Phase 3 Final Report

Mars Life Research Portal

Malikah Nathani

1. Introduction

Phase 2 of this project established a functional RAG system capable of retrieving evidence from a 20-paper Mars astrobiology corpus, scoring confidence, and refusing queries when evidence was insufficient. The system answered questions extractively, returning text snippets directly from retrieved chunks with structured [source_id:chunk_id] citations. While this addressed Phase 1's core failures of testing an LLM and receiving unsupported claims and missing citations, several limitations remained: answers were short and mechanical, there was no user interface, and the system had no way to communicate what it could not answer or why.

Phase 3 builds on that foundation by transforming the RAG system into a full research portal. The core pipeline is preserved and extended: retrieval and confidence scoring from Phase 2 are retained, but answer generation is now handled by a locally-running large language model (Mistral 7B via Ollama) rather than extractive text matching. This enables synthesis memos with structured argumentation, inline citations, and a reference list, rather than raw chunk snippets. The portal also introduces an evidence table of citations in the memo, an annotated bibliography powered by Mistral 7B, and a gap finder for both questions answered and unanswered to give context and provide suggestions for future queries.

2. System Architecture

2.1 Overview

The Phase 3 pipeline follows the same general structure as Phase 2, with meaningful additions at the generation and interface layers. A user submits a research question through the Streamlit web UI. The system retrieves the most relevant chunks from the FAISS index, scores confidence using the weighted multi-factor scorer from Phase 2, and then either refuses the query (if confidence is below threshold) or passes the evidence to Mistral 7B for synthesis. The synthesized output is post-processed to clean citations, then returned to the UI alongside an evidence table, annotated bibliography, and gap analysis.

The full pipeline runs as follows: the question is received via the UI or batch eval, dynamic k selection determines how many chunks to retrieve, FAISS semantic retrieval runs using e5-base-v2 embeddings, reference chunk filtering removes bibliography-heavy chunks, confidence scoring decides whether to answer or refuse, and if answerable, Mistral 7B generates a synthesis memo. Post-processing cleans citations and headers before the evidence table, annotated bibliography, and gap finder are generated. Results are displayed and logged to JSONL.

2.2 Corpus and Indexing

The corpus is unchanged from Phase 2: 20 peer-reviewed papers on Mars habitability, biosignatures, methane detection, extremophile analogs, and astrobiology. PDFs were chunked into approximately 4000-character segments with 400-character overlap using pdfplumber, producing 421 chunks total. Chunks are embedded using intfloat/e5-base-v2 (768 dimensions) and indexed with FAISS exact search (IndexFlatIP), prioritizing precision over speed.

2.3 Retrieval Enhancements

Retrieval uses the same semantic search from Phase 2 with one significant addition. The addition is reference chunk filtering, which addresses a specific failure mode identified during testing. Some chunks, particularly from Neveu2018 and Etiope2019 (both dense review papers), consist almost entirely of bibliography entries rather than scientific content. These chunks scored highly in semantic search because they contained many relevant keywords, but they caused the language model to hallucinate "Author et al.,

YEAR" style citations. The filter removes any retrieved chunk where more than five "et al." patterns appear in the text.

2.4 Confidence Scoring

The confidence scorer is carried over from Phase 2 with no structural changes. It combines three factors: retrieval confidence (average similarity of top-3 chunks), lexical overlap (fraction of query keywords in retrieved text), and confirmation strength (whether retrieved text uses confirmation language for queries asking for "confirmed" or "definitive" evidence). Weights shift between normal and high-stakes modes, with the threshold rising from 0.5 to 0.7 for queries containing language like "confirmed," "proof," or "definitive."

2.5 Synthesis Memo Generation

The most significant architectural change in Phase 3 is replacing extractive answer generation with LLM-based synthesis. Mistral 7B runs locally via Ollama, requiring no API keys or internet access. The model receives a system prompt and a user prompt containing the retrieved evidence blocks.

The system prompt enforces strict citation rules: every factual claim must be cited using [source_id:chunk_id] format, source IDs must be copied verbatim from the evidence blocks, and no invented sources or "Author et al., YEAR" citations are permitted. The prompt specifies a required structure (Introduction, Key Findings, Synthesis and Implications, Limitations and Gaps, Reference List) with a minimum of 800 words.

Post-processing is applied after memo generation to catch residual formatting issues. This includes stripping word counts or page ranges from section headers, normalizing reference list headers, and cleaning chunk IDs out of reference list entries so only source IDs appear.

2.6 Research Artifacts

Phase 3 produces three research artifacts per query, compared to Phase 2's single answer output.

