**📄 Proposal Content**

**Title:** **Proposal for Enhancing the UniversalRAG Framework with Persistent Storage and Hybrid Search**

**Author:** Kratu **Date:** 1 September 2025

**1. Executive Summary**

This proposal outlines a plan to extend the UniversalRAG architecture with:

* Persistent FAISS vector store storage for rapid restarts.
* Hybrid search (BM25 + embeddings) for improved retrieval accuracy.
* Automated ingestion pipelines for continuous corpus updates.

The goal is to improve retrieval precision, reduce cold‑start latency, and enable long‑term scalability.

**2. Background**

The current UniversalRAG implementation supports:

* Multi‑format ingestion (web, PDF, Markdown, CSV, DOCX).
* Adaptive chunking and embedding with all-mpnet-base-v2.
* FAISS in‑memory vector store for retrieval.

However, the system requires re‑embedding on every restart, and retrieval accuracy can degrade for certain query types.

**3. Proposed Enhancements**

**3.1 Persistent Vector Store**

* Store FAISS index on disk.
* Implement load‑on‑startup to bypass re‑embedding.
* Schedule periodic index backups.

**3.2 Hybrid Search**

* Combine BM25 keyword search with dense embeddings.
* Merge results via reciprocal rank fusion.
* Improve performance for factoid and keyword‑heavy queries.

**3.3 Automated Ingestion**

* Monitor RSS feeds and file directories.
* Trigger ingestion jobs on new content.
* Maintain freshness without manual intervention.

**4. Implementation Plan**

| **Phase** | **Task** | **Duration** | **Owner** |
| --- | --- | --- | --- |
| 1 | Add FAISS persistence | 1 week | Backend |
| 2 | Integrate BM25 hybrid search | 2 weeks | Backend |
| 3 | Build ingestion scheduler | 1 week | DevOps |
| 4 | QA & Benchmarking | 1 week | QA Team |

**5. Risks & Mitigation**

| **Risk** | **Likelihood** | **Impact** | **Mitigation** |
| --- | --- | --- | --- |
| Index corruption | Low | High | Automated backups |
| BM25 latency | Medium | Medium | Cache frequent queries |
| Ingestion overload | Low | Medium | Rate limiting |

**6. Expected Outcomes**

* **50% faster startup** due to persistent index.
* **+10% retrieval accuracy** on benchmark datasets.
* **Reduced manual maintenance** via automation.

**7. Conclusion**

By implementing persistence, hybrid search, and automation, UniversalRAG will become more robust, accurate, and scalable — ready for production‑grade deployments