Phase 2: Ground the Domain (Research-Grade RAG)

1. Introduction

The objective of Phase 2 was to move from prompt experimentation in Phase 1 to a fully grounded, research-grade retrieval-augmented generation (RAG) system over my domain corpus. While Phase 1 focused on prompt structure and model behavior, Phase 2 required building an end-to-end system that supports traceable citations, reproducible evaluation, and measurable improvements. The emphasis in this phase was not on interface design, but on retrieval quality, structured grounding, logging, and trust behavior.

The core goal was to construct a baseline RAG pipeline, evaluate it across a structured 20-query set, implement at least one meaningful enhancement, and then compare the baseline and enhanced systems in a controlled manner. The guiding principle throughout this phase was that every major claim generated by the system must be supported by retrievable evidence from the corpus, and if the corpus does not support a claim, the system must explicitly say so rather than speculate.

This report documents the corpus construction process, the baseline RAG implementation (Version 1), the trust-tightening enhancement (Version 2), evaluation methodology, measurable improvements, and the remaining failure modes and tradeoffs observed.

2. Corpus and Data Manifest

The corpus consists of fifteen sources related to the research domain, including peer-reviewed papers, technical reports, and reputable AI research publications. At least eight of these are formal academic or technical works, satisfying the requirement that the corpus include credible and research-grade material. The remaining sources supplement the academic literature with high-quality technical documentation and structured research commentary relevant to the research question.

Each source is represented in a structured data manifest (`data/data_manifest.csv`). The manifest includes the following metadata fields: `source_id`, `title`, `authors`, `year`, `source_type`, `venue`, `url_or_doi`, `raw_path`, and a short `relevance_note`. The relevance note briefly explains why the source is included and how it contributes to the research question.

To ensure reproducibility and citation traceability, all raw artifacts (PDFs or text snapshots) are stored locally under `data/raw/`. The system does not rely on live URLs. Every citation generated by the RAG system references a `source_id` and, when applicable, a `chunk_id`,

both of which map directly to stored and processed text. This guarantees that every citation in an answer can be resolved to specific ingested content.

3. Baseline RAG Implementation (Version 1)

The baseline system (Version 1) follows a standard RAG pipeline architecture. During ingestion, each source is parsed and cleaned, and both raw and processed versions are stored. The processed text is then chunked using a fixed chunk size with overlap. Where possible, section boundaries from research papers are preserved to avoid fragmenting arguments across chunks.

The chunked text is embedded using a sentence-transformer embedding model and indexed in FAISS. At query time, the system retrieves the top-k most similar chunks based on vector similarity. These retrieved chunks are passed to the generation model along with a structured prompt that requires inline citations.

The generation step enforces structured citations in the format `(source_id, chunk_id)`. Free-form bibliographies are not allowed. All queries, retrieved chunks, generated answers, and prompt version identifiers are logged to disk. The system includes a one-command execution pipeline (`run_phase2.py`) that runs the full evaluation and scoring process, ensuring reproducibility.

Version 1 emphasized answer completeness and synthesis depth. The system attempted to answer nearly every query, including direct, synthesis, and edge-case prompts, while including structured citations wherever possible.

4. Evaluation Design

To evaluate the system rigorously, I constructed a set of twenty queries divided into three categories: ten direct factual queries, five synthesis or multi-hop queries, and five ambiguity or edge-case queries. The direct queries tested factual retrieval and explanation of specific methods or concepts. The synthesis queries required comparing multiple sources or integrating information across documents. The edge-case queries tested whether the system would correctly abstain when the corpus lacked supporting evidence.

Each answer was manually evaluated on three dimensions: groundedness (faithfulness to retrieved evidence), citation correctness (accuracy and traceability of citations), and usefulness (clarity and completeness of the response). In addition to numeric scores, each output was tagged with specific failure modes such as fabricated_reference, citation_mismatch, over_abstention, or missing_structured_citations.

This evaluation framework allowed for structured comparison between Version 1 and Version 2.

5. Version 1 Results

Version 1 demonstrated strong synthesis capabilities and generally high usefulness across direct and multi-hop questions. In several synthesis cases, the system successfully integrated information from multiple sources and produced well-structured, citation-backed explanations. The system rarely refused to answer and was generally willing to attempt a response even when evidence was partial.

