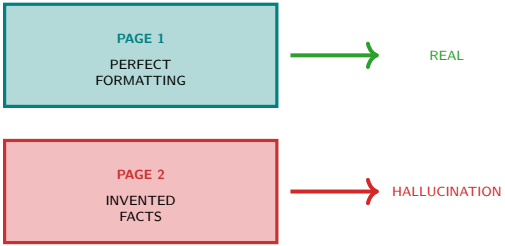
Why can an AI write a perfect-sounding financial analysis that is completely wrong?

The hallucination problem:

- An analyst asks the model for a regulatory summary
- The model generates a polished two-page report
- Citations look authentic, numbers are precise
- Compliance review reveals: half the facts are invented

Why this happens:

- Language models are trained to predict plausible next words
- They generate text that sounds confident
- Confidence does not correlate with correctness
- Models cannot distinguish fact from fiction

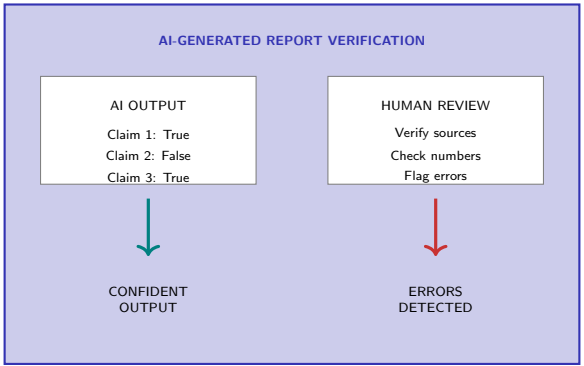


Key Insight

Language models express hallucinations with the same confidence as correct statements. Verification is mandatory.

Lesson 5.2: Generative AI and LLMs — Core tension: models can read, write, and reason about documents, but hallucinate with complete confidence.

Have you ever used AI to write something – and had to check every claim?



Reflection

Language models can produce fluent text faster than humans, but every claim must be verified against authoritative sources.

Hallucination is not a bug that will be fixed. It is inherent to how language models generate text.

What are the different ways LLMs are being deployed in financial services?

Document Processing:

- Extract key terms from contracts
- Summarize regulatory filings
- Parse financial statements
- Classify documents by type

Report Generation:

- Draft earnings summaries
- Generate compliance reports
- Create client communications

Code Assistance:

- Generate SQL queries from natural language
- Auto-generate unit tests
- Document existing code

Use Case	Risk Level
Document extraction	Low
Email drafts	Low
Report summaries	Medium
Code generation	Medium
Regulatory interpretation	High
Trading signals	High

Deployment pattern:

- Start with low-stakes tasks
- Add human review layers
- Reserve high-stakes for mature systems

Key Insight

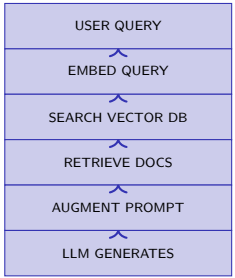
Deploy LLMs where errors are tolerable or human-reviewable first. Reserve high-stakes decisions for later.

Document processing and report generation are the most mature use cases. Trading signal generation remains experimental.

How does a retrieval-augmented generation system ground an LLM's answers in real documents?

The RAG pipeline:

1. User asks: What is our currency exposure?
2. Query is converted to a vector representation
3. Vector database finds the most relevant document chunks
4. Retrieved chunks are inserted into the prompt
5. LLM generates an answer grounded in the documents
6. System provides citations to source documents



Why RAG for finance:

- Reduces hallucination by grounding in actual documents
- No need to retrain when documents change
- Provides audit trail with citations

Key Insight

RAG is the most widely adopted LLM architecture in finance because it balances accuracy, cost, and auditability.

Retrieval-augmented generation transforms a language model from a creative writer into a research assistant with citations.

How do fine-tuning and prompting architectures differ for financial LLM deployment?

Prompting:

- Design input instructions to guide the model
- No training data required
- Fast to implement
- Cost: only API calls
- Example: few-shot learning with examples

Fine-tuning:

- Adjust model parameters on domain-specific data
- Requires thousands of labeled examples
- Weeks to implement
- Cost: thousands to hundreds of thousands
- Example: specialized model for regulatory text

Dimension	Prompt	Fine-tune
Data needed	None	Thousands
Time to deploy	Hours	Weeks
Cost	Low	High
Flexibility	High	Low
Control	Medium	High

Decision rule:

- Start with prompting
- Add RAG for factual grounding
- Fine-tune only for high-volume specialized tasks

Key Insight

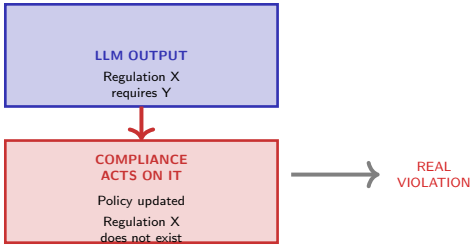
Most financial institutions succeed with prompting plus RAG. Fine-tuning is reserved for specialized high-volume tasks.

