Phase 3: Personal Research Portal (PRP) Product

1. Introduction

Phase 3 extends the research-grade RAG system developed in Phase 2 into a structured Personal Research Portal (PRP) product. While Phase 2 focused on retrieval quality, citation discipline, and trust calibration, Phase 3 required transforming the system into a usable research tool with evaluation instrumentation, artifact exports, and comparative system analysis.

The central objective of this phase was not to improve UI polish, but to formalize the system as a research instrument. The PRP must support structured research workflows: running controlled query sets, logging outputs, exporting artifacts, comparing prompt variants, and measuring grounding behavior systematically. The emphasis was on architectural clarity, evaluation rigor, and reproducibility.

This report documents the final architecture, major design decisions, structured evaluation results, observed tradeoffs, system limitations, and proposed next steps.

2. Overall Architecture

The Phase 3 PRP maintains the core RAG architecture from Phase 2 but adds a product layer and formal evaluation instrumentation.

The system is composed of five primary components:

1. **Corpus Layer**
2. **Retrieval Layer**
3. **Generation Layer**
4. **Evaluation Layer**
5. **Product / Export Layer**

Each layer is modular and versioned.

2.1 Corpus Layer

The corpus from Phase 2 remains unchanged. It includes research-grade sources related to AI coding assistants and their effects on productivity, code quality, skill formation, and technical debt.

All raw artifacts are stored locally and referenced via structured metadata. Each chunk has a unique `(source_id, chunk_id)` identifier, ensuring traceability from output back to raw text.

The system does not rely on live URLs. Every citation resolves to stored content.

2.2 Retrieval Layer

Retrieval uses dense embeddings and vector similarity search over chunked documents. During evaluation, `top_k=6` chunks are retrieved for each query.

Each retrieved result includes:

- `source_id`
- `chunk_id`
- rank

This bounded retrieval context is injected into generation, enforcing evidence restriction.

No retrieval changes were introduced in Phase 3. Instead, the focus shifted toward generation behavior and evaluation instrumentation.

2.3 Generation Layer

The generation module enforces strict grounding rules:

- Use only provided evidence chunks.
- No external knowledge.
- No free-form bibliography.
- Inline citations must follow exact `(source_id, chunk_id)` format.
- If no evidence exists, output a structured “Insufficient evidence” statement.

In Phase 3, I introduced multiple system prompt variants to test how instruction strictness affects structural compliance.

Three prompt versions were evaluated:

- Prompt 1: Flexible structured citation enforcement.
- Prompt 2: Strict formatting constraints with explicit output template.
- Prompt 3: Highly conservative trust-tightened version.

This allowed controlled comparison of structural behavior under different constraint regimes.

2.4 Evaluation Layer (New in Phase 3)

Phase 3 introduces a formal evaluation script (`phase3_eval.py`) that:

- Runs a standardized 10-question research set.
- Logs full generated responses.
- Extracts inline citations.
- Parses reference sections.
- Validates citation format.
- Checks reference consistency.
- Logs retrieved chunks.
- Records prompt version.

Each evaluation produces a timestamped JSON artifact stored under:

```
None
logs/phase3_eval/<prompt_version>/
```

This transforms the system into a reproducible research instrument rather than a one-off RAG demo.

2.5 Product & Artifact Layer

The PRP now supports:

- Structured query threads
- Exportable evaluation artifacts
- Prompt version tracking
- Logged answer + citation extraction
- Structured result comparison

This enables side-by-side system comparisons and formal analysis of failure modes.

The system is now capable of supporting research workflows rather than ad-hoc question answering.

3. Design Choices

Phase 3 design decisions prioritized reproducibility and evaluability over conversational smoothness.

3.1 Instrumentation Over UI

Rather than investing time in interface polish, I prioritized:

- Logging
- Structured JSON exports

- Prompt version control
- Metric calculation
- Citation extraction validation

This aligns with the requirement that the PRP function as a research tool rather than a chat interface.

3.2 Structural Enforcement Through Prompting

Instead of post-processing citations, I chose to enforce strict formatting directly via system rules. This allowed me to test the limits of prompt-based structural compliance.

This decision revealed important reliability limitations, discussed below.

3.3 Comparative Prompt Testing

Rather than relying on a single prompt, I treated prompts as system variants and evaluated them systematically.

This mirrors experimental methodology:

- Controlled variable: system rules
- Fixed variables: corpus, retrieval, query set
- Measured outputs: citation validity, reference consistency, abstention behavior

This approach allowed quantitative comparison.

