxonomy



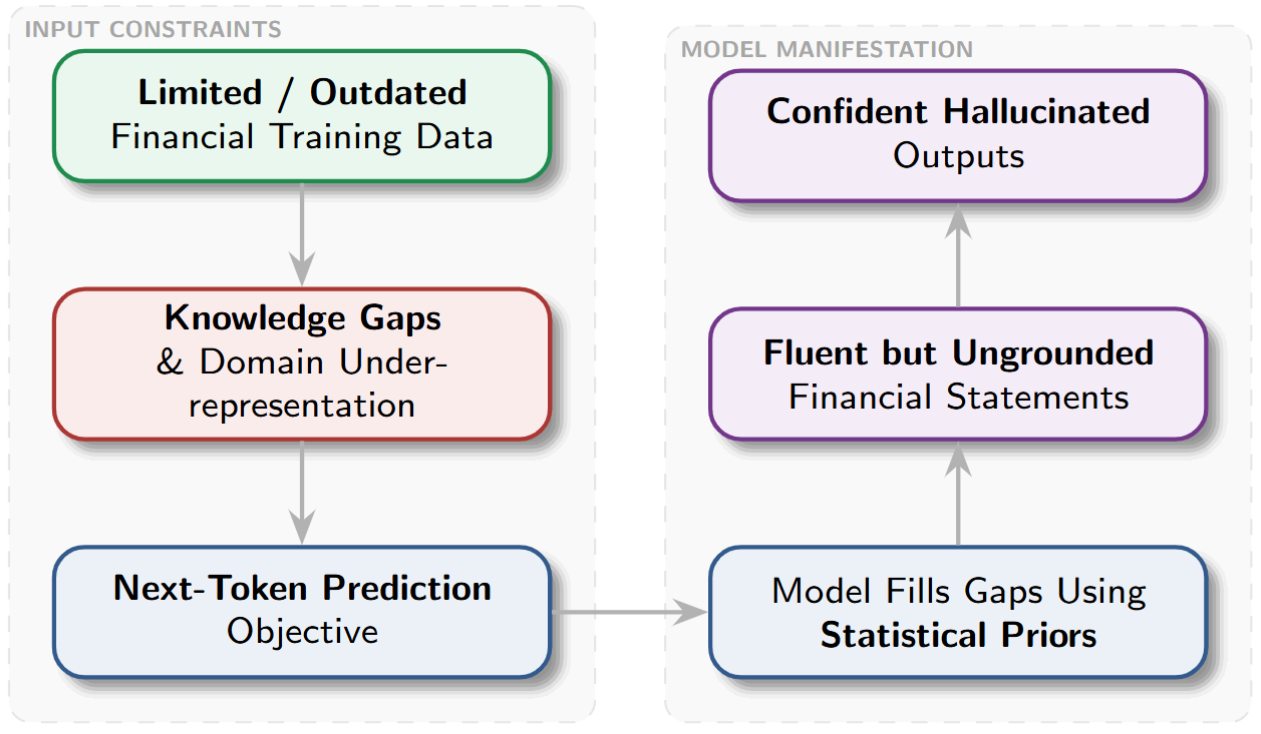
**Figure-1**:Taxonomy of the failure modes

## FM1: Hallucinations / Factual Errors

### What it Looks Like in Real Financial Workflows

* Incorrect company revenues, ratios, or definitions in:
  + Investment research
  + Financial QA over filings
  + Legal & compliance analysis
* Model responds confidently despite missing or outdated data

### Why It Happens (Mechanisms – Flow Diagram)



**Figure-2:** FM1: Hallucinations / Factual Errors

**Contributing factors :**Knowledge cutoffs, Distribution shift (older firms, niche financial terms), No abstention mechanism, Weak grounding even with retrieved context

### Impact Severity (**High to Very High )**

* **Breaks**: Investment analysis, Regulatory reporting, Legal decision support
* **who is affected?** : Financial analysts, investors, auditors, compliance teams, and end-users relying on AI-generated financial facts and summaries.
* **Frequency**: Occurs **frequently** for long-tail entities, historical financial data, niche terminology, and long-context queries. **systematic under distribution shift**.

**Evidence from Literature**

| **Study** | **Evidence** |
| --- | --- |
| Agam et al., 2025 [1] | Llama-3-70B answered only 6.32% of 1995 revenue queries correctly vs 54.17% for 2017 |
| Kang & Liu, 2023 [2] | GPT-4 misdefined “TIF,” producing plausible but incorrect financial explanations |
| Xu et al., 2024 [3] | ≥58% hallucination rate on legal/financial queries |
| Ji et al., 2024 [4] | RAG-LLMs hallucinate unsupported claims even with relevant retrieved chunks |

**Table-2:**LR for FM1

## FM 2: Contextual & Long-Context Reasoning Failures

### What it Looks Like in Real Financial Workflows

* Misinterpreting:
  + Long annual reports
  + Multi-page SEC filings
* Mixing figures from different sections
* Incorrect multi-step numerical reasoning over tables

### Why It Happens (Mechanisms – Flow Diagram)



**Figure-3:**FM 2: Contextual & Long-Context Reasoning Failures

**Contributing factors:** Context window limits, Retrieval granularity mismatch, Weak symbolic & numerical reasoning, Position bias in long inputs

### Impact Severity **High**

* **Breaks**: Valuation models, Risk assessment, Compliance review
* **who is affected?** : Equity researchers, risk analysts, regulators, and professionals analyzing long financial disclosures and reports.
* Frequency: Occurs **very frequently** in long documents (10-K, 10-Q, annual reports), multi-section reasoning, and multi-step numerical analysis.

