OPERATION LEDGER-MIND

ENGINEERING REPORT 2026

PREPARED FOR
Alpha-Yield Capital

LEAD AI ARCHITECT
Rumesh Mohan

DATE
February 09, 2026

Executive Summary

This engineering report evaluates the efficacy of **Parametric Memory (Fine-Tuning)** versus **Non-Parametric Memory (Advanced Retrieval-Augmented Generation - RAG)** for automated financial intelligence. The assessment utilized the Uber Technologies 2024 Annual Report as the primary dataset, subjecting both architectures to a rigorous evaluation across 487 test questions.

> *"The architectural recommendation for Alpha-Yield Capital is a Hybrid RAG-FT System, balancing high-volume synthesis with compliance-critical precision."*

Quantitatively, the Fine-Tuned model (Llama-3-8B) demonstrated superior performance over the RAG system, achieving an average **LLM-as-a-Judge score of 4.01/5** and a **ROUGE-L overlap of 0.608**, significantly higher than the RAG system's 2.54/5 and 0.121 respectively. This resulted in a decisive **61.1% victory margin** for the Fine-Tuned model across **346 of 486** evaluated questions.

However, despite its quantitative lead, qualitative analysis revealed critical vulnerabilities within the Fine-Tuned model, particularly concerning **citation confabulation (34% of failures)**, **numerical metadata omission (28%)**, and **parametric mixing (23%)**. These issues highlight instances where the model blended facts from disparate sections or generated plausible but incorrect information.

Consequently, the architectural recommendation for Alpha-Yield Capital is a **Hybrid RAG-FT System**, designed to leverage the strengths of both approaches while mitigating their respective weaknesses, ensuring both high-volume analytical synthesis and compliance-critical precision.

Methodology

The evaluation framework was meticulously designed to simulate a production-grade financial analysis environment, ensuring the robustness and real-world applicability of our findings. The methodology primarily focused on three key areas: high-fidelity data generation, optimized fine-tuning processes, and the implementation of an advanced retrieval pipeline.

1. Synthetic Data Generation Pipeline

The cornerstone of this engineering project was the development of a high-fidelity synthetic dataset specifically engineered to stress-test both parametric and non-parametric architectures under conditions mirroring actual financial intelligence tasks. We meticulously transformed Uber's extensive **132-page 2024 Annual Report** into a structured evaluation suite comprising a total of **1,633 question-answer pairs** (**1,146 training pairs** and **487 test pairs**).

Teacher-Student Architecture

- **Teacher Model (Gemma 3:4B via Ollama)**: This model was tasked with generating a diverse range of queries through structured prompting. Our prompting strategy mandated a specific category distribution:
    - **Hard Facts (40%)**: Verifiable data points such as financial figures, dates, and legal references
    - **Strategic Summaries (30%)**: High-level business insights, competitive analysis, and forward-looking statements
    - **Stylistic/Creative outputs (30%)**: Testing tone consistency and professional financial writing
- **Student Model (Llama-3.3-70B via OpenRouter)**: Acting as the answer generator, this more capable model received both the chunk context and teacher-generated questions to produce ground-truth responses. The separation of question generation from answer generation was critical to prevent model-specific biases.

The document processing involved a **1,500-character fixed-size chunking strategy with zero overlap**. This decision was crucial in preventing information leakage between the training and test sets, thereby maintaining the integrity of our evaluation. The utilization of two separate LLMs in the Teacher-Student architecture was a deliberate choice to prevent "circular reasoning," a phenomenon where a model might generate questions it is pre-optimized to answer.

2. Fine-Tuning Configuration

The parametric memory system, dubbed **"The Intern,"** leveraged Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA) to memorize Uber's 2024 strategic narrative.

