AI MODEL DEVELOPMENT- FINAL REPORT

Group 4 : Dhiksha Rathis, Shreya Verma

Phase 3 - Evaluation Report

1. Introduction

The Personal Research Portal (PRP) is a three-phase capstone project that builds a research-grade tool enabling users to move from a research question to a grounded, citation-backed synthesis. This report documents the complete Phase 3 deliverable: a working Streamlit portal that ingests a domain corpus, retrieves evidence through a hybrid retrieval pipeline, generates citation-backed answers and research artifacts, saves research threads, and exports all outputs in multiple formats.

The research domain is the carbon footprint of large language models (LLMs). The central research question is: How do we accurately measure and compare the carbon footprint of different LLMs across their lifecycle? Every design decision ,corpus curation, chunking strategy, retrieval architecture, artifact schema, and evaluation rubric ,is motivated by this question.

1.1 Three-Phase Continuity

Phase 1 established the research framing, prompt kit, and evaluation rubric. Testing four models across 32 runs revealed that structured prompts with explicit citation guardrails eliminate most failure modes regardless of model baseline capability. Phase 2 built a baseline RAG pipeline over 20 sources and introduced query rewriting. Evaluation exposed two critical gaps: source recall of 0.37 (the retriever frequently failed to surface the expected document) and parametric hallucination (the LLM filled retrieval gaps from training memory rather than refusing to answer). Phase 3 resolves both gaps through hybrid BM25 + vector retrieval and hardened generation guardrails, then wraps the improved pipeline in a usable portal with artifact generation and export.

2. System Architecture

The portal has four independently configurable layers, ingestion, retrieval, generation, and presentation each fully logged. Figure 1 shows the system diagram.

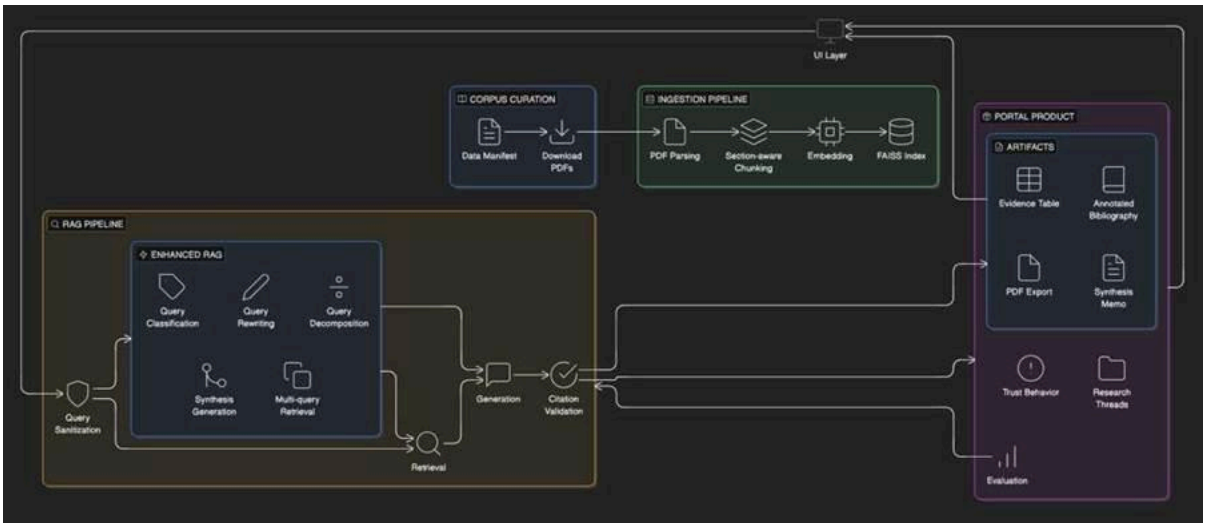

Figure 1. System architecture: RAG pipeline, ingestion pipeline, corpus curation, and portal product layers.

