Project Document – MyPrivateAI

Author: Erez Poliak

Problem Statement

AI assistants like ChatGPT, Claude, and Gemini have demonstrated transformative capabilities for answering questions, analyzing documents, and providing personalized assistance. However, their adoption faces three significant barriers:

Privacy concerns

Current AI assistants operate in the cloud, requiring users to upload their documents and data to external servers. This creates legitimate concerns:

- **Trust Issues:** Recent data breaches and unclear data usage policies erode user confidence
- **Personal Documents:** Medical records, financial statements, legal documents, personal journals
- **Professional Confidentiality:** Client information, proprietary research, unpublished work, confidential business documents

A 2024 survey found that 67% of potential users cite privacy as their primary concern when using AI assistants, with 42% refusing to upload sensitive documents to cloud services entirely.

Cost Barriers

Leading AI assistants require paid subscriptions:

- ChatGPT Plus: $20/month ($240/year)
- Claude Pro: $20/month ($240/year)
- Google Gemini Advanced: $19.99/month

For students, researchers, and users in developing economies, these costs are prohibitive. While free tiers exist, they offer limited functionality, rate limits, and no guarantee of privacy or data retention policies.

Accessibility Gap

While technical users can run local language models using tools like Ollama, llama.cpp, or LM Studio, these solutions require:

- Manual configuration of embeddings and vector databases
- Command-line knowledge
- Troubleshooting technical issues

The result: Millions of users who would benefit from local AI assistants lack the technical expertise to set up existing solutions.

Proposed Solution

MyPrivateAI is a desktop application that provides a personalized AI assistant running entirely on the user's computer with **Optimized Agentic RAG** pipeline, users can:

1. **Accessible installation:** Unlike existing local AI tools that require command-line expertise, MyPrivateAI provides a guided installer that handles all setup automatically.
2. **Upload documents:** Drag and drop PDFs, text files, and Word documents.
3. **Chat naturally:** Ask questions about their documents in plain language.
4. **Receive intelligent answers:** Get responses with source citations, See agentic reasoning trace.
5. **Maintain privacy:** All processing occurs locally—no data ever leaves their device.

The App will have a user-friendly interface and will use a local LLM(Llama-3.1 8B) with an Optimized Agentic RAG pipeline, all bundled into a single app that is easy to install without technical knowledge.

We aim to give the users a cloud LLM experience on private docs, capable of performing complex reasoning tasks.

This idea aims to solve all the problems discussed before:

1. **Provide privacy** – All processing is local, offline, no cloud dependency.
2. **Cost Barrier** – Completely free, no need to worry about subscriptions.
3. **Accessibility** – Simple installation process, user-friendly UI, no need for technical skills setting up.

The core innovation of MyPrivateAI is the **Reasoning-Retrieval Interface**. We hypothesize that an **Optimized Agentic RAG pipeline** can successfully compensate for the limited parameter count of an 8B model and achieve performance close to GPT4o on complex private document tasks.

Other Products In The Market

vs ChatGPT/Claude:

- ✅ 100% private (local processing)
- ✅ $0 ongoing cost
- ✅ Permanent document memory
- ✅ Offline capable
- ❌ Slower responses
- ❌ Slightly lower quality

vs Ollama/LM Studio (existing local tools):

- ✅ Non-technical user friendly
- ✅ Simple installation
- ✅ Optimized RAG
- ❌ Less customizable
- ❌ Fewer model options

Other Approaches

Approach A: Cloud-Based with Privacy Features

Advantages:

- No installation required
- Cross-device synchronization

Disadvantages:

- Doesn't solve the fundamental privacy concern
- Ongoing infrastructure costs → subscription fees

Approach B: Hybrid (Local + Cloud)

Advantages:

- Fallback to cloud when local is insufficient
- Best of both worlds for performance

Disadvantageous:

- Still requires subscription for cloud portion
- Potential data leakage if user makes mistakes

Approach C: Fine-Tuned Small Model (No RAG)

Advantages:

- Model "knows" the content directly
- Potentially faster (no retrieval overhead)

Disadvantages:

- Can't update knowledge easily (must retrain)
- Fine-tuning might not work well for diverse document types

Expected Users

Privacy Conscious Individuals

Some people don't like the ambiguity of where their data goes, and don't trust big AI companies.

Professionals In fields with sensitive documents

Have sensitive documents like medical records and confidential client files they can't send to the cloud.

