Exercises: Lesson 5.2 – Generative AI and LLMs in Finance

Module 5: The Automation Problem

Prof. Dr. Joerg Osterrieder

Exercise 1: Design a RAG Pipeline for Regulatory Q&A

Scenario: A European bank's compliance team receives 200 questions per week about internal policies and external regulations (MiFID II, GDPR, PSD2). Currently, junior analysts spend 4–6 hours per question researching answers. The bank has 5,000 policy documents and access to all relevant EU regulations.

Tasks:

1. Draw a complete RAG architecture diagram showing: document ingestion, chunking strategy, embedding model, vector database, retrieval, augmented prompt construction, LLM generation, and validation.
2. Specify your recommended chunk size and overlap. Justify your choice for regulatory documents.
3. What embedding model would you recommend (general-purpose vs. multilingual vs. domain-specific)? Why?
4. Design three guardrails specific to this compliance use case.
5. Estimate the weekly cost using a medium-tier model ($3/1M input tokens, $15/1M output tokens), assuming each question requires retrieving 5 chunks of 500 tokens each, plus a 200-token query and 800-token response.

Exercise 2: Hallucination Risk Assessment Matrix

Scenario: An investment bank is evaluating LLM deployment across six use cases. For each use case, assess the hallucination risk.

Use cases:

1. Generating first drafts of equity research summaries
2. Answering client questions about portfolio performance
3. Drafting marketing emails for new fund launches
4. Summarizing earnings call transcripts
5. Calculating Value-at-Risk (VaR) from market data
6. Reviewing loan covenants for compliance violations

Tasks:

a) For each use case, rate the hallucination risk (Low / Medium / High / Critical) across three dimensions: factual accuracy, numerical precision, and regulatory consequence.

b) Rank the six use cases from safest to most dangerous for LLM deployment.

c) For the two highest-risk use cases, propose specific mitigation strategies.

Exercise 3: LLM Deployment Cost-Benefit Analysis

Scenario: A mid-size asset management firm ($20B AUM) is considering deploying an LLM for automated client report generation. Currently, 15 analysts spend approximately 30% of their time writing monthly client reports.

Given data:

- Average analyst salary: $120,000/year (fully loaded cost: $180,000)
- 500 client reports per month, each taking 3 hours manually
- LLM deployment cost (Year 1): Integration $250K, API costs $60K/year, monitoring $40K/year
- LLM reduces report writing time by 65%, but requires 15 minutes of human review per report

Tasks:

a) Calculate the current annual cost of manual report generation.

b) Calculate the Year 1 total cost of the LLM-assisted approach (including remaining human time).

c) Calculate the net savings (or loss) in Year 1 and Year 2.

d) What non-financial benefits and risks should be considered beyond the cost calculation?

Exercise 4: Prompting Strategy Comparison

Task: For the following financial task, write three different prompts: zero-shot, few-shot (with 3 examples), and chain-of-thought. Then evaluate which is most appropriate.

Financial task: Classify the sentiment of the following earnings call excerpt as Positive, Negative, or Neutral, and extract the key financial metric mentioned.

Test excerpt: "While revenue grew 8% year-over-year to $14.2 billion, operating margins contracted by 150 basis points due to elevated restructuring charges. We expect margin recovery in the second half as cost actions take effect."

Requirements:

a) Write the zero-shot prompt (2–3 sentences of instruction)

b) Write the few-shot prompt with 3 labeled examples from different sentiment categories

c) Write the chain-of-thought prompt that asks the model to reason step-by-step

d) Which strategy would you recommend for processing 10,000 earnings calls? Justify with cost and accuracy considerations.

Exercise 5: Embedding Quality for Financial Search

Scenario: A vector database contains embeddings of 50,000 research notes from a sell-side broker. An analyst queries: "What is the outlook for European semiconductor companies given recent export controls?"

The retrieval returns these top-5 results:

1. "ASML Q3 earnings beat on strong EUV demand" (similarity: 0.87)
2. "US-China tech export restrictions tighten further" (similarity: 0.84)
3. "Semiconductor supply chain analysis: European fabs" (similarity: 0.82)
4. "European defense stocks rally on NATO spending" (similarity: 0.79)
5. "History of semiconductor industry cycles 1990–2020" (similarity: 0.77)

Tasks:

a) Which results are relevant? Which are false positives? Explain why the false positives scored high.

b) Propose two techniques to improve retrieval precision for this query.

c) If you could re-chunk the documents, what chunk size and metadata would you add to improve results?