4. Evaluation Design

The evaluation uses 10 structured research questions across four themes:

- Productivity
- Code quality
- Technical debt
- Developer experience

Each query is processed with:

- `top_k=6` retrieval
- Structured generation
- Citation extraction
- Consistency checking

Metrics:

- Citation Valid Rate
- Reference Consistency Rate
- Insufficient Response Rate

This allows structural grounding evaluation independent of subjective usefulness scoring.

5. Evaluation Results

Across the three prompt variants:

Prompt	Citation Valid Rate	Reference Consistency Rate
Prompt 1	1.0	0.0
Prompt 2	0.9	0.0
Prompt 3	1.0	0.0

5.1 Primary Observation

While citation validity remained high (0.9–1.0), reference consistency remained 0% across all prompts.

In other words, the system frequently generated answers grounded in retrieved evidence, but it failed to produce a properly structured “References:” section matching inline citations.

This reveals a structural compliance failure.

5.2 Prompt Behavior Differences

Prompt 1:

- High recall.
- Verbose responses.
- Inconsistent formatting.
- Some citation drift.

Prompt 2:

- More disciplined formatting.
- Slight reduction in citation validity.
- Still no structured reference block.

Prompt 3:

- Highly conservative.
- Increased refusal behavior.
- Reduced synthesis depth.
- Still failed structured reference compliance.

This mirrors Phase 2 tradeoffs: tightening trust reduces coverage.

6. Failure Modes

Phase 3 revealed several structural failure modes.

6.1 Reference Section Generation Failure

Despite explicit instructions, the model rarely produced:

```
None
References:
(source_id, chunk_id)
(source_id, chunk_id)
```

This suggests that enforcing strict post-answer structural formatting through prompting alone is unreliable.

6.2 Over-Conservatism

The strictest prompt variant frequently refused to answer, even when evidence existed. This replicates the high-precision, low-recall behavior observed in Phase 2 Version 2.

6.3 Structural vs Semantic Grounding Gap

Citation validity checks confirm that referenced chunk IDs exist. However, structural compliance does not guarantee semantic alignment between claims and cited text.

This highlights the difference between:

- Format correctness
- True faithfulness

7. Tradeoff Analysis

Phase 3 further reinforces a core insight:

Groundedness, structural compliance, and synthesis capability cannot be simultaneously maximized.

Increasing structural rigidity:

- Reduces hallucination
- Increases refusal
- Decreases answer completeness

Reducing constraints:

- Improves coverage
- Increases formatting drift
- Risks unsafe citation behavior

This demonstrates that trust calibration is not binary but continuous.

8. Engineering Quality and Reproducibility

The PRP now satisfies research-grade engineering standards:

- All artifacts logged
- Prompt versions tracked
- Retrieval results stored
- Evaluation reproducible via single script
- Metrics automatically computed
- Dependencies fixed
- Raw data stored locally

The system is reproducible and auditable.

9. Limitations

Several limitations remain. First, structural enforcement via prompting is unreliable. A future version should programmatically generate the References section rather than relying on model formatting. Second, citation validation checks only structural existence, not semantic alignment. Third, retrieval remains embedding-only and does not incorporate hybrid ranking or cross-encoder reranking. Finally, UI remains minimal. The system functions as a research tool but not yet as a polished research portal interface.

10. Next Steps

Future improvements could include:

1. Programmatic post-processing of citations to guarantee reference consistency.

2. Semantic citation alignment scoring.
3. Hybrid retrieval (BM25 + dense embeddings).
4. Cross-encoder reranking.
5. Evidence highlighting interface.
6. Version comparison visualization within the UI.

These enhancements would move the PRP from research prototype toward production-grade research infrastructure.

11. Conclusion

Phase 3 successfully transformed a research-grade RAG system into a structured Personal Research Portal with evaluation instrumentation and comparative system analysis. While citation validity remained strong across prompt variants, structural reference consistency remained at zero, revealing a persistent compliance failure mode. This highlights a key limitation of prompt-only structural enforcement in grounded generation systems.

Rather than presenting a perfectly compliant system, this phase exposes meaningful reliability tradeoffs and demonstrates how they can be measured systematically. The resulting PRP is reproducible, instrumented, and capable of supporting structured research workflows. The primary remaining challenge is improving structural enforcement and semantic citation alignment without sacrificing synthesis capability. That tradeoff defines the next stage of system development.