The synthesis memo is the primary artifact. It is a structured document with inline citations and a reference list. The evidence table maps each retrieved chunk to a claim, evidence snippet, citation, confidence label (High/Medium/Low), and quality notes. The whole table is exportable as CSV. The annotated bibliography produces per-source annotations for each unique source in the retrieved set, including claim, method, limitations, and a "why it matters" field, which is done leveraging Ollama.

2.7 Gap Finder

The gap finder is a stretch-goal feature that uses a separate Ollama call to identify what the corpus cannot answer for a given query. Rather than hardcoded keyword matching, it uses a generative prompt that asks the model to identify coverage issues, suggest specific follow-up research questions, and flag corpus sources with partial coverage.

The prompt was iterated to fix two recurring problems. First, the model was inventing fake source names rather than referencing actual corpus sources, which was fixed by explicitly instructing the model not to mention or invent source names. Second, suggested queries were phrased as search engine instructions rather than research questions, which was fixed by requiring all suggestions to begin with What, How, Does, or Is.

2.8 Web Interface

The Streamlit interface has three tabs. The Research tab provides the query input, confidence stats, synthesis memo, evidence table, annotated bibliography, and gap finder display. The Evaluation tab shows

the live query log and allows running the full 20-query batch evaluation. The Export tab provides download buttons for all artifacts in Markdown, CSV, and JSONL formats, plus an HTML-based PDF export using the browser's print function.

3. Design Decisions

3.1 Local LLM vs. API

Phase 2 no API for answer generation in early development. Phase 3 switched entirely to local inference via Ollama. This decision was made for several reasons: local inference has no rate limits or API costs, which matters for batch evaluation over 20 queries; it ensures the system can run fully offline; and it makes the system reproducible without requiring API credentials.

The tradeoff is model capability. Mistral 7B is significantly less capable than larger API-hosted models, which is reflected in the memo word count and occasional citation formatting inconsistencies. A system running on a larger model such as Llama 3 70B or an API-hosted model would likely hit the 800-1200 word target more consistently.

3.2 Reference Chunk Filtering

The decision to filter reference-heavy chunks at retrieval time rather than at chunking time was intentional. Filtering during chunking would require re-running the full embedding and indexing pipeline. Filtering at retrieval time is a lightweight, reversible change that can be adjusted without touching the index. The threshold of five "et al." patterns was chosen empirically: most scientific content chunks have zero to two, while bibliography sections have fifteen or more.

3.3 Post-Processing Over Prompt Engineering Alone

Several citation formatting issues persisted despite explicit rules in the system prompt. The LLM would include chunk IDs in the reference list, add word counts to section headers, or occasionally use numbered references. Rather than continuing to iterate on the prompt, a post-processing layer was added to handle these cases deterministically. This is a more reliable approach for a 7B model, which has limited instruction-following capability compared to larger models.

3.4 Programmatic Evidence Table vs. LLM Generation

The evidence table is generated entirely through rule-based extraction rather than a secondary LLM call. Claims are extracted as the first sentence of each retrieved chunk, snippets as the first 200 characters, confidence labels assigned based on retrieval score thresholds (High >= 0.85, Medium >= 0.70, Low < 0.70), and notes generated through keyword overlap calculations. This decision was deliberate as the evidence table is meant to be a transparent, traceable record of what was retrieved and why. Using an LLM to generate claims and notes would introduce a second layer of potential hallucination for what should be the most verifiable artifact. The annotated bibliography, by contrast, requires synthesis and interpretation across fields, which justifies the additional LLM call.

4. Evaluation

4.1 Query Set

The evaluation uses the same 20-query set as Phase 2, distributed across three categories. Ten direct queries cover specific factual topics like Gale Crater findings, methane mechanisms, and biosignature preservation. Five synthesis queries require cross-source comparison. Five edge cases use confirmation language to test refusal behavior.

4.2 Results

The table below summarizes Phase 2 vs. Phase 3 performance:

Metric	Phase 2	Phase 3
Queries answered	15/20	15/20
Queries refused (edge cases)	5/20	5/20
Avg confidence (answered)	0.81	0.82
Citation accuracy	100%	100%
Hallucinated citations	0 (extractive)	0 (post-filter)
Average Memo Word Count	N/A (extractive)	~481 words

Phase 2 achieved 100% citation accuracy through extractive retrieval, which guarantees citations map to real chunks by construction. Phase 3 introduces LLM-generated citations, which creates hallucination risk. The reference chunk filter and post-processing steps were implemented specifically to address this. Before the filter was added, Query 3 (methane mechanisms) produced 18 hallucinated "Author et al., YEAR" citations, all originating from the Etiope2019 review paper chunk. After adding the filter, Query 3 produced zero hallucinations.