Exercise 6: Model Selection for a Bank

Scenario: A retail bank needs to deploy LLMs for three different use cases simultaneously. Annual query volumes and accuracy requirements differ:

Use Case	Queries/Year	Avg Tokens	Min Accuracy
Customer FAQ chatbot	2,000,000	1,500	85%
Internal policy Q&A (RAG)	50,000	4,000	92%
Regulatory report review	5,000	12,000	97%

Available models: Small ($0.15/1M tokens, 88% accuracy), Medium ($3/1M tokens, 93% accuracy), Large ($15/1M tokens, 97% accuracy).

Tasks:

a. Assign the optimal model to each use case. Justify your choices.

b. Calculate the total annual API cost for your recommended configuration.

c. What is the cost if you naively use the Large model for all three? Calculate the savings from your optimized approach.

Exercise 7: LLM Governance Framework Design

Scenario: You are the newly appointed Head of AI Governance at a bank. The CEO wants to move from Stage 1 (Experimental) to Stage 3 (Production) within 12 months. Currently, employees use ChatGPT informally with no oversight.

Tasks:

a) Draft a 5-point LLM acceptable use policy covering: permitted use cases, prohibited use cases, data handling rules, output review requirements, and incident reporting.

b) Design a model risk management process for LLMs (analogous to SR 11-7). Include: validation, monitoring, documentation, and escalation triggers.

c) Identify three metrics you would track monthly to measure responsible AI adoption.

d) What organizational changes (roles, committees, training) are needed to support this transition?

Exercise 8: End-to-End LLM Implementation Case Study

Scenario: A private equity firm wants to use LLMs to accelerate due diligence. For each potential acquisition target, analysts review 200–500 documents (financial statements, legal contracts, market reports). Current due diligence takes 4–6 weeks per target.

Design a complete LLM-powered due diligence assistant:

a. Architecture: Choose between fine-tuning, RAG, or hybrid. Draw the system architecture and justify your choice.

b. Document processing: How would you handle the variety of document types (PDFs, spreadsheets, legal contracts)? Specify the ingestion pipeline.

c. Risk mitigation: What are the three most dangerous failure modes for this system? Design safeguards for each.

d. Evaluation: Define five metrics to evaluate the system before and after deployment.

e. Business case: Estimate the ROI assuming the firm evaluates 20 targets per year, each requiring $200K in analyst time. The LLM system costs $500K in Year 1. What is the break-even point?

Answer Key – Exercise 1

(1) Architecture should include: Document ingestion → Text extraction → Chunking → Embedding model → Vector DB → Query embedding → Similarity search → Top-k retrieval → Prompt construction (query + context + instructions) → LLM → Validation layer → Response.
(2) Recommended: 500–800 tokens per chunk with 100-token overlap. Regulatory documents have cross-referenced sections; overlap prevents losing context at chunk boundaries.
(3) Multilingual domain-specific embeddings (e.g., a model fine-tuned on legal/regulatory text in EU languages). General-purpose models miss regulatory jargon.
(4) Guardrails: (i) Citation enforcement – every answer must reference a specific document section; (ii) Confidence threshold – flag answers below 0.7 confidence for human review; (iii) Recency check – verify cited regulation version is current, not superseded.
(5) Per query: Input = 200 (query) + 5 × 500 (chunks) + 300 (prompt template) = 3,000 tokens. Output = 800 tokens. Weekly: 200 queries. Cost = (200 × 3,000 / 1M × \$3) + (200 × 800 / 1M × \$15) = \$1.80 + \$2.40 = **\$4.20/week** (\$218/year). The API cost is negligible; the real cost is integration and maintenance.

Answer Key – Exercise 2

(a) Risk assessment:

Use Case	Factual	Numerical	Regulatory	Overall
Marketing emails	Low	Low	Low	**Low**
Earnings summaries	Medium	Medium	Low	**Medium**
Equity research drafts	High	High	Medium	**High**
Client portfolio Q&A	High	Critical	High	**Critical**
VaR calculation	Medium	Critical	High	**Critical**
Loan covenant review	Critical	High	Critical	**Critical**

(b) Safest to most dangerous: Marketing emails → Earnings summaries → Equity research drafts → Client portfolio Q&A → VaR calculation → Loan covenant review.

(c) For VaR: Use LLMs only for report formatting; actual calculations must come from validated quantitative models. For loan covenants: RAG with exact document retrieval, mandatory human legal review, and dual-LLM cross-check.