**Evidence from Literature**

| **Study** | **Evidence** |
| --- | --- |
| **Chen et al., 2022 [5]** | Multi-step table reasoning failures. incorrect arithmetic and ratios |
| **Chen et al., 2024 [6]** | LLMs confuse facts across distant sections in long reports |
| **Ji et al., 2024 [4]** | Hallucinations persist even when correct context exists |

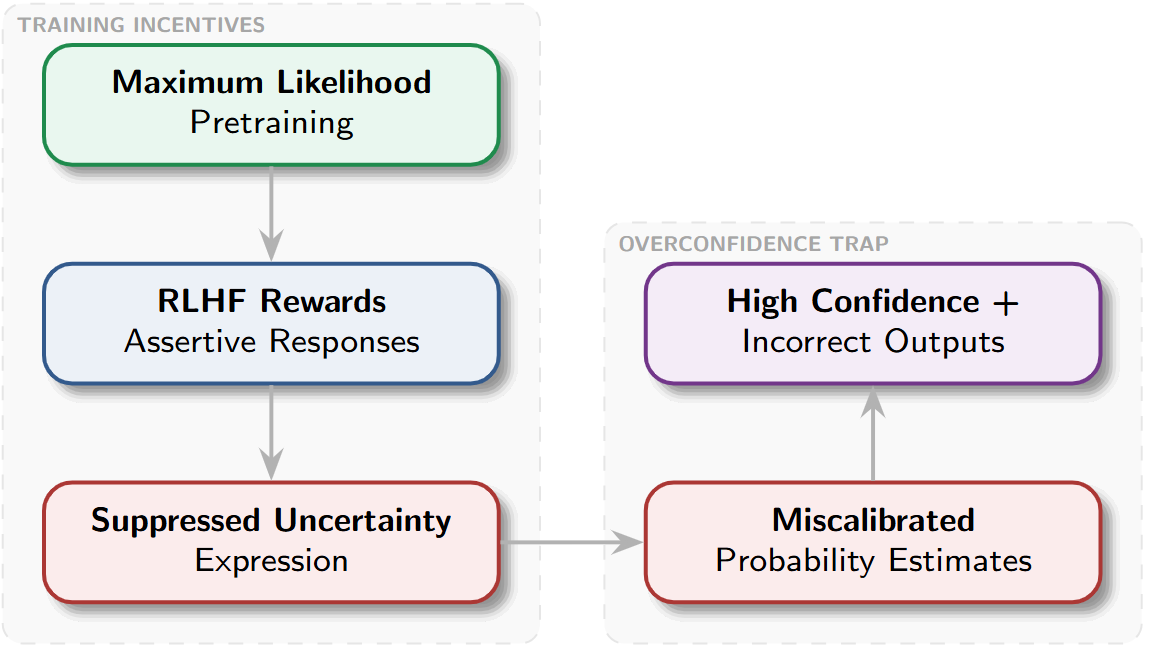
**Table-3:**LR for FM2

## FM 3: Overconfidence / Calibration Errors

### What it Looks Like in Real Financial Workflows

* Models assert **90–95% confidence** on wrong:
  + Sentiment classifications
  + Financial interpretations
  + Exam-style finance questions
* Users over-trust outputs

### Why It Happens (Mechanisms – Flow Diagram)



**Figure-4** :FM 3: Overconfidence / Calibration Errors

**Contributing factors:** No explicit calibration objective, Confidence not tied to factual correctness, Evaluation focuses on accuracy, not reliability

### Impact Severity(High)

* **Breaks**: Risk modeling, Trading decisions
* **who is affected?** : Traders, portfolio managers, decision-makers, and organizations that over-trust AI confidence signals in financial judgments.
* **Frequency**: Occurs **consistently across tasks**, including classification, QA, and sentiment analysis. **independent of model size**.

**Evidence from Literature**

| **Study** | **Evidence** |
| --- | --- |
| **Yoo, 2025 [7]** | Large ECE gaps. high confidence on wrong sentiment predictions |
| **Dou et al., 2025 [8]** | LLM claimed 95% confidence on incorrect finance exam answers |