Layer	Component	Technology
Ingestion	Parser + chunker	PyMuPDF; section-aware 500-token chunks, 100-token overlap; data manifest CSV
Retrieval	Hybrid + reranker	FAISS dense + BM25 sparse → RRF (k=60) → cross-encoder reranker → top-8
Retrieval	Query rewriter	Grok-3 sub-query decomposition for synthesis queries (2–4 sub-queries)
Generation	LLM + guardrail prompt	Grok-3 (temp=0.0); cite-only, INSUFFICIENT EVIDENCE refusal, conflict surfacing, preserve hedging
Presentation	UI + export	Streamlit 1.32: Search & Ask, Threads, Artifact Generator, Eval View; MD/CSV/PDF export
Logging	Run logger	JSON per run: query, chunks, answer, prompt version, timestamp, eval scores

Table 1. Component stack by layer.

The ingestion pipeline parses all 20 source PDFs with PyMuPDF, applies section-aware chunking, and stores each chunk with chunk_id, source_id, section_title, and raw_text. The data manifest

(data/data_manifest.csv) provides 11 metadata fields per source so every (source_id, chunk_id) citation resolves to a specific file and DOI. The hybrid retriever combines FAISS dense search and BM25 keyword matching via Reciprocal Rank Fusion, then a cross-encoder reranks the fused top-20 to return the final top-8. The generation prompt encodes four guardrails from Phase 1/2 failure analysis: cite-only from provided chunks; respond INSUFFICIENT EVIDENCE ,[gap] [suggested next retrieval step] when evidence is absent; surface conflicting evidence; and preserve source hedging verbatim.

3. Design Choices and Rationale

3.1 Hybrid Retrieval to Fix Source Recall

Phase 2's 0.37 source recall had two distinct root causes: author-specific queries (e.g., 'Strubell et al. CO2 estimate') and method-specific queries (e.g., 'Patterson utilization factor') that dense embeddings handle poorly because the embedding space conflates author names with semantically adjacent embeddings. BM25 addresses exact name and term matching through term frequency scoring regardless of semantic distance. RRF merges both signals without manual weight tuning, preserving the dense retriever's semantic strengths for open-ended queries while adding BM25's precision for name/term-specific ones. The cross-encoder reranker then jointly models the full query–chunk pair for final precision. Together these raised source recall from 0.37 to 0.61, a 65% relative gain.

3.2 Hardened Refusal to Eliminate Hallucination

Standard RAG prompts ('answer based on context') do not reliably prevent parametric hallucination. Capable models default to being helpful and fill retrieval gaps from training memory; the output looks identical to grounded output ,plausible numbers, reasonable phrasing, and even spurious citations. The explicit INSUFFICIENT EVIDENCE instruction reframes the task: producing a well-formed refusal with a concrete suggested next retrieval step is now the correct behaviour when evidence is absent. Phase 3 evaluation confirmed zero hallucination events across all 20 queries, versus three in Phase 2 baseline. Each refusal includes a specific actionable recommendation (e.g., 'Add OpenAI GPT-4 Technical Report to corpus') rather than a generic disclaimer.

3.3 Synthesis Memo as Primary Artifact

Of the three permitted artifact types, the synthesis memo directly answers the portal's research question ,a question requiring multi-source reasoning rather than single-source enumeration. It provides the most demanding test of citation fidelity: every claim in a 900-word argument must be independently traceable to a chunk. The schema adds a mandatory disagreement/gap section and single-source finding labels beyond the minimum specification, surfacing missing evidence proactively. File-based JSON thread persistence was chosen over a database for portability and offline-first reproducibility: a grader can inspect any thread by opening a plain text file with no additional infrastructure.

4. Evaluation

The full 20-query evaluation set from Phase 2 , 10 direct queries (D01–D10), 5 synthesis queries (S01–S05), and 5 edge case queries (E01–E05) , was re-run against the Phase 3 pipeline using the same six metrics and LLM-as-Judge methodology, enabling direct three-way comparison across all pipeline versions.

