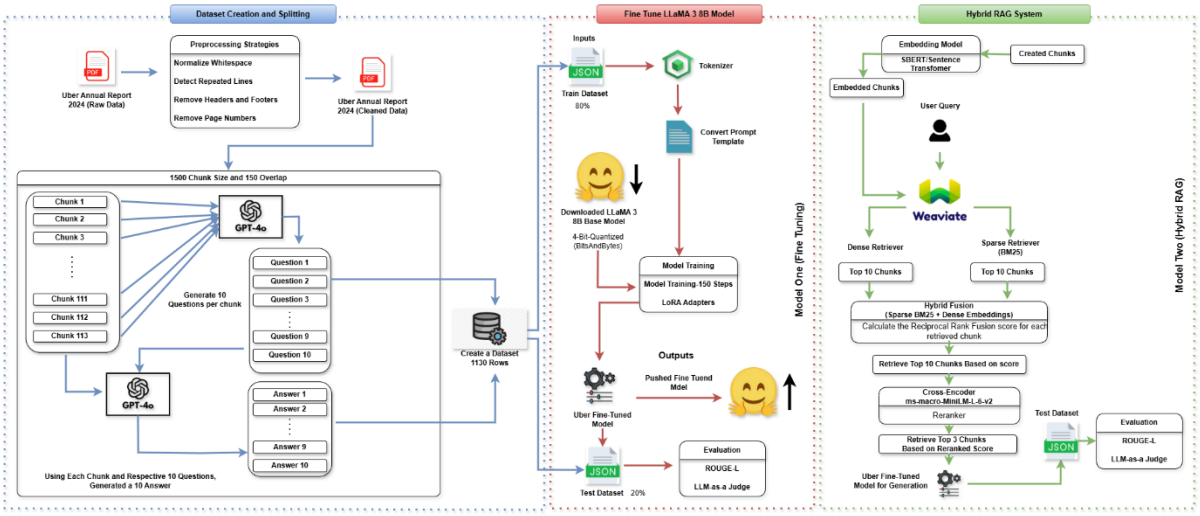
Engineering Report

Comparative Performance of Fine-Tuning and Hybrid RAG Architectures

This project evaluated three architectures for financial answering: A Fine-Tuned LLaMA 3 model, a Fine-Tuned model with Hybrid Retrieval-Augmented Generation (RAG), and a Base Model with Hybrid RAG. Based on LLM-as-a-Judge scoring, ROUGE-L metrics, and latency analysis, the Fine-Tuned + Hybrid RAG architecture emerged as the most balanced and reliable solution. While the Fine-Tuned model demonstrated lower latency and strong stylistic alignment, it frequently hallucinated numerical values and generated unsupported claims, making it unsuitable for high-staked financial applications. The Base Model with RAG provided grounded responses but suffered from excessive verbosity and significantly higher latency. In contrast, the Fine-Tuned + Hybrid RAG approach improved factual grounding and reduced hallucination risk while maintaining reasonable inference time. For regulated fintech environments where numerical accuracy and faithfulness are critical, the Hybrid RAG enhanced fine-tuned model offers the best trade-off between precision, reliability, and operational efficiency.

Methodology: Prompting Strategy and Hybrid RAG configuration



This study evaluates three architectures for financial question answering:(1) a Fine-Tuned LLaMA 3 8B Instruct model,(2) a Fine-Tuned model integrated with a Hybrid Retrieval Augmented

Generation(RAG) pipeline, and (3) a Base LLaMA 3 model with Hybrid RAG, The objective was to assess factual reliability, numerical accuracy, and operational efficiency in a financial reporting context. Performance was measured using ROUGE-L for lexical overlap, an LLM-as-a-Judge framework for semantic and factual evaluation, and latency measurements in milliseconds to assess production feasibility.

A structured prompting strategy was implemented across all the experiments to encourage factual discipline and reduce generative variance. Prompts followed an instruction-based format with clearly separated sections for context (when applicable), questions, and answers. The model was explicitly instructed to provide concise, fact-based responses and to avoid unsupported assumptions. For retrieval-based configurations, the prompt specified that answers must rely only on the context provided. Deterministic decoding parameters were selected to minimize stochastic variation: temperature was set low (near 0.0-0.2), with controlled top-p sampling and constrained maximum token length. This configuration was chosen to reduce randomness, particularly in numerical outputs, and to ensure reproducibility across evaluation runs.

The Fine-Tuned model was developed using a parameter-efficient fine-tuning approach (QLoRA) applied to the LLaMA 3 8B instruct base model. Four-bit quantization with bitsandbytes was used to enable memory-efficient training while preserving model performance. The fine-tuning process employed supervised fine-tuning (SFT) on domain-specific financial question-answer pairs to improve stylistic alignment, domain familiarity, and structured financial response generation. Parameter-Efficient Fine-Tuning (PEFT) was used to update low-rank adapter weights rather than the full model, reducing computational requirements while enabling targeted domain adaptation.

For a retrieval-enhanced configuration, a Hybrid RAG architecture was implemented. Financial documents were segmented into manageable chunks to balance contextual completeness and retrieval precision. Semantic embeddings were generated using a sentence-transformer model, selected for its ability to capture contextual similarity beyond keyword matching. These embeddings were indexed to a Weaviate vector database to enable efficient similarity-based retrieval. At inference time, the top-k most relevant chunks (e.g., k=3) were retrieved using Cross Encoder and concatenated into the prompt as contextual grounding. This approach aims to improve factual faithfulness by anchoring generations to source material. The value of k was chosen to

balance recall and noise: increasing k improves the likelihood of retrieving relevant information but may introduce irrelevant context that could confuse generation.