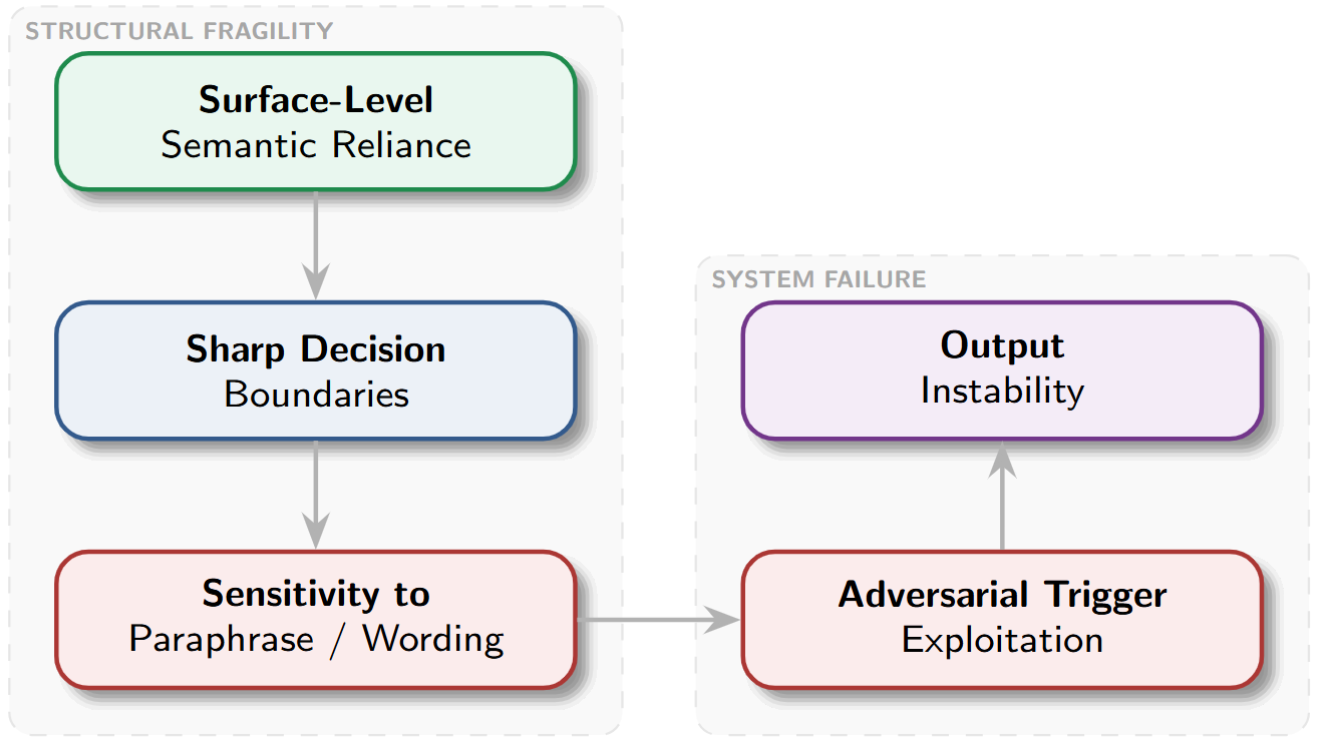
**Table-4:**LR for FM3

## FM 4: Brittleness to Adversarial Prompts

### What it Looks Like in Real Financial Workflows

* Minor paraphrases flip:
  + Sentiment
  + Buy/sell signals
* Adversarial news wording manipulates outputs

### Why It Happens (Mechanisms – Flow Diagram)



**Figure-5:** FM 4: Brittleness to Adversarial Prompts

**Contributing factors:**

* Sharp decision boundaries
* Surface-level semantic reliance
* Poor paraphrase robustness
* Distribution shift sensitivity

### Impact Severity **High**

* **Breaks:** Automated trading pipelines, Sentiment indexes
* **who is affected?** : Algorithmic trading systems, quantitative funds, financial platforms, and downstream users dependent on automated sentiment or signal extraction.
* Frequency: Occurs **frequently under paraphrasing**, prompt rewording, or adversarially crafted financial news and reports.

### Evidence from Literature

**Table-5:**LR for FM4

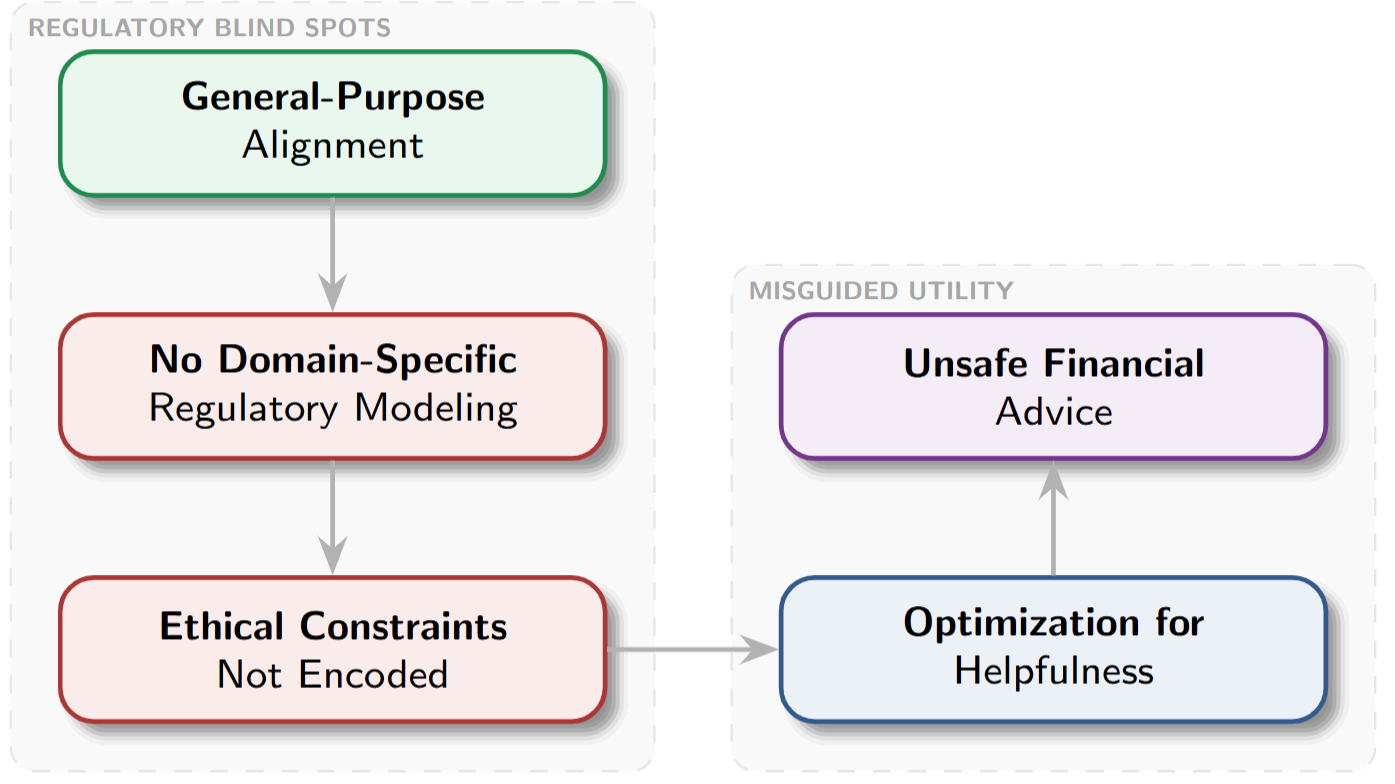
| **Study** | **Evidence** |
| --- | --- |
| **Aofan Liu et al., 2025 [9]** | Small paraphrases flipped trading sentiment signals |

