Phase 2 Evaluation Report: Personal Research Portal (PRP)

Name: Lindberg Simpson

Domain: Remote Work and Productivity

1. Introduction

Phase 2 of the Personal Research Portal (PRP) implemented a systematic evaluation protocol using a fixed set of 20 queries designed to stress-test the system's ability to handle specific facts, complex synthesis, and ambiguous edge cases.

This report details the design of the query set, the implementation of automated metrics, and an analysis of the quantitative results derived from eval_run10.jsonl. It further documents specific failure modes and quantifies the impact of system enhancements, specifically structured citations and evidence scoring, demonstrating how the system addresses the reliability issues identified in the previous phase.

2. System and Corpus Overview

The PRP is built upon a specific corpus defined in data_manifest.csv. The system ingested a mix of peer-reviewed literature and working papers focused on the remote work landscape. The corpus includes high-density academic texts and quantitative reports, requiring the ingestion pipeline to handle complex formatting and metadata.

The transition from Phase 1 to Phase 2 involved moving from manual context injection to automated chunking and retrieval. The system now utilizes a vector store to index these documents, allowing for semantic search against user queries. A key design constraint carried over from Phase 1 was the requirement for "enforceable constraints." Consequently, the Phase 2 system was engineered to reject general knowledge answers in favor of responses grounded strictly in the retrieved chunk IDs (e.g., SRC001_chunk_0).

3. Query Set Design

To evaluate the system's performance comprehensively, a dataset of 20 queries was constructed (queries.jsonl). This set was stratified into three distinct categories to simulate different levels of user intent and difficulty: Direct, Synthesis, and Edge Cases.

3.1 Direct Queries (10 Queries)

The majority of the evaluation set consists of Direct Queries. These are designed to test the precision of the retrieval system and the model's ability to extract specific statistics, definitions, or methodological details. Examples include Q01: *"According to the Global*

*Survey… what is the average daily commute-time savings?"*and Q04: *"What share of paid workdays in the U.S. were full days worked from home?"*

These queries serve as a baseline for functionality. Success here requires the system to identify the exact source mentioned) and extract a numerical or categorical fact without hallucination. Phase 1 showed that models often gloss over specific numbers, so these queries police that behavior.

3.2 Synthesis Queries (5 Queries)

Synthesis Queries represent the higher-order research task of integrating findings across multiple documents. These queries force the RAG system to retrieve chunks from different source IDs and combine them into a coherent narrative. An example is Q13: *"Compare the findings of Gibbs et al. (SRC010) regarding 'focus time' with the findings in Emanuel et al. (SRC009)."* These queries test the system's ability to handle conflicting or complementary evidence. In the domain of productivity research, where one study may claim productivity gains and another losses, the system must present both viewpoints rather than smoothing them into a single, potentially inaccurate average.

3.3 Edge / Ambiguity Queries (5 Queries)

The final category, Edge Cases, tests the system's "trust behavior" and calibration. These queries ask for information that is either debatable, requires classification, or is entirely absent from the corpus. For example, Q20 asks: *"What is the exact causal effect (in %) of remote work on productivity in the U.S. economy as a whole?"* Since the corpus likely contains specific firm-level studies but not a definitive "U.S. economy" causal percentage, the correct behavior is for the system to state the lack of evidence or provide heavily qualified partial evidence. These queries are critical for detecting "hallucination," where the model invents a plausible-sounding answer to satisfy the user despite a lack of grounding data.

4. Evaluation Methodology and Metrics

Phase 2 implemented two automated metrics computed for every query response:

4.1 Groundedness (Mean: 0.409)

Groundedness measures the factual adherence of the generated answer to the retrieved context. In this implementation, it was calculated as the fraction of answer sentences that have a direct textual substring overlap with the retrieved chunks.

It is important to note that this is a conservative, "lower-bound" metric. Because it relies on substring matching, paraphrased content receives zero credit even if factually correct. Therefore, a score of 0.409 does not indicate 60% hallucination, but rather that roughly 41% of the generated output is a direct, verbatim quote from the source text. This metric rigorously penalizes creative writing and rewards strict adherence to the source material, aligning with the project's goal of accuracy over fluency.

4.2 Answer Relevance (Mean: 0.821)

Answer Relevance measures topical alignment by calculating the fraction of query keywords that appear in the generated response. A score of 0.821 indicates that the system is highly effective at staying on topic. The answers consistently address the entities and concepts raised in the user's prompt, suggesting that the retrieval mechanism is surfacing relevant chunks even if the generation step (measured by Groundedness) sometimes varies.

5. Quantitative Results

The evaluation run (eval_run10.jsonl) processed all 20 queries, providing a clear quantitative picture of the system's current capabilities.