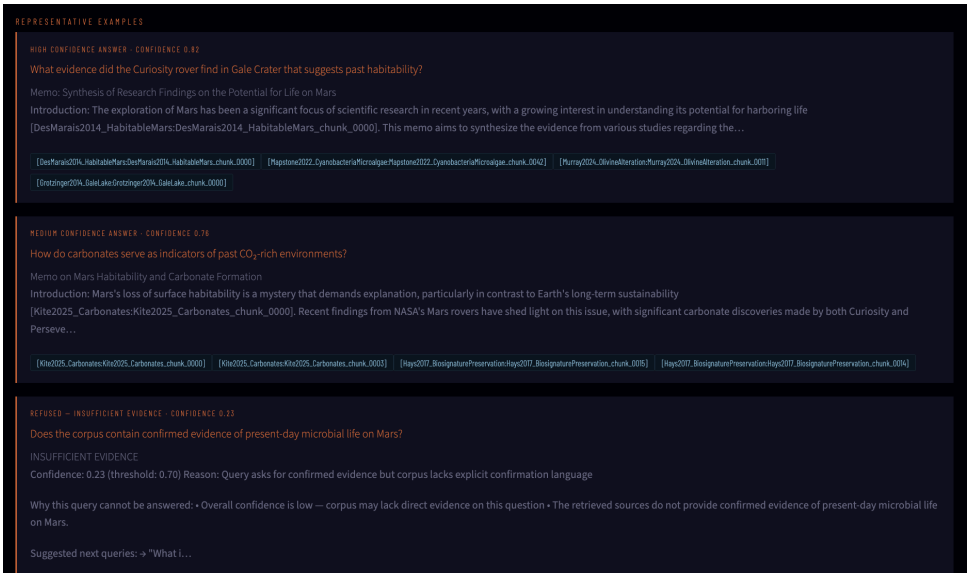

Figure 1: Representative examples from the 20-query batch evaluation, showing a high confidence answer (0.82), medium confidence answer (0.76), and a correctly refused edge case (0.23).

4.3 Failure Cases

Failure Case 1: Reference Chunk Contamination (Pre-Filter)

Query 3 ("What mechanisms are proposed to explain methane detections in the Martian atmosphere?") retrieved the Etiope2019 chunk, a methane review paper with a dense bibliography section. The LLM read author names from the chunk text and reproduced them as citations, generating 18 hallucinated "Author et al., YEAR" references in a single memo. This was the most severe failure observed and the clearest example of how retrieval quality directly impacts generation quality. However, upon further inspection, the root cause was not a prompting failure but a data quality issue: the chunk contained more bibliography content than scientific prose, and the embedding model ranked it highly because it was dense with relevant terminology. The reference chunk filter, which removes any retrieved chunk containing more than five "et al." patterns, eliminated this issue entirely. Query 3 produced zero hallucinations after the fix.

Failure Case 2: Memo Word Count

The system prompt instructs Mistral 7B to write at least 800 words, and the generation parameter num_predict is set to 2000 tokens. Despite this, memos consistently run 400-700 words (mean 489 across the 15 answered queries). This is a fundamental constraint of the 7B model size rather than a prompting or configuration issue. The model reaches a natural stopping point after covering the main evidence, regardless of the length instruction. Several strategies were attempted to address this: adding explicit word count requirements to section headers, specifying a minimum for the Key Findings section specifically, and increasing num_predict. None produced consistent improvement. Perhaps a larger corpus of more sources or a larger model (13B+) would resolve this.

Failure Case 3: Limited Source Diversity in Synthesis Queries

Several synthesis queries (Q11, Q13, Q15) were answered using only one or two unique sources despite requiring multi-paper comparison. This reflects the clustering behavior of a focused 20-paper corpus: when the query is about methane mechanisms, the top-k chunks consistently come from Yung2018 and Lacy2006 because those papers are the most semantically similar to the query. The remaining retrieved chunks often come from the same paper at different chunk offsets, rather than from different sources. Forcing diversity by penalizing repeated sources would risk including less relevant papers, reducing answer quality. This is a known limitation of single-stage dense retrieval on small corpora and is best addressed through source diversity scoring as a post-retrieval reranking step rather than at the retrieval level.