Model Architecture

- **Base Model**: Llama-3-8B-Instruct (unsloth/llama-3-8b-Instruct-bnb-4bit)
- **Quantization**: 4-bit NF4 (Normal Float 4) with double quantization enabled via BitsAndBytesConfig
- **LoRA Parameters**:
    - Rank (r): 16
    - Alpha (α): 16 (scaling factor)
    - Target Modules: q_proj, k_proj, v_proj, o_proj (all attention layers)
    - Dropout: 0.0

Training Configuration

- **Optimizer**: AdamW 8-bit with paged optimizers
- **Learning Rate**: 2e-4 with cosine annealing schedule
- **Training Steps**: 120 (exceeding the minimum 100 requirement)
- **Batch Size**: 1 per device with 8-step gradient accumulation (effective batch size: 8)
- **Final Training Loss**: 0.34
- **Hardware**: Google Colab T4 GPU (16GB VRAM)

The choice of 4-bit quantization was essential for fitting the 8B parameter model within T4 GPU memory constraints while maintaining inference quality. The LoRA configuration prioritized all attention mechanisms, as these layers are most critical for understanding contextual relationships in financial documents.

3. Hybrid RAG Parameters

The non-parametric system, architected as **"The Librarian,"** implemented a sophisticated Four-Stage Hybrid Retrieval Pipeline to ensure comprehensive and accurate information retrieval.

Stage 1 - Dense Vector Search

This initial stage utilized **all-MiniLM-L6-v2** (384-dimensional embeddings) for semantic intent matching. This model was chosen for its efficiency and effectiveness in capturing the conceptual meaning of queries, allowing for the retrieval of documents semantically similar to the user's input, even if exact keywords were not present. **Top-20 candidates** were retrieved based on cosine similarity scores.

Stage 2 - Lexical Keyword Search (BM25)

To complement the semantic search, **Okapi BM25** was employed for exact entity matching. This stage was critical for identifying precise terms, such as specific financial figures (e.g., "$1.97 billion") or proper nouns (e.g., "Dara Khosrowshahi"), ensuring that key factual elements were not overlooked. **Top-20 candidates** were retrieved based on term frequency and inverse document frequency scoring.

Stage 3 - Reciprocal Rank Fusion (RRF)

The results from the dense and lexical searches were then aggregated using Reciprocal Rank Fusion with **k=60**. RRF effectively combines ranked lists from multiple retrieval methods using the formula:

RRF(d) = Σ[1/(k + rank_i(d))]

This gives higher scores to documents that consistently rank well across different approaches, thereby improving the overall relevance of the retrieved documents.

Stage 4 - Cross-Encoder Reranking

The top-20 results from RRF were reranked using a Cross-Encoder model (**cross-encoder/ms-marco-MiniLM-L-6-v2**). Unlike bi-encoders that compute query and document embeddings independently, cross-encoders jointly process query-document pairs, enabling more nuanced relevance scoring. The **final top-10 chunks** were selected for context injection into the generation prompt, ensuring maximum precision while maintaining manageable context window sizes.

Vector Database

Weaviate Cloud (WCD) was selected for its native support of hybrid search, automatic vectorization, and production-grade scalability. The schema was configured with 1,500-character chunks indexed with both dense vectors and BM25 inverted indices.

WEAVIATE_URL="11oidykjrlycnjfcbkyiw.c0.asia-southeast1.gcp.weaviate.cloud"
WEAVIATE_API_KEY="Vi9TSEpFMjdXVmM1c0x4VV8rNTNKUnJScU05Yk5QL05wVUhjaDRaU1hXVUpXZE4vdU82VnJBbHJYT0RnPV92MjAw"

The "Hallucination" Audit: Fine-Tuning Failures

Despite its quantitatively higher aggregate scores, the Fine-Tuned model exhibited systematic and concerning failures within high-stakes financial contexts, particularly regarding numerical accuracy. Our comprehensive audit identified three primary failure modes.

Case 1: Citation Confabulation

One of the most critical failures observed was citation confabulation, where the Fine-Tuned model generated plausible but ultimately incorrect legal or numerical references. For instance, when queried for specific legal references related to the Sarbanes-Oxley Act (**15 U.S.C. 7262(b)**), the model would produce boilerplate text that appeared legitimate but was factually inaccurate. A notable example involved the model referencing **"Section 12(b) check-box boilerplate"** instead of the actual, specific legal code required for internal control attestation.

This is a **catastrophic failure** in a financial context, as incorrect legal citations can have severe compliance implications. While not directly a numerical error, this confabulation demonstrates the model's tendency to generate convincing but false information, which could easily extend to numerical data if not properly grounded. The model memorized the style and structure of legal text but failed to link it to the correct, verifiable content, indicating a superficial understanding rather than deep comprehension. In our audit, citation confabulation accounted for **34% of all failures**, representing the highest-risk failure mode.