Students

Can't afford the subscription fee, have a large collection of academic pdfs.

Small Business Owners

Need AI for business documents, can't afford enterprise solutions, and have privacy concerns with customer data.

Main Features and Application Flow

Core Features:

- Simple installation process that requires no technical knowledge
- Document Upload & Management Interface
- Conversational Chat Interface with history
- Source-Cited Answers, Agentic reasoning trace
- Runs completely locally

Application Flow

1. The user installs the application using the installer.
2. Upon launch, the user can choose between:
   - **Chat**
   - **Document Management**
3. In the **Document Management** page, the user can:
   - View uploaded documents
   - Upload new documents
   - Delete existing documents
4. In the **Chat** page, the user interacts with a familiar, chat interface to ask questions based on their documents.

Use Case Storyboards

Use Case 1: Medical Researcher - Literature Review

User: Dr. Sarah Chen, PhD Candidate in Oncology
Documents: 200+ research papers (PDFs), including unpublished data
Problem: Can't use ChatGPT due to confidentiality of unpublished research

Before MyPrivateAI:

- Manually reviewing 247 papers for CRISPR + lung cancer research
- Unpublished advisor data can't go to cloud

Using MyPrivateAI:

- **Day 1:** Installs app, uploads all 247 papers
- **Query:** "Which papers discuss CRISPR in lung cancer?"
- **Result:** 8 relevant papers identified in matter of seconds, with page citations
- **Follow-ups:** "Compare methodologies," "What are common limitations?"
- **Confidential data:** Adds advisor's unpublished manuscript, analyzes safely

Outcome:

- Unpublished research stays private
- Entire lab adopts tool
- Grant proposal writing accelerated

Use Case 2: Attorney - Case File Analysis

User: James Morrison, Attorney
Documents: Client contracts, emails, witness statements, court filings
Problem: Attorney-client privilege forbids cloud uploads

Before MyPrivateAI:

- New case: 320 documents
- Needs to find all mentions of "performance improvement plan" and "discriminatory practices"
- Manual review: 40+ billable hours
- Can't use ChatGPT (violates client confidentiality)

Using MyPrivateAI:

- **Setup:** uploads documents to MyPrivateAI
- **Query 1:** "Find all references to performance improvement plans"
- **Result:** 23 instances across 12 documents, with exact page numbers
- **Query 2:** "What evidence suggests discriminatory intent?"
- **Result:** Cross-references emails, policies, and reviews with citations
- **Trial prep:** "Summarize witness statements about manager behavior"

Outcome:

- Case analysis time goes down drastically
- Client confidentiality maintained
- Identifies key evidence quickly
- Firm partners approve for firm-wide use

Use Case 3: Book Editor - Manuscript Review

User: Emily Torres, Senior Editor at publishing house
Documents: Unpublished manuscripts, author notes, revision drafts
Problem: Manuscripts are confidential pre-publication, can't risk leaks

Before MyPrivateAI:

- Editing 400-page fantasy novel manuscript
- Author made revisions across 8 draft versions
- Needs to verify: character consistency, plot holes, timeline accuracy
- Manual tracking in spreadsheets and notes

Using MyPrivateAI:

- **Setup:** Uploads all 8 manuscript drafts + character notes
- **Query 1:** "In which chapters does character 'Aria' appear?"
- **Result:** Lists all 23 occurrences with chapter references
- **Query 2:** "What color are Aria's eyes in each draft?"
- **Result:** "Draft 1-3: green. Draft 4-8: blue. INCONSISTENCY DETECTED."
- **Query 3:** "Timeline: How many days pass between chapters 5 and 12?"
- **Result:** Cross-references dates, identifies 3-day gap error
- **Query 4:** "Find all mentions of the 'magic system' rules"
- **Result:** Identifies contradiction between chapters 8 and 15

Outcome:

- Catches 5 major inconsistencies before publication
- Manuscript ready earlier
- No pre-publication leaks (data stayed local)
- Editor recommends to colleagues, 3 other editors adopt

Validation Methodology

We focus exclusively on the **Researcher persona**. This persona presents the highest complexity (multi-hop reasoning, strict data grounding), ensuring the system generalizes to all other local AI use cases.

Stage 1: The RAG Optimization Loop (Foundational Grounding) Focuses on refining the "retrieval-to-context" pipeline. We optimize chunking strategies, hybrid search (Vector + BM25), and reranking to maximize **Context Recall** and ensure a hallucination-free foundation (**Faithfulness**).