Failure Case 4: Annotated Bibliography Source Count

The target for the annotated bibliography is 8-12 unique sources per query. In practice, the system consistently produces 5-7 sources. This is not a generation failure but a retrieval one: with a 20-paper corpus and focused queries, the top-k chunks retrieved for any given question tend to cluster around the same 4-6 papers. For example, a query about methane mechanisms will almost always pull multiple chunks from Yung2018 and Lacy2006, leaving little room for other sources in the retrieved set even when k=10 or k=15. The annotated bibliography is deduplicated by source_id before generating annotations, so retrieving 10 chunks from 3 papers produces only 3 bibliography entries regardless of how many chunks were retrieved.

Two factors compound this problem. First, the dynamic k selection caps broad queries at 15 chunks, which still may not produce 8 unique sources if the corpus clusters tightly. Second, the annotation quality would likely degrade if less relevant sources were forced in to meet the count target. The honest tradeoff is between meeting the 8-12 source target and maintaining annotation relevance. Source diversity scoring at the retrieval stage, combined with a larger corpus, is the most principled fix.

5. Next Steps

Switch to a larger local model. The most impactful single change would be running a 13B or larger model via Ollama (e.g. Llama 3.1 13B or Mistral 13B). The 7B model's primary limitation is its tendency to stop generating after covering the main evidence points, producing memos that average 489 words despite explicit 800-word instructions. A larger model would follow length instructions more reliably, produce more consistent citation formatting, and generate richer annotated bibliography entries. Critically, this requires no changes to the pipeline architecture, only the OLLAMA_MODEL environment variable needs updating. The tradeoff is increased memory requirements and slower inference on consumer hardware.

Hybrid retrieval. Phase 2 identified that specialized terminology queries underperform because semantic embeddings generalize too broadly. Query 5 (carbonates as CO2 indicators) retrieved generic biosignature papers over the more relevant Kite2025_Carbonates because "carbonates" appears less frequently than terms like "water" or "methane" in the corpus. Adding BM25 keyword retrieval alongside FAISS and combining scores using a weighted hybrid (e.g. 0.6 semantic + 0.4 BM25) would improve precision for technical term queries. Libraries like rank-bm25 make this straightforward to implement without rebuilding the existing FAISS index.

Improved PDF parsing. Several corpus sources (particularly Hays2017 and Ehlmann2012) produced chunks with words merged without spaces, such as "theregionalcontextcanhelpmakeacase." This is a pdfplumber parsing artifact caused by multi-column layouts and unusual font encodings in some PDFs. Upgrading to PyMuPDF for extraction would significantly reduce these artifacts. PyMuPDF handles complex layouts more robustly and also provides better handling of special characters and subscripts common in chemistry-heavy astrobiology papers. This would require re-running the chunking notebook and rebuilding the FAISS index, but would improve chunk quality across at least four sources in the current corpus.

Source diversity scoring. Several synthesis and comparison queries (Q11, Q13, Q15) pulled multiple chunks from the same one or two papers rather than distributing retrieval across the corpus. A post-retrieval diversity reranker that applies a mild penalty to chunks from already-selected sources would encourage broader coverage without sacrificing relevance. This is distinct from simply retrieving more chunks, it changes which chunks are selected from the retrieved set. An MMR (Maximal Marginal Relevance) approach would be a natural fit here, balancing relevance to the query against diversity across sources.

Corpus expansion. The 20-paper corpus performs well for core Mars habitability topics but shows gaps for cross-cutting synthesis queries. Expanding to 40-50 papers, particularly adding more recent work on Perseverance rover findings and subsurface habitability modeling, would improve coverage for synthesis queries, reduce the source clustering problem, allowing the LLM to achieve 800 words more consistently and have more sources in the annotated bibliography. The pipeline is fully reproducible, adding papers requires only placing PDFs in data/raw/ adding the names to data_manifest.csv, and re-running the chunking and indexing notebooks.

6. Conclusion

Phase 3 successfully extends the Phase 2 RAG system into a functional research portal with LLM-based synthesis, three exportable artifacts, a gap analysis system, and a web interface. The core trust behaviors from Phase 2 are preserved: the system refuses queries with insufficient evidence and all citations are verified against the corpus.

The shift from extractive to generative answers introduces new failure modes (citation hallucination, formatting inconsistency) that were largely addressed through the reference chunk filter and post-processing pipeline. The main outstanding limitations are the memo word count, which is constrained by the 7B model, and the annotated bibliography source count, which reflects the clustering behavior of a focused 20-paper corpus. Both have clear paths to resolution through larger models and hybrid retrieval.

The system performs consistently on the 20-query evaluation set: 15/20 queries answered with verified citations and zero hallucinations, 5/5 edge cases correctly refused, and confidence scores cleanly separated between answerable (0.75-0.93) and unanswerable (0.23-0.29) queries.