Case 2: Numerical and Metadata Omission

A more direct and equally problematic failure mode was the frequent omission of precise numerical metadata. The Fine-Tuned model consistently struggled to provide exact page numbers or specific financial figures, even when the information was present in its training data.

For example, when asked about "Executive Compensation," the model could correctly identify the heading or topic (**"Item 11. Executive Compensation"**) but would fail to provide the corresponding page reference (specifically, **"Page 127"**). This parametric decay makes the

output unverifiable without manual intervention, which is unacceptable in financial reporting where auditability and precision are paramount.

In another instance, when asked for specific revenue figures or growth percentages, the model would often generalize or omit the exact numbers, providing qualitative statements like **"significant revenue growth"** instead of the precise figure **"$37.281 billion, representing 16% year-over-year growth."** This suggests a weakness in retaining and accurately recalling specific numerical details, preferring broader conceptual understanding over precise data points. The absence of critical metadata renders the information less trustworthy and requires significant human effort to validate. Omission failures represented **28% of total failures** in our audit.

Case 3: Parametric Mixing (Verbosity Injection)

The third significant failure mode, parametric mixing, often led to the injection of unrelated statistics and verbose, confusing responses, particularly when numerical data was involved. The model frequently blended information from disparate sections of the document, leading to a loss of contextual accuracy.

For example, in a query regarding personnel retention challenges in competitive markets, the model injected unrelated statistics about global employee counts (e.g., **"31,100 employees across 70+ countries"**) from a completely different section of the annual report discussing company scale. While these numbers might be factually correct in isolation, their inclusion in an irrelevant context made the response misleading and verbose.

This "verbosity injection" indicates an overgeneralization of attention, where the model struggles to filter out extraneous numerical information and focus solely on the relevant data for the given query. In a financial report, such mixing of data can lead to misinterpretations and incorrect conclusions, undermining the reliability of the analysis. The model's inability to maintain strict contextual boundaries for numerical data is a clear limitation in its fine-tuned performance. Parametric mixing accounted for **23% of failures**.

Audit Analysis Summary

Failure Mode	Frequency	Impact on Fintech Operations
Confabulation	34%	High Risk: Regulatory penalties
Omission	28%	Medium Risk: Impedes auditability
Verbosity Mixing	23%	Low Risk: Increases noise
Refusal	15%	Low Risk: Safe failure mode

Error Rate: 24.8% (121 of 486 questions exhibited critical failures)

The audit reveals a fundamental trade-off: the Fine-Tuned model excels at synthesis and coherence but fails catastrophically at precision-critical tasks. For financial intelligence systems, even a 24.8% error rate on numerical accuracy is unacceptable for compliance-facing workflows.

Conclusion: Architectural Recommendations

The strategic choice between deploying a Fine-Tuning (FT) approach and a Retrieval-Augmented Generation (RAG) system for a Fintech client hinges on a careful evaluation of the specific use case, balancing the need for synthesis with the imperative for precision and compliance.

When to Recommend Fine-Tuning (FT)

Fine-Tuning is highly recommended for scenarios demanding high-volume analytical synthesis and tasks where a broad understanding of patterns and trends is more critical than pinpoint factual recall. For a Fintech client, this would include applications such as:

- **Earnings Call Preparation**: Generating comprehensive summaries and insights from vast amounts of financial reports and market data to prepare executives for earnings calls. The FT model's ability to synthesize information efficiently (average latency: **2,524ms**) makes it ideal for quickly distilling key takeaways where exact citations are not immediately critical.

- **General Market Sentiment Analysis**: Analyzing large datasets of news articles, social media, and financial reports to gauge overall market sentiment. The FT model can identify overarching themes and sentiments, providing a holistic view without needing to cite every single source explicitly.

- **Internal Research and Development**: For exploratory data analysis or generating initial hypotheses where the primary goal is to uncover potential patterns or strategic insights. The model's strong ROUGE-L score (**0.608**) indicates superior content overlap with expected analyst outputs.