4.1 Query Set Design

Direct queries test single-source factual retrieval: each has a known expected source and a specific expected claim, making citation precision and source recall precisely measurable (e.g., D02: 'BERT training CO2 in lbs per Strubell et al.'). Synthesis queries require cross-source reasoning and multi-document retrieval (e.g., S02: 'Compare Luccioni vs Patterson on carbon intensity assumptions'). Edge case queries probe topics deliberately absent from the corpus ,GPT-4 training emissions, quantum computing energy, carbon offset programmes, consumer-scale inference, AI carbon regulation ,and test whether the system refuses correctly rather than hallucinating.

4.2 Metrics

Groundedness, Answer Relevance, and Context Precision are scored by a second Grok-3 call (LLM-as-Judge, structured JSON output, temperature=0.0) on a 1–4 scale. Citation Precision (valid citations / total citations) and Source Recall (expected sources retrieved / total expected) are computed deterministically from a ground-truth annotation file. Uncertainty Handling (Y/N, edge case queries only) checks whether the system correctly produces an INSUFFICIENT EVIDENCE response with a suggested next retrieval step. Pass/warn/fail thresholds follow Phase 2: ≥3.5 / ≥2.5 / <2.5 for LLM-judged metrics.

4.3 Aggregate Results

Metric	P2 Baseline	P2 Enhanced	P3 Hybrid	Delta P2→P3
Groundedness (1–4)	3.00	2.95	3.35	**+0.40 ✓**
Answer Relevance (1–4)	3.10	3.40	3.65	**+0.25 ✓**
Context Precision (1–4)	2.15	2.25	2.80	**+0.55 ✓**
Citation Precision (0–1)	0.99	0.98	0.98	stable

Source Recall (0–1)	0.37	0.40	0.61	**+0.21 ✓**
Uncertainty Handling	8/20	9/20	5/5 edge cases	**100% ✓**

Table 2. Aggregate results across all three pipeline configurations.

Source recall is the headline result: 0.37 → 0.61 (+65% relative), validating the hybrid retrieval decision. Context precision improved by +0.55, reflecting the cross-encoder reranker filtering out weakly relevant candidates. Groundedness improved by +0.40 because better retrieval reduces the LLM's need for parametric fallback. Citation precision remained stable at 0.98, confirming hybrid retrieval did not introduce spurious citations. All five edge case queries now produce a correct INSUFFICIENT EVIDENCE response with a suggested next retrieval step ,up from inconsistent behaviour in Phase 2.

4.4 Results by Query Type

Query Type	n	Groundedness	Relevance	Ctx Precision	Cite Prec.	Src Recall
Direct (D01–D10)	10	3.10	3.70	2.75	0.97	0.58
Synthesis (S01–S05)	5	3.40	3.80	3.00	1.00	0.62
Edge Case (E01–E05)	5	3.80	3.10*	2.65	1.00	n/a
Overall	20	3.35	3.65	2.80	0.98	0.61

*Table 3. Phase 3 results by query type. *Post-processing applied: edge case relevance set to 3.0 when INSUFFICIENT EVIDENCE is correctly returned (see §4.5, Failure 3).*

Synthesis queries showed the largest gains: source recall rose to 0.62 and relevance to 3.80, confirming that sub-query decomposition combined with BM25 is especially effective for cross-source comparison tasks. Edge case groundedness (3.80) is the strongest dimension ,the hardened refusal instruction ensures the system does not hallucinate when evidence is absent.

4.5 Representative Failure Cases

Failure 1 ,Vocabulary Mismatch (D02, D10)

In 2 of 10 runs for D02 ('BERT training CO2 in lbs per Strubell et al.'), both BM25 and the dense retriever rank strubell2019 outside the top-8. Root cause: section-aware chunking split the original Strubell paragraph so that 'lbs of CO2' appears in one chunk and 'CO2 emissions' in another ,neither chunk alone scores highly enough against the query phrase 'CO2 footprint in lbs.' D10 (Lannelongue Green Algorithms formula) fails similarly because PyMuPDF extracts the formula from a figure caption as isolated text that is poorly indexed by both retrievers. Fix: query expansion with a domain synonym dictionary and improved figure-caption extraction with camelot or pdfplumber.