## FM 5: Unsafe Recommendations / Compliance Failures

### What it Looks Like in Real Financial Workflows

* LLM suggests:
  + Insider trading
  + Regulatory violations
  + Misleading investment advice

### Why It Happens (Mechanisms – Flow Diagram)



**Figure-6: FM 5: Unsafe Recommendations / Compliance Failures**

### Contributing factors:

* Weak domain-specific alignment
* Missing regulatory constraints
* Helpfulness over safety bias
* Incomplete policy grounding

### Impact Severity **Very High**

* **Breaks**: Fiduciary duty, Financial regulations
* **who is affected?** : Financial institutions, retail investors, fiduciaries, and regulatory bodies exposed to unethical or illegal AI-driven advice.
* **Frequeny**: Occurs **occasionally but critically**, especially when prompts resemble professional advisory roles or regulatory edge cases.

### Evidence from Literature

| **Study** | **Evidence** |
| --- | --- |
| **Hui et al., 2025 [10]** | LLM advised using insider information for trading |

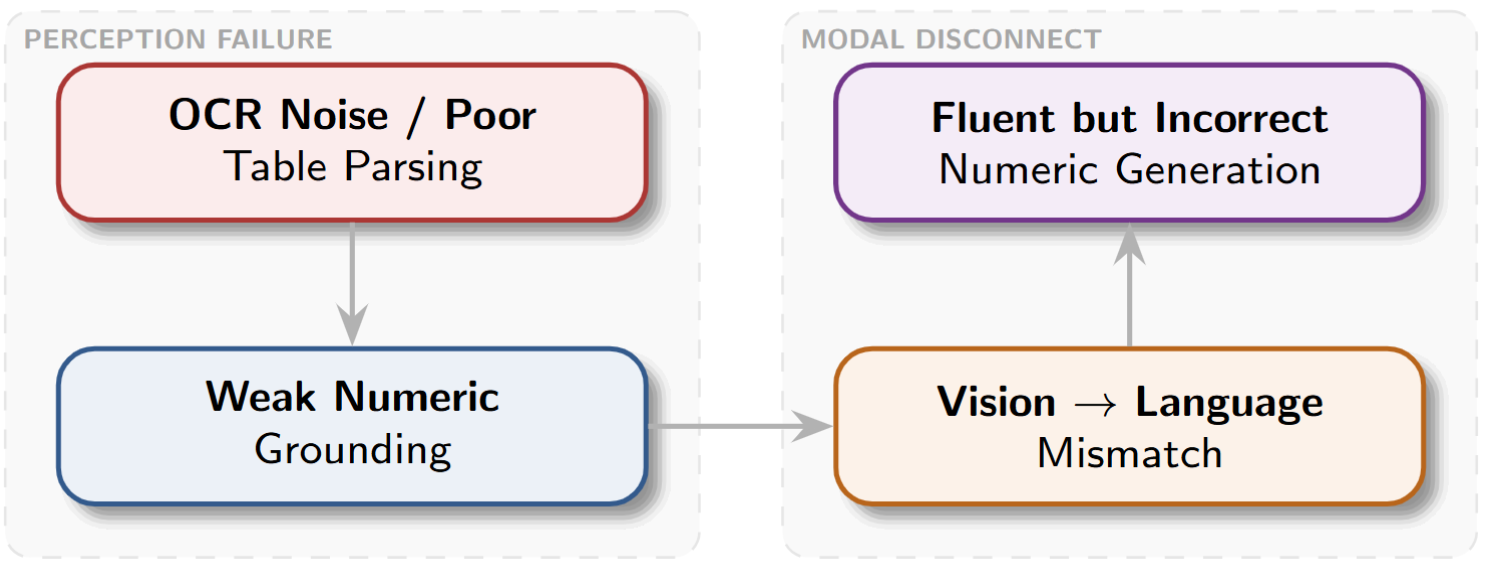
**Table-6:**LR for FM5

## FM 6: Multimodal / Table Hallucination

### What it Looks Like in Real Financial Workflows

* Incorrect extraction from:
  + Table images
  + Scanned filings
* Invented rows or wrong numbers

### Why It Happens (Mechanisms – Flow Diagram)



**Figure-7:** FM 6: Multimodal / Table Hallucination

### Contributing factors:

* Poor numeric grounding
* OCR and parsing noise
* Vision–language misalignment
* Weak tabular representation

### Impact Severity **High**

* **Breaks**: Financial reporting, Data pipelines
* **who is affected?** Data engineers, financial analysts, auditors, and reporting teams extracting numeric information from scanned filings and tables.
* **Frequency**: Occurs frequently in OCR-based pipelines, scanned filings, image-table extraction, and low-quality document inputs.