When to Recommend Retrieval-Augmented Generation (RAG)

RAG becomes indispensable for compliance-critical applications and scenarios involving dynamic data where verifiable provenance and absolute numerical precision are non-negotiable. For Fintech clients, these use cases are paramount:

- **Regulatory Compliance Reporting**: Any report or analysis that must adhere to strict regulatory guidelines, such as SEC filings, 10-K/10-Q reports, or internal audit reports. RAG's ability to provide exact citations with page numbers and ground its responses in specific source documents is crucial for demonstrating compliance and auditability. The system's superior latency (**1,696ms average**) also enables faster retrieval for time-sensitive compliance queries.

- **Dynamic Data Analysis**: When dealing with rapidly changing financial data, such as real-time market feeds, trading data, or live transaction records. RAG can retrieve the most up-to-date information directly from authoritative sources, ensuring that decisions are based on current data rather than memorized patterns that may be outdated.

- **Quantitative Financial Modeling and Valuation**: Tasks requiring precise dollar amounts, specific dates, or exact legal citations for financial models, valuations, or contract analysis. The need for verifiable provenance means that every piece of numerical information must be traceable back to its original source, a capability that RAG excels at with its four-stage retrieval pipeline.

BONUS: Cost Analysis

Assuming **500 daily users** making **10 queries each** (5,000 daily queries, **150,000 monthly**):

Fine-Tuning Deployment

- **Infrastructure**: AWS g5.xlarge instance (NVIDIA A10G, 24GB VRAM)
  - On-Demand: $1.006/hour × 730 hours = $734.38/month
  - Reserved (1-year): ~**$440/month** (40% savings)
- **Inference Cost**: ~$0.002 per query with 4-bit quantization
  - Monthly: 150,000 × $0.002 = **$300**
- **Total Monthly Cost**: $440 + $300 = **$740**

RAG Deployment

- **Vector Database**: Weaviate Cloud Standard tier
  - $25/month for 500k vectors + overage = ~**$45/month**
- **Embedding API**: OpenAI text-embedding-3-small
  - $0.00002 per 1k tokens × 150k queries × 500 tokens avg = **$150/month**
- **LLM Generation**: Groq Llama-3-70B (free tier sufficient for this volume)
  - **$0/month** (alternatively: $0.59/1M tokens = ~$265/month on OpenRouter)
- **Total Monthly Cost**: $45 + $150 + $0 = **$195**

Hybrid Orchestration

- 80% queries routed to FT: ($740 × 0.8) = **$592**
- 20% verification via RAG: ($195 × 0.2) = **$39**
- **Total Monthly Cost**: **$631**

Cost Recommendation: Deploy the Hybrid Orchestration Layer at **$631/month**, providing **24% cost savings** versus pure FT ($740) while maintaining compliance precision through selective RAG verification.

Final Recommendation for Alpha-Yield Capital

Given the strengths and weaknesses of both architectures, we recommend a **Hybrid Orchestration Layer** for Alpha-Yield Capital. This layered approach is designed to maximize efficiency and accuracy:

1. **Primary Layer (Fine-Tuning)**: This layer would handle approximately **80% of standard analytical queries**, leveraging the FT model's speed and synthesis capabilities for routine tasks such as earnings summaries, market commentary, and strategic insights.
2. **Verification Layer (RAG)**: For outputs containing critical numerical data (dollar amounts, percentages, dates) or requiring legal citations, the RAG system would be triggered for verification. A simple regex pattern (e.g., `/\$[\d,\.]+|\d+%|Page \d+/`) can identify high-risk content requiring verification.
3. **Human-in-the-Loop**: A final human review step would flag any discrepancies between FT outputs and RAG verifications for analyst review, providing an ultimate safeguard against errors. This tier is essential for SEC filings and board presentations.

This hybrid strategy ensures that Alpha-Yield Capital benefits from the rapid analytical synthesis of fine-tuning (2.5s average latency for 80% of queries) while maintaining the absolute precision and verifiability essential for financial intelligence (1.7s verification latency for critical 20%). The system achieves a **61.1% accuracy improvement** over pure RAG while maintaining compliance standards through selective verification, representing the optimal balance for production deployment.

Confidentiality & Disclaimer

Thank you for your attention.

Alpha-Yield Capital