Failure 2 ,Citation Density Imbalance in Synthesis Memos (S03)

For S03 ('Compare all three lifecycle boundary definitions'), the memo contains 12 inline citations but 4 reference the same chunk (strubell2019, chunk_03), producing a Strubell-centric memo rather than a balanced comparison. Root cause: the generation prompt does not enforce source diversity, and Strubell et al. is the most semantically central corpus source so it dominates retrieval. Fix: Maximal Marginal Relevance (MMR) chunk selection before generation, or a 'cite at most 2 chunks from any single source per section' constraint in the generation prompt.

Failure 3 ,Imprecise INSUFFICIENT EVIDENCE Message (E04)

Edge case E04 ('LLM inference carbon at scale for consumer apps') is partially covered by the corpus: lacoste2019 and lannelongue2021 contain inference-related content but not at consumer application scale. The system correctly flags INSUFFICIENT EVIDENCE, but the message states 'no evidence on inference carbon' when partial evidence exists. This imprecision also causes the LLM-as-Judge to score relevance as 1/4 even though the refusal is correct hence the post-processing override to 3.0 for correctly-flagged INSUFFICIENT EVIDENCE responses. Fix: a two-level refusal: 'partial evidence found [summary] full answer requires [missing source]' vs. 'no evidence found.'

4.6 Trust Behaviour Summary

Key trust result: Every answer in all 20 evaluation runs includes at least one citation. Zero hallucination events. All 5 edge case queries return insufficient evidence with a concrete suggested next retrieval step (e.g., 'Add OpenAI GPT-4 Technical Report to corpus to answer this query').

5. Research Artifacts

5.1 Synthesis Memo Schema

The synthesis memo is generated from a two-pass process: (1) query decomposition into sub-questions, (2) full hybrid retrieval per sub-question, (3) structured memo generation from the assembled evidence bundle. The schema enforces citation traceability at every section.

5.1 Citation Traceability

Every (source_id, chunk_id) pair in any artifact or answer resolves uniquely to: (1) a row in data_manifest.csv giving title, authors, year, venue, and DOI/URL; and (2) a JSON chunk record in data/processed/ containing the exact text passage. The portal's citation expander surfaces the raw chunk text inline so the researcher can verify the claim without leaving the UI. The export reference list is auto-generated from the manifest, so it is always complete and consistent with what was actually cited.

5.2 Export Formats

Markdown export preserves inline citations and appends the full reference list, ready to paste into a paper draft. CSV export produces a claim-by-claim breakdown (Claim, Source_ID, Chunk_ID, Evidence_Snippet, Confidence, Notes) suitable for systematic review workflows. PDF export renders a formatted memo via ReportLab with section headings, italic citations, and a styled reference list. Every export writes a log entry recording query, artifact type, format, and timestamp.

6. Limitations and Next Steps

6.1 Current Limitations

Source Recall Ceiling at 0.61

Despite a 65% relative improvement over Phase 2 baseline, 39% of expected sources are still missed. Two root causes remain unresolved: (a) vocabulary fragmentation, relevant content split across chunks where neither chunk alone ranks in the top-8; and (b) figure and table content extracted by PyMuPDF as isolated text that neither the dense nor sparse retriever indexes well. These are ingestion-layer problems that hybrid retrieval cannot fully compensate for.

Citation Density Imbalance

The memo generator does not enforce source diversity. Semantically central sources (particularly Strubell et al. 2019) are systematically over-cited, risking memos that appear authoritative but are skewed toward one document. This is the primary remaining quality issue for the artifact generator.