Answer Key – Exercise 3

(a) Current cost: 500 reports/month $\times$ 3 hrs $\times$ 12 months = 18,000 hrs/year. At \$180K/2,080 hrs = \$86.54/hr. Total = 18,000 $\times$ \$86.54 = **\$1,557,692/year**.
(b) Year 1 LLM approach:

- Remaining human writing: 18,000 hrs $\times$ 0.35 = 6,300 hrs = \$545,192
- Human review: 500 $\times$ 12 $\times$ 0.25 hrs = 1,500 hrs = \$129,808
- Integration: \$250,000 (one-time)
- API: \$60,000; Monitoring: \$40,000
- Total Year 1 = \$545,192 + \$129,808 + \$250,000 + \$60,000 + \$40,000 = **\$1,025,000**

(c) Year 1 savings: \$1,557,692 $-$ \$1,025,000 = **\$532,692**. Year 2 (no integration cost): \$775,000, savings = **\$782,692**.
(d) Non-financial: faster delivery to clients, consistency of report quality, analyst satisfaction (less repetitive work). Risks: hallucinated performance numbers, client trust concerns, regulatory scrutiny of AI-generated reports.

Answer Key – Exercises 4–5

Exercise 4:
(a) Zero-shot: "Classify the sentiment of the following earnings call excerpt as Positive, Negative, or Neutral. Also extract the key financial metric. Excerpt: [text]"
(b) Few-shot: Provide 3 labeled examples (one per sentiment), then the target excerpt.
(c) Chain-of-thought: "Step 1: Identify positive signals. Step 2: Identify negative signals. Step 3: Weigh the balance. Step 4: Classify and extract the key metric."
(d) For 10,000 calls: Few-shot. It balances accuracy ($\sim$82%) with cost (3 examples add $\sim$300 tokens per call vs. chain-of-thought adding $\sim$500+ output tokens). Fine-tuning is justified only if this is a recurring quarterly task.

Exercise 5:
(a) Relevant: Results 1, 2, 3. False positives: Result 4 ("European" + "stocks" creates surface similarity) and Result 5 ("semiconductor" matches but historical focus is irrelevant).
(b) (i) Hybrid search combining semantic + keyword filters (require "export control" or "regulation"); (ii) Metadata filtering by date range and sector tags.
(c) Chunk size: 400–600 tokens. Add metadata: sector, geography, date, document type, key entities. This enables pre-filtering before semantic search.

Answer Key – Exercise 6

(a) Assignment: FAQ chatbot → Small model (88% > 85% requirement, massive volume). Policy Q&A → Medium model (93% > 92%, moderate volume). Regulatory review → Large model (97% meets requirement, low volume, highest stakes).
(b) Optimized cost:

- FAQ: 2M × 1,500 / 1M × \$0.15 = \$450
- Policy: 50K × 4,000 / 1M × \$3 = \$600
- Regulatory: 5K × 12,000 / 1M × \$15 = \$900
- **Total: \$1,950/year**

(c) All-Large cost:

- FAQ: 2M × 1,500 / 1M × \$15 = \$45,000
- Policy: 50K × 4,000 / 1M × \$15 = \$3,000
- Regulatory: 5K × 12,000 / 1M × \$15 = \$900
- Total: \$48,900. **Savings: \$46,950 (96% reduction).**

Answer Key – Exercises 7–8

Exercise 7 (key points):
(a) Policy: Permitted = internal research, document drafting, code assistance. Prohibited = client-facing without review, PII in prompts, trading decisions. Data = no client data to external APIs. Review = all outputs verified before external use. Incidents = report hallucinations and data leaks within 24 hours.
(b) MRM process: Initial validation (benchmark testing), ongoing monitoring (accuracy drift, hallucination rate), documentation (model cards), escalation (auto-flag if accuracy drops below threshold).
(c) Metrics: (i) Hallucination rate per use case, (ii) User adoption rate, (iii) Incident count.
(d) Roles: AI Governance Committee, Model Risk team expanded for LLMs, mandatory AI literacy training.

Exercise 8 (key points):
(a) RAG hybrid: RAG for document retrieval + fine-tuned model for domain-specific extraction. Justification: documents change per target (no static training set), but format consistency benefits from fine-tuning.
(c) Failure modes: (i) Missing critical risk factor (mitigation: completeness checklist), (ii) Hallucinated financial figure (mitigation: cross-reference with structured data), (iii) Outdated legal status (mitigation: timestamp validation).
(e) ROI: Current cost = $20 \times \$200K = \$4M$/year. If LLM reduces analyst time by 50% = \$2M savings. Year 1 net = $\$2M - \$500K = \$1.5M$. Break-even within Year 1 at approximately target #5.