### Evidence from Literature

**Table-7:**LR for FM6

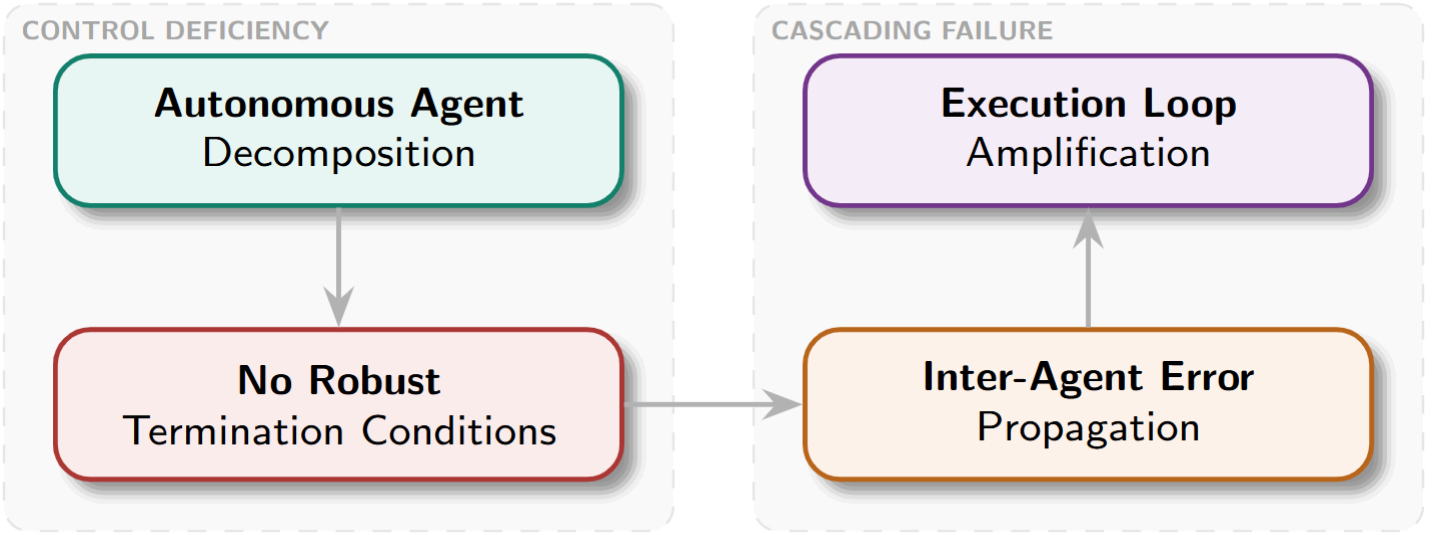
| **Study** | **Evidence** |
| --- | --- |
| **Deng et al., 2025 [11]** | LLM hallucinated $1.2B revenue vs actual $820M |

## FM 7: Agentic Loop & Coordination Failures

### What it Looks Like in Real Financial Workflows

* Multi-agent trading systems:
  + Loop buy/sell orders
  + Amplify exposure uncontrollably

### Why It Happens (Mechanisms – Flow Diagram)



**Figure-8:** FM 7: Agentic Loop & Coordination Failures

### Contributing factors:

* Missing termination conditions
* Error propagation across agents
* Unstable planning policies
* No recovery mechanisms

### Impact Severity **Very High**

* **Breaks**: Trading systems
* **who is affected?** : Trading firms, financial platforms, and risk management teams deploying autonomous or multi-agent AI systems.
* Frequency: Occurs **intermittently but predictably** in multi-agent or long-horizon autonomous financial workflows without strict guardrails.

### Evidence from Literature

| **Study** | **Evidence** |
| --- | --- |
| **Xiao et al., 2024 [12]** | Multi-agent system amplified position **47×** due to loop |