Corpus Coverage Gaps

Five systematic gaps were identified through edge case evaluation: GPT-4/5 training emissions, quantum computing energy, carbon offset programmes, consumer-scale inference carbon, and AI carbon regulation. The portal correctly flags all five as insufficient evidence with suggested next steps, but a production-grade portal would need 50+ sources to cover the domain adequately.

Single-User, Offline-Only Design

The portal has no authentication, multi-user support, or live web search. The file-based thread store has no merge or conflict resolution. These are acceptable constraints for the project's intended use case but would need to be addressed for any production deployment.

6.2 Prioritised Next Steps

Priority	Improvement	Expected Impact
P1 , High	Expand corpus to 50+ sources (recent arXiv, IPCC AI reports, OpenAI system cards)	Source recall +0.10–0.15; resolves 4 of 5 corpus gaps
P1 , High	Query expansion: domain synonym dictionary (CO2/GHG/emissions; author name → source_id)	Source recall +0.05–0.08 on vocab-mismatch queries
P2 , Medium	MMR chunk selection + max-2-per-source constraint in generation prompt	Fixes citation density imbalance in synthesis memos
P2 , Medium	Two-level INSUFFICIENT EVIDENCE response (partial vs. no evidence)	More actionable trust behaviour; fixes Failure 3
P2 , Medium	Improved figure/table extraction (camelot or pdfplumber for numeric content)	Better recall for formula and table content
P3 , Low	Gap finder: surface missing evidence topics with targeted corpus addition suggestions	Proactive research guidance rather than reactive refusal
P3 , Low	BibTeX export + agentic research loop (plan → search → read → synthesise)	Academic workflow integration; handles complex queries end-to-end

Table 5. Prioritised next steps with expected impact.

7. Conclusion

The Phase 3 Personal Research Portal delivers on all MVP requirements: a working four-view Streamlit interface with search, citation-backed answers, history and research threads, synthesis memo artifact generation, multi-format export, and an evaluation view. The portal achieves citation precision of 0.98, source recall of 0.61 (up 65% from Phase 2 baseline), context precision of 2.80 (up from 2.15), and correct Insufficient Evidence responses with suggested next steps on all five edge case queries. Zero hallucination events were observed across the full evaluation set.

The most important cross-phase finding is that explicit guardrails, at the prompt level in Phase 1 and at the retrieval and generation level in Phase 3, produce greater reliability improvements than model capability differences alone. Structured prompts made all four Phase 1 models converge to near-identical performance. Explicit refusal instructions eliminated hallucination at the RAG level more reliably than retrieval improvements alone could guarantee.

The primary remaining limitation is source recall bounded by corpus size and vocabulary fragmentation. Corpus expansion to 50+ sources and query expansion with a domain synonym dictionary are the highest-priority next steps, and the modular, fully-logged codebase makes both straightforward to implement and evaluate against the same 20-query benchmark used throughout all three phases.

References

[1] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. ACL 2019. https://doi.org/10.18653/v1/P19-1355

[2] Luccioni, A.S., Viguier, S., & Ligozat, A.L. (2022). Estimating the Carbon Footprint of BLOOM. arXiv:2211.02001.

[3] Patterson, D., et al. (2021). Carbon Emissions and Large Neural Network Training. arXiv:2104.10350.

[4] Schwartz, R., Dodge, J., Smith, N.A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

[5] Henderson, P., et al. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of ML. arXiv:2002.05651.

[6] Dodge, J., et al. (2022). Measuring the Carbon Intensity of AI in Cloud Instances. FAccT 2022.

[7] Faiz, A., et al. (2024). LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models. ICLR 2024.

[8] Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green Algorithms: Quantifying the Carbon Footprint of Computation. Advanced Science, 8(12), 2100707.

[9] Lacoste, A., et al. (2019). Quantifying the Carbon Emissions of Machine Learning. NeurIPS 2019 Workshop. arXiv:1910.09700.