However, several serious trust-related failures were observed. In four cases, the system exhibited fabricated reference behavior, including invented bibliography-style references or inconsistent citation formatting. In other instances, citations were present but did not cleanly support all claims, resulting in citation mismatches. One direct query included unsupported quantitative claims without structured citations.

These behaviors indicate that Version 1 operated as a high-recall system: it prioritized answering the question over strict evidence discipline. While this increased coverage and synthesis richness, it introduced citation integrity risks.

6. Enhancement: Version 2 (Trust Tightening)

Version 2 introduced stricter grounding and trust constraints. The generation prompt was modified to explicitly forbid fabricated citations and require abstention when sufficient evidence was not present in the retrieved chunks. The system was instructed to state “insufficient evidence” rather than speculate. Additionally, free-form reference-style outputs were eliminated, and citation behavior was constrained to structured `(source_id, chunk_id)` formats only.

No new retrieval mechanisms were introduced in this enhancement. Instead, the focus was on tightening trust behavior, improving citation discipline, and strengthening refusal calibration.

7. Version 1 vs. Version 2 Comparison

The comparison between versions reveals a clear tradeoff.

First, fabricated reference behavior was reduced significantly. Version 1 exhibited four clear fabrication failures, whereas Version 2 reduced this number to two. This represents approximately a 50% reduction in fabricated citation behavior, a measurable improvement in trust calibration.

Second, abstention behavior changed dramatically. Version 1 rarely refused to answer, even in ambiguous or unsupported cases. Version 2, by contrast, frequently abstained, particularly on synthesis queries. Several queries that Version 1 attempted to answer were labeled “insufficient evidence” in Version 2, even when partial evidence existed in the corpus. This indicates a shift toward high-precision, low-recall behavior.

Third, citation discipline improved overall in Version 2. Hallucinated bibliography behavior disappeared, and citation formats became more consistent. However, answer coverage decreased, particularly in multi-hop synthesis tasks.

Overall, Version 1 can be characterized as a high-recall, lower-precision system, while Version 2 behaves as a higher-precision, lower-recall system. The enhancement improved trust behavior but reduced synthesis depth and coverage.

8. Representative Failure Cases

One representative Version 1 failure involved fabricated bibliography behavior in a synthesis query. The system generated references that were not directly tied to ingested corpus entries. This violates the traceability requirement of research-grade systems and poses a significant reliability risk. Version 2 successfully eliminated this behavior.

Another failure case involved over-abstention in Version 2 during a synthesis task. Evidence supporting the question existed across multiple sources, but the system declined to answer due to overly strict grounding rules. This demonstrates that aggressive anti-hallucination constraints can suppress legitimate synthesis.

A third failure case in Version 1 involved citation mismatch. While citations were present and formatted correctly, they did not fully support all claims made in the answer. This highlights that citation presence alone does not guarantee faithfulness; alignment between claims and cited evidence must be verified.

9. Tradeoff Analysis

Phase 2 highlights a fundamental tension in research-grade RAG systems: groundedness, usefulness, and trust calibration cannot be optimized simultaneously without tradeoffs.

Version 1 prioritized usefulness and synthesis capability but allowed unsafe citation behavior. Version 2 improved citation integrity and refusal discipline but reduced answer completeness. Tightening hallucination resistance directly constrained synthesis depth.

This demonstrates that designing trustworthy RAG systems requires careful calibration between precision and recall. Over-penalizing speculative reasoning can degrade legitimate synthesis, while under-constraining citation behavior risks fabricated evidence.

10. Reproducibility and Engineering Quality

The system satisfies the required production patterns. All queries, retrieved chunks, prompt versions, and outputs are logged. The evaluation pipeline can be executed with a single command, which runs retrieval, generation, scoring, and summary metric reporting. Dependencies are pinned, and all raw data artifacts are stored locally. This ensures that graders can reproduce evaluation results and verify citation traceability.

11. Conclusion

Phase 2 successfully transformed a prompt-based research workflow into a structured, reproducible RAG system with measurable grounding behavior. The baseline system demonstrated strong synthesis ability but exhibited unsafe citation behavior. The enhanced system reduced fabricated references and improved trust calibration, though at the cost of increased abstention and reduced synthesis depth.

The resulting system is stable, reproducible, and citation-traceable. The primary remaining challenge is balancing strict grounding enforcement with synthesis capability. This tradeoff will inform the design decisions in Phase 3, where the RAG system will be wrapped into a usable Personal Research Portal product.