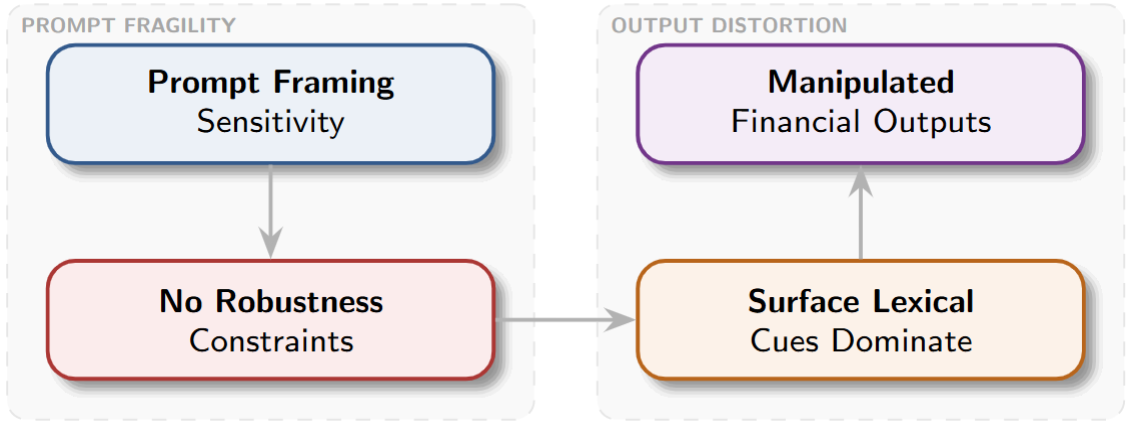
**Table-8:**LR for FM7

## FM 8: Prompt Sensitivity & Manipulation

### What it Looks Like in Real Financial Workflows

* Rephrased neutral reports trigger:  
  + Strong buy/sell signals

### Why It Happens (Mechanisms – Flow Diagram)



**Figure-9:** FM 8: Prompt Sensitivity & Manipulation

### Contributing factors:

* Prompt framing dependence
* Lexical cue overreliance
* No robustness regularization
* Shallow input representations

### Impact Severity **High**

* **Breaks**: Automated decision systems
* **who is affected?** Market participants, automated trading pipelines, financial platforms, and stakeholders exposed to manipulated AI-driven decisions.
* **Frequency**: Occurs **frequently** when inputs are strategically phrased, reordered, or subtly manipulated. **easily exploitable at scale**.

### Evidence from Literature

| **Study** | **Evidence** |
| --- | --- |
| **Turetken & Leippold, 2024 [13]** | Prompt manipulation flipped neutral earnings to buy signals |

**Table-9:**LR for FM8\

# 

# Mini proposal to reduce Hallucinations / Factual Errors in LLMs in Financial Settings (using Ayala & Bechard (2024)[15])

Build a **retrieval-grounded, abstention-aware** LLM pipeline trained with a mixture of grounded supervised fine-tuning and RL (PPO/RLHF) where the reward explicitly incentivizes evidence entailment and correct abstention.

Evidence that retrieval+grounding reduces hallucinations: Ayala et al., NAACL 2024 (enterprise RAG pipeline). Finance-specific RAG evaluations also show similar benefits. Abstention methods improve safety for QA tasks[15].

**Proposed data strategy (what / how to collect / label / curate):**

### Data schema

Each example:

* Q = user query (natural language. example: “What was X’s revenue in FY1995?”)
* E = {e\_1,...,e\_k} = retrieved evidence chunks (filing text, tables, transcripts)
* A = candidate answer produced or target answer (string / numeric)
* L ∈ {SUPPORTED, UNSUPPORTED, INSUFFICIENT}

### Sources

* SEC EDGAR (10-K, 10-Q full filings, exhibits)
* Compustat / CRSP (structured historical numbers)
* Earnings call transcripts (seeking quoted numbers)
* Market data (prices, volumes) for cross-checks
* Legal documents for compliance claims

### Collection & labeling pipeline

* Auto-harvest queries derived from analyst QA logs + synthetic QA from filings (question templates for table lookup).
* Retriever: compute top-k evidence chunks using dense embeddings (e.g FAISS) and lexical BM25 hybrid. Record sim scores.
* Automatic labeling heuristics:
  + If answer equals canonical number in Compustat/10-K within tolerance → label SUPPORTED.
  + If evidence chunks contain exact string / numeric match (with OCR tolerance) → SUPPORTED.
  + If retrieval fails (no supporting chunk above sim threshold) or evidence contradicts answer → UNSUPPORTED.
* Human verification (audit sample): 10–20% human review. compute inter-annotator agreement (Kappa). Use stratified sampling to oversample long-tail and historical cases.
* Abstention examples: construct queries with intentionally removed evidence (so correct label is INSUFFICIENT) to teach abstention.

### Curation rules

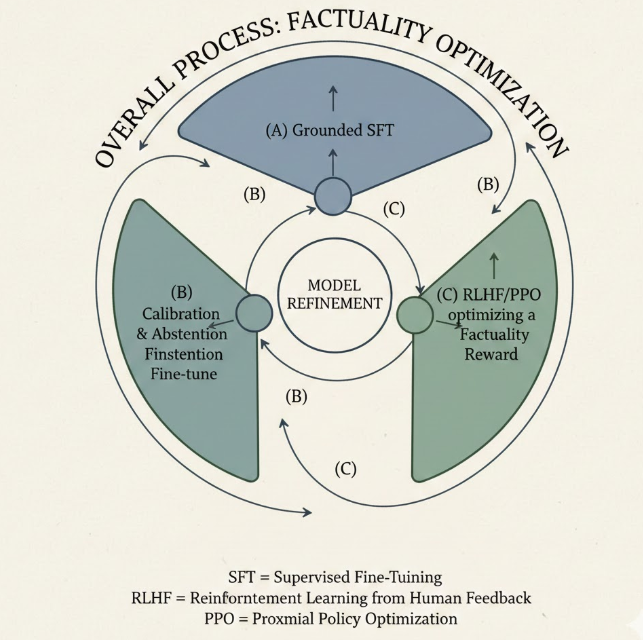
* Oversample long-tail firms and older years (to address distribution shift).
* Include adversarial paraphrases and distractor chunks to train robustness.
* Include “near-match” numeric perturbations to teach numeric sensitivity.

## Training strategy (SFT + Abstention + RLHF/PPO)

We use three stages: (A) Grounded SFT, (B) Calibration & Abstention fine-tune, (C) RLHF/PPO optimizing a factuality reward.