5.1 Performance by Category

The data reveals distinct performance characteristics across the three query types:

Direct Queries: Achieved the highest relevance (0.872) and moderate groundedness (0.416). This confirms that when a specific fact is requested, the system usually finds the right topic and extracts quotes effectively.

Synthesis Queries: These showed slightly stronger groundedness (0.442) than direct queries, though lower relevance (0.792). This suggests that when forced to combine sources, the model relies more heavily on copying text directly from chunks to ensure accuracy, which boosts the substring-match score.

Edge Cases: These were the weakest performers (Groundedness: 0.361; Relevance: 0.749). The drop in relevance is significant as when the model struggles to find a direct answer to an ambiguous question, it tends to drift away from the specific keywords in the query, often offering tangential information instead of a direct "I don't know."

5.2 Best and Worst Performers

The strongest grounding was observed in Q01 (0.516) and Q15 (0.483). These queries requested specific definitions or numbers clearly present in the text (e.g., commute times in Aksoy et al.), allowing the model to simply lift the relevant sentence.

The clearest failure was Q20, which yielded a groundedness score of 0.264 and relevance of 0.548. This query asked for a precise causal percentage for the whole U.S. economy. The low scores reflect the system's struggle as it retrieved tangentially related chunks but could not quote a specific answer because none existed, leading to a low-match, low-relevance response that likely drifted into speculation.

6. Failure Case Analysis

A qualitative review of the logs reveals three distinct failure modes that define the current system's limitations.

Failure 1: Fabricated Specific Statistic (Q01)

While Q01 had high overall grounding, a close inspection of the output reveals a subtle hallucination. The query asked for the average daily commute-time savings. The model correctly identified the topic but introduced a specific numerical average that was a composite of different data points rather than a single number found in the text. This "math hallucination" is a known risk as the model attempted to calculate an average on the fly rather than retrieving a pre-calculated one. This inflated the relevance score but compromised factual integrity.

Failure 2: Hallucinated Structural References (Q07)

In answering Q07, the model referenced specific table numbers (e.g., "Table 3") that did not exist in the retrieved chunks. This is a "structural hallucination." The model, trained on academic papers, expects data to be in tables and will sometimes invent a table reference to make the answer sound more authoritative. This weakens citation reliability, as a user checking "Table 3" would find nothing.

Failure 3: Edge-Case Speculation (Q20)

This was the most significant failure regarding "trust behavior." When asked for the "exact causal effect" of remote work on the US economy, the system failed to cleanly state "no evidence found." Instead, it drifted into speculation, piecing together unrelated statistics to form an answer that sounded plausible but was not supported by the corpus. This resulted in the lowest groundedness score of the batch. This indicates that the "refusal" threshold in the prompt instructions needs to be tuned more aggressively for Phase 3.

7. Enhancements and Measurable Improvements

To address the issues seen in Phase 1, Phase 2 implemented three specific enhancements.

7.1 Structured Citations

In Phase 1, the model often alluded to sources vaguely (e.g., "One study suggests..."). Phase 2 implemented a strict output schema requiring inline citations formatted as (SourceID, ChunkID). The evaluation shows a 95% compliance rate with this format. This is the single largest improvement in the system. It transformed the output from a generic summary into a verifiable research tool. Users can now instantly audit a claim by cross-referencing the SRC ID with the manifest.

7.2 Evidence Strength Scoring

The system now appends a confidence tier (High/Medium/Low) to retrieved items based on their vector similarity scores. This enhancement improves transparency. While it does not prevent the model from trying to answer, it signals to the user when the retrieval layer is struggling. In the logs for Q20 (the failed edge case), the retrieval scores were noticeably lower, providing a warning signal that the subsequent answer should be viewed with skepticism.

7.3 Automated Evaluation Metrics

The move from manual inspection to automated scoring (Groundedness/Relevance) is a methodological enhancement that allows for regression testing. This enhancement successfully identified the systemic grounding ceiling of ~0.41, providing a concrete baseline that must be beaten in Phase 3.

8. Conclusion

The system in Phase 2 can ingest a complex domain corpus, understand structured queries, and generate answers with a level of citation discipline that was absent in Phase 1.

The quantitative results demonstrate that the system excels at Direct Queries (Relevance > 0.87) but struggles with Edge Cases where evidence is missing (Groundedness drops to < 0.30). The implementation of structured citations and automated logging has converted "hallucination" into a measurable error rate.

Moving to Phase 3, the priority must shift from simply retrieving text to better handling the "absence of evidence." Future enhancements must focus on "negative constraints" by teaching the model to refuse to answer when the retrieval scores fall below a safety threshold to fully realize the goal of a trustworthy Personal Research Portal.