Prompt engineering is the lowest-cost way to customize LLM behavior. It requires no training data and no compute infrastructure.

What happens when an LLM hallucinates a regulatory requirement that does not exist?

Hallucination cascade:

- Compliance officer asks for a summary of new rules
- LLM generates a polished report citing specific regulations
- Team implements policy changes based on the summary
- Audit reveals: the cited regulation does not exist
- Real compliance violations result from following fake rules



Why regulatory hallucinations are dangerous:

- Acting on fake requirements wastes resources
- Missing real requirements creates violations
- LLMs cannot reliably distinguish real from plausible

Key Insight

Hallucinated regulatory requirements can cause real compliance violations. Never trust LLM legal interpretations without verification.

Regulatory interpretation remains the weakest area for LLMs. These tasks require maximum human oversight.

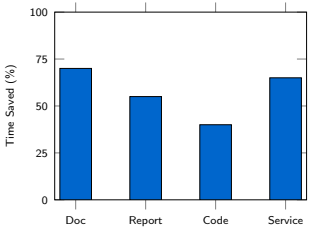
Where are LLMs creating measurable value in financial workflows?

High-value applications:

- Document processing: sixty to eighty percent time reduction
- Report generation: cost savings of five to fifteen per document
- Code assistance: thirty to fifty percent developer time savings
- Customer service: handle equivalent of hundreds of human agents

Common pattern:

- High-volume repetitive tasks
- Text-heavy workflows
- Human review already in place
- Low cost of error



Key Insight

LLMs create value where text processing is the bottleneck and human review catches errors.

Document processing is the low-hanging fruit: high volume, structured input, and human reviewers already in place.

Who benefits from LLM automation in finance and who is displaced?

Who benefits:

- Analysts freed from repetitive document review
- Institutions processing high document volumes
- Customers receiving faster responses
- Developers accelerated by code assistants

Who may be displaced:

- Junior analysts doing routine summaries
- Document processing specialists
- Entry-level customer service roles
- Workers without technical reskilling

Equity considerations:

- Training data biases can amplify
- Access to LLM tools may be unequal

Beneficiaries	At Risk
Senior analysts	Junior analysts
Fast processing	Routine roles
Cost savings	Job displacement
Strategic work	Repetitive tasks

Organizational response:

- Reskilling programs
- Human-AI collaboration models
- Focus on judgment-heavy roles

Key Insight

LLM automation shifts work from routine to judgment. Institutions must invest in reskilling.

Automation creates winners and losers. Responsible deployment requires transition support for displaced workers.

Three tests to evaluate whether an LLM application is safe for a financial use case

Test 1: Can every claim be verified against a source document?

- Use RAG to ground outputs
- Require citations for all facts
- Validate citations are real

Test 2: What is the cost of a hallucination in this context?

- Low-stakes: draft emails, internal summaries
- High-stakes: regulatory filings, trading signals
- Deploy in low-stakes first

Test 3: Is there a human review step before the output reaches a decision?

- Human-in-the-loop for all high-stakes tasks
- Automated review for low-stakes
- Clear escalation paths

The LLM Safety Assessment:

Criterion	Pass?
Citation verification	☐
Hallucination cost analysis	☐
Human review step	☐
Performance monitoring	☐
Bias and fairness check	☐
Regulatory compliance	☐

All boxes must be checked before high-stakes deployment.

Key Insight

Safety is not about perfect accuracy. It is about appropriate guardrails for the stakes involved.

In regulated finance, guardrails are not optional. They are the difference between a useful tool and a compliance liability.

Your Challenge

Design a RAG Pipeline for Financial Use

You are designing a retrieval-augmented generation system to answer questions about regulatory filings for a financial institution.

Tasks:

1. Describe the six stages of your RAG pipeline from user query to final answer
2. Identify three types of hallucination that could occur in this context
3. For each hallucination type, propose one specific mitigation strategy
4. Explain how you would validate that citations provided by the LLM point to real document passages

Hallucination types to consider:

- Fabricated facts (numbers, dates, requirements that do not exist)
- Invented citations (references to documents that do not exist)
- Temporal confusion (mixing data from different reporting periods)

Learning Goal

Apply RAG architecture principles to a real financial scenario and identify safety measures.

RAG provides auditability by anchoring every answer to source documents. The retrieval step is the key differentiator.