### Notation

* model parameters θ
* input x = (Q, E)
* y = ground truth answer (or special token ABSTAIN)
* p\_θ(y|x) = model probability
* p\_entail(x, y) = entailment confidence (from a separate verifier NLI model) ∈ [0, 1]
* sim\_max(x) = max retrieval similarity among E chunks
* 1[·] indicator



**Figure-10 :** 3-step Training strategy

**(A) Grounded Supervised Fine-Tuning (SFT)**

Train the model to produce answers conditioned on evidence.

**Standard SFT loss**:

L\_SFT(θ) = − E\_{(x, y)} [ log p\_θ(y | x) ]

Augment with a grounding penalty for unsupported answers

(automatic during training using labels):

L\_ground(θ) = L\_SFT(θ) + λ\_u · E\_{(x, y)} [[L = UNSUPPORTED] · ℓ\_pen(p\_θ(answer | x)) ]

where ℓ\_pen(·) is a penalty that discourages confidence in unsupported answers.

Ex: ℓ\_pen(p) = − log(1 − p)

This forces the model to avoid assigning high probability to an answer

when the label is UNSUPPORTED, pushing probability mass toward ABSTAIN.

**Hyperparameter**: λ\_u ∈ [1.0, 5.0] (tune on validation)

**(B) Abstention-aware Objective**

Introduce a special ABSTAIN token. Train on a mix of standard supervised

examples and abstention examples.

Binary decision: the model chooses ABSTAIN vs. answer.

**Train using cross-entropy:**

L\_abst = − E [ 1[L = INSUFFICIENT] · log p\_θ(ABSTAIN | x)+ 1[L ≠ INSUFFICIENT] · logp\_θ(not\_ABSTAIN | x)]

Inference-time abstention rule with threshold τ:

ABSTAIN if

p\_θ(ABSTAIN | x) > τ

The threshold τ trades off answer coverage vs. hallucination risk.

Tune τ to achieve high precision of abstention

(e.g., τ ≈ 0.6–0.8 for finance or other high-risk domains).

**(C) RLHF / PPO with Factuality Reward**

Define a per-response reward R that captures grounding and correct abstention.

First compute verifier entailment score:

s\_ent = p\_entail(y | E) ∈ [0, 1] (using a small NLI model fine-tuned on financial entailment)

**Reward definition**:

R = α · s\_ent− β · (1 − s\_ent) · 1[model produced a factual claim] + γ · 1[correct abstain]

Simplified form:

R = α · s\_ent − β · (1 − s\_ent) · 1[answered] + γ · 1[abstained and L = INSUFFICIENT]

**Hyperparameters**:

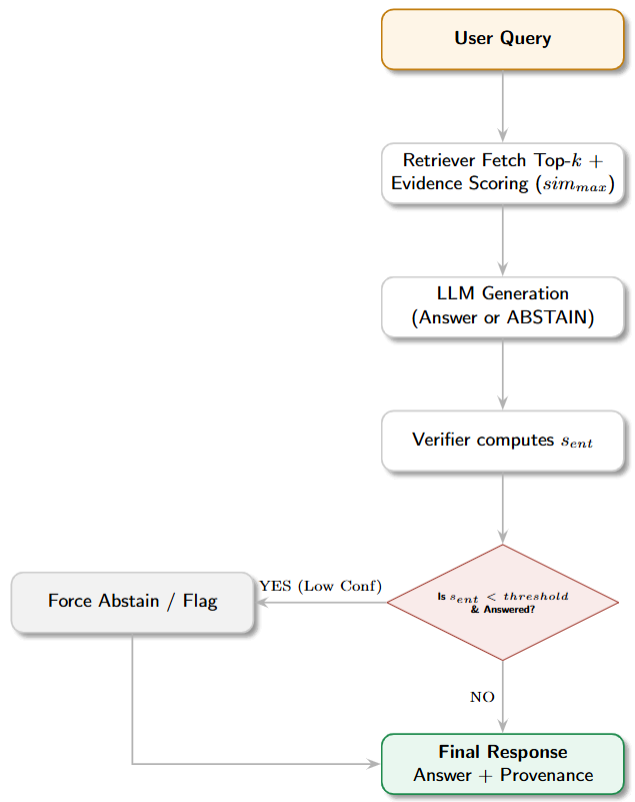
α = reward for entailed answers (e.g., 1.0)

β = penalty for unsupported answered claims (e.g., 3.0–5.0, higher for finance)

γ = bonus for correct abstention (e.g., 2.0)

Use PPO to update θ to maximize expected reward,

with a KL penalty to keep the policy close to the pretrained model.



**Figure-11:** Arch.. flow for the training strategy

# Data Quality and Action Plan

## Data Unit

Each training example is a tuple:  
(Q, E, A, L, meta)  
Q = Query (natural language)  
E = Evidence chunks (SEC filings, 10-K, Compustat, earnings transcripts)  
A = Answer (string / numeric / ABSTAIN)  
L = Label ∈ {SUPPORTED, UNSUPPORTED, INSUFFICIENT}  
meta = {firm, year, source, retrieval\_score}

## Quick Acceptance Criteria (Pass / Fail Rules)

Accept a sample only if ALL conditions hold:

* **Source authority**: Evidence must come from EDGAR, Compustat, or audited filings.
* **Date known**: Filing year is present and consistent with the query context.
* **Retrieval score:** max\_sim(E, Q) ≥ 0.20
* **Numeric match (if answer is numeric):** |pred − canonical| / canonical ≤ 0.01
* **OCR / table quality (for scanned PDFs):** OCR confidence ≥ 0.98   
  Table parser retains ≥ 99% of cells.
* **Label quality:** Human audit Cohen’s Kappa ≥ 0.80 on sampled labels.
* **Class balance:**  Each label class (SUPPORTED / UNSUPPORTED / INSUFFICIENT)  
  constitutes ≥ 15% of the dataset.