Retrieved context was clearly delimited within the prompt to distinguish source material from instructions. This design sought to reduce prompt ambiguity and encourage the model to extract information directly from retrieved evidence. The hybrid approach combines parametric knowledge (from fine-tuning) with non-parametric retrieval to mitigate hallucinations while maintaining domain fluency.

Evaluation was conducted through a multi-metric pipeline. ROUGE-L measured lexical alignment with reference answers, LLM-as-a-Judge assessed semantic correctness and faithfulness, and latency was recorded using execution time measurements. This multi-dimensional methodology enables comprehensive comparison across accuracy, reliability, and operational performance, providing an engineering-oriented assessment of each architecture's suitability for fintech deployment.

The Hallucination Audit: Numerical Failures in Tuning (Fine-Tuned Model

A focused audit of the Fine-Tuned model reveals a consistent and concerning pattern of numerical hallucination, particularly in questions requiring precise financial figures. While the model often demonstrated strong stylistic alignment and domain fluency, it struggled to reproduce exact values from the source material, and even minor numeric deviations constituted critical failures, making this behavior especially problematic.

One clear example involves the balance of intangible assets as of December 31, 2024. The reference value was $1,425, yet the Fine-Tuned model generated $1,421, and in another configuration produced entirely different figures, such as $2,142. Although the numbers appear superficially similar, this deviation reflects parametric approximation rather than factual recall. LLM-as-a-Judge evaluation penalized this response heavily (2/5), correctly identifying it as a factual error. Notably, ROUGE-L scores remained relatively high in some cases, highlighting a key limitation of lexical-overlap metrics: they may fail to detect small but financially significant numerical inaccuracies.

More severe hallucinations were observed in aggregate financial metrics. For example, when asked about total Platform Participant direct transaction costs for the Mobility segment, the correct

answer was $6,884, yet the Fine-Tuned model responded with $1,444. This large discrepancy suggests that the model did not retrieve or recall the specific reported value but instead generated a plausible-looking number consistent with similar financial patterns seen during training. Such behavior indicates that fine-tuning improves structural familiarity with financial reporting language but does not guarantee accurate memorization of quantitative details.

The model also failed in count-based questions. When asked how many jurisdictions were being evaluated for further licenses, the correct answer was four, yet the model responded with 12 jurisdictions. This represents a categorical factual fabrication rather than a minor approximation error. In financial disclosures, incorrect counts can materially misrepresent the company's strategy or regulatory exposure, further underscoring the risk of relying solely on parametric knowledge.

Another critical failure mode occurred when the source text explicitly stated that no specific information was provided. For instance, when asked what percentage of Gross Booking came from large metropolitan areas, the correct response was that the text did not contain this information. However, the Fine-Tuned model confidently generated approximately 80%, introducing unsupported numerical content. This demonstrates a strong bias toward answering rather than abstaining, even when evidence is absent. In compliance-sensitive environments, this tendency to fabricate rather than defer represents a major governance risk.

Finally, in regulatory threshold questions, the model substituted concrete monetary values for qualitative thresholds. The correct answer regarding income classification in Mexico was “more than one minimum salary a month,” yet the model produced specific peso amounts such as 20,000 Mexican pesos. This substitution reflects generative interpolation rather than evidence-based reasoning.

Overall, the audit demonstrates that the Fine-Tuned model exhibits systematic numerical instability, including approximation errors, fabrication of unsupported figures, and failure to abstain when information is missing. While linguistically coherent, its quantitative reliability is insufficient for high-stakes financial applications without retrieval grounding or additional verification safeguards.

Conclusion: When to Recommend Fine-Tuning vs. RAG for a Fintech Client

The experimental results demonstrate that architecture selection for a fintech client must be driven primarily by risk tolerance, regulatory exposure, and the criticality of numerical accuracy. Across evaluations, the Fine-Tuned model showed strong stylistic fluency, low latency, and coherent domain-specific language. However, it repeatedly produced incorrect numerical values, fabricated unsupported figures, and failed to abstain when information was unavailable. In financial services, where even minor numerical deviations can have regulatory, legal, or reputational consequences, this behavior represents a material risk. Therefore, fine-tuning alone is not recommended for high-stakes financial question answering involving audited financial statements, compliance disclosures, or regulatory thresholds.

Retrieval-Augmented Generation (RAG), particularly in the Hybrid configuration, combined with fine-tuning, demonstrated improved grounding and more consistent alignment with reference answers. Although numeric instability was not fully eliminated, retrieval reduced the frequency of unsupported claims and improved structural faithfulness. The Fine-Tuned + Hybrid RAG architecture offered the best balance between factual reliability and operational efficiency, with moderate latency and stronger contextual anchoring compared to the base model with RAG alone. In regulated fintech environments such as investor reporting, compliance review, or financial analytics dashboards, Hybrid RAG is the preferred architecture because it supplements parametric knowledge with verifiable document retrieval.

Fine-tuning is most appropriate when the task prioritizes tone, formatting consistency, or domain-specific language generation rather than exact numerical extraction. Examples include customer support automation, policy summarization, internal documentation drafting, or conversational interfaces where approximate reasoning is acceptable. In contrast, RAG should be mandatory when answering fact-based, document-grounded, or audit-sensitive queries.

For production deployment in fintech, the evidence suggests adopting a Hybrid RAG architecture augmented with additional safeguards, such as answer abstention logic and numeric validation layers. This layered approach mitigates hallucination risk while preserving efficiency. Ultimately, in financial systems where correctness outweighs creativity, retrieval-grounded generation is not optional; it is a governance requirement.