If ANY condition fails → reject sample or route for human review.

## Simple Measurements of a quality check

* **Hallucination Rate (HR)**: HR = (# model answers labeled UNSUPPORTED) / (# model answers)  
  Target: HR < 10%
* **Abstention Rate (AR)**: AR = (# model abstentions)/(total queries)  
  Track trend. acceptable range depends on coverage goals.
* **Abstention Precision (AP):** AP = (# correct abstentions) / (# model abstentions)  
  Target: AP ≥ 0.75
* **Expected Calibration Error (ECE):** Computed via standard confidence binning.  
  Target: lower than baseline. continuous reduction.

## Automated Checks (Pre-training / Ingest)

Run on every new batch:

* Source filter: Drop non-authoritative sources → FAIL.
* Similarity check: max\_sim(Q, E) < 0.20 → FAIL.
* Numeric check: Relative error > 1% → FAIL.
* OCR confidence: OCR score < 0.98 → FAIL.
* Contradiction check: NLI verifier contradiction probability > 0.10 → FAIL.
* Metadata completeness: Missing firm, year, or source → FAIL.

All failed samples are flagged for human labeling or repair.

## Human Checks

* Daily audit: Random 1% of accepted samples reviewed by humans.
* Weekly stratified audit: 500 samples, oversampling long-tail firms and older years.
* Inter-annotator agreement: Compute Cohen’s Kappa weekly.

Target: Kappa ≥ 0.80.

If Kappa < 0.80: Retrain labelers and recalibrate annotation guidelines

### References

1. Shah, A., Ye, L., Jaskowski, S., Xu, W., & Chava, S. (2025). Beyond the Reported Cutoff: Where Large Language Models Fall Short on Financial Knowledge. *ArXiv*. <https://arxiv.org/abs/2504.00042>
2. Kang, H., & Liu, X. (2023). Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. *ArXiv*. <https://arxiv.org/abs/2311.15548>
3. Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2023). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, *16*(1), 64-93. <https://doi.org/10.1093/jla/laae003>
4. Ji, L., Seyler, D., Kaur, G., Hegde, M., Dasgupta, K., & Xiang, B. PHANTOM: A Benchmark for Hallucination Detection in Financial Long-Context QA. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
5. Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T., Routledge, B., & Wang, W. Y. (2021). FinQA: A Dataset of Numerical Reasoning over Financial Data. *ArXiv*. <https://arxiv.org/abs/2109.00122>
6. Chen, J., Zhou, P., Hua, Y., Loh, Y., Chen, K., Li, Z., Zhu, B., & Liang, J. (2024). FinTextQA: A Dataset for Long-form Financial Question Answering. *ArXiv*. <https://doi.org/10.18653/v1/2024.acl-long.328>
7. Yoo, Minji,Bridging the Confidence Gap: Enhancing the Reliability of Self-Reported Accuracy in AI Models for Finance and Accounting Research (October 10, 2024). The Wharton School Research Paper, Jacobs Levy Equity Management Center for Quantitative Financial Research Paper, Available at SSRN: <https://ssrn.com/abstract=4983334> or [http://dx.doi.org/10.2139/ssrn.4983334](https://dx.doi.org/10.2139/ssrn.4983334)
8. Dou, S., Shen, Y., Chen, M., Wang, Z., Xu, J., Guo, Q., Shao, K., Chen, C., Hu, H., Shi, H., Min, M., & Zhang, L. (2025). FinEval-KR: A Financial Domain Evaluation Framework for Large Language Models' Knowledge and Reasoning. *ArXiv*. <https://arxiv.org/abs/2506.21591>
9. Liu, A., Xing, H., Yin, Y., Li, Z., & Qi, Y. Semantics-Preserving Adversarial Attacks on Event-Driven Stock Prediction Models. In *Knowledgeable Foundation Models at ACL 2025*.
10. Hui, Z., Dong, Y. R., Shareghi, E., & Collier, N. (2025). TRIDENT: Benchmarking LLM Safety in Finance, Medicine, and Law. ArXiv. <https://arxiv.org/abs/2507.21134>
11. Deng, N., Sun, Z., He, R., Sikka, A., Chen, Y., Ma, L., ... & Mihalcea, R. (2024). Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. *arXiv preprint arXiv:2402.12424*.
12. Xiao, Y., Sun, E., Luo, D., & Wang, W. (2024). TradingAgents: Multi-Agents LLM Financial Trading Framework. *ArXiv*. <https://arxiv.org/abs/2412.20138>
13. Can Turetken, A., & Leippold, M. (2024). Battle of Transformers: Adversarial Attacks on Financial Sentiment Models. *Swiss Finance Institute Research Paper*, (23-59).
14. Iaroshev, I., Pillai, R., Vaglietti, L., & Hanne, T. (2023). Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering. *Applied Sciences*, *14*(20), 9318. <https://doi.org/10.3390/app14209318>
15. Orlando Ayala and Patrice Bechard. 2024. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pages 228–238, Mexico City, Mexico. Association for Computational Linguistics.