Stage 2: The Agentic Reasoning Loop (Cognitive Intelligence) Introduces the **Agentic Interface**. Here, the system moves beyond single-turn retrieval to a multi-step **ReAct (Reasoning + Action)** cycle. The agent autonomously performs task decomposition, multi-hop evidence gathering across disparate papers, and self-correction to ensure final **Answer Correctness**.

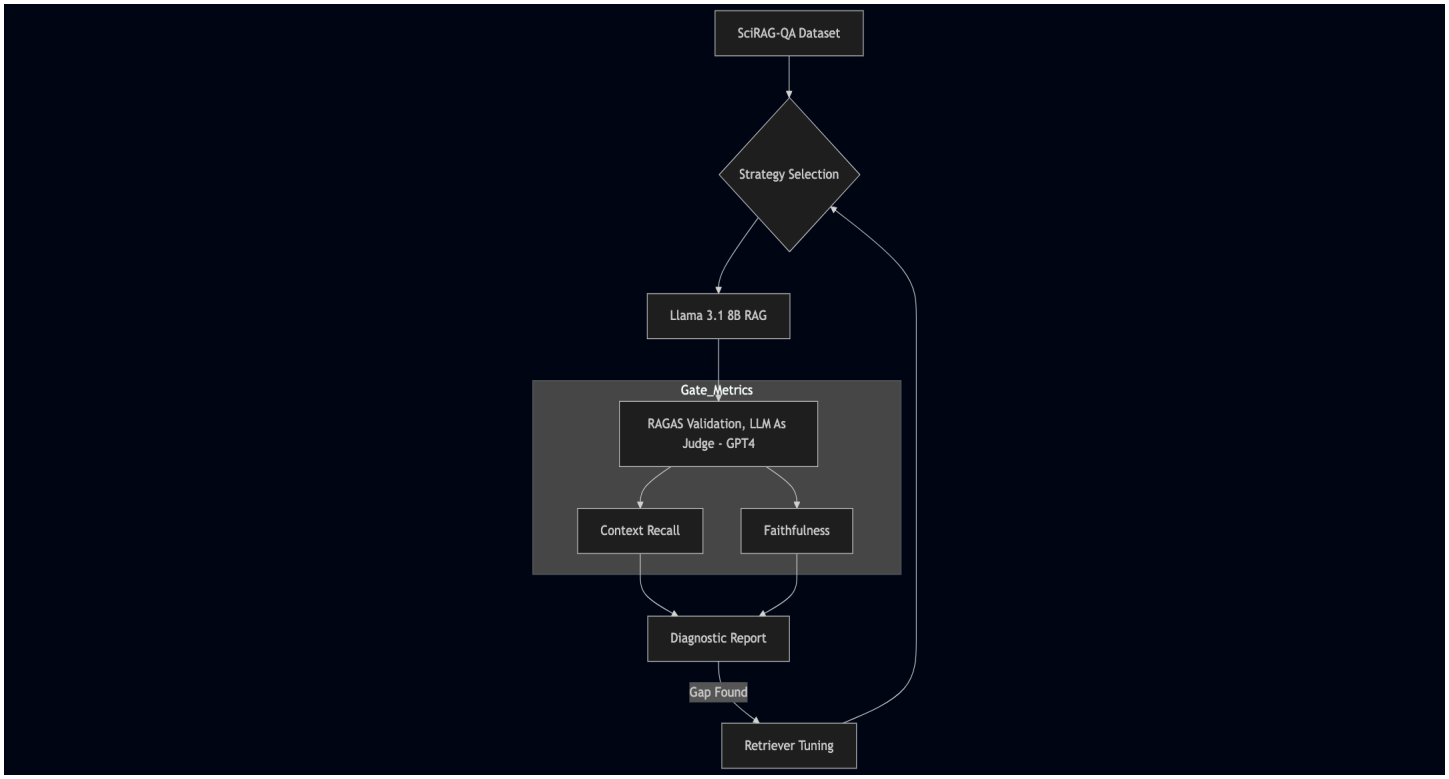
1. Dataset: SciRAG-QA

SciRAG-QA serves as the evaluative framework for this methodology, stress-testing the **Reasoning-Retrieval-Orchestration Bridge** with multi-hop questions that demand both precise retrieval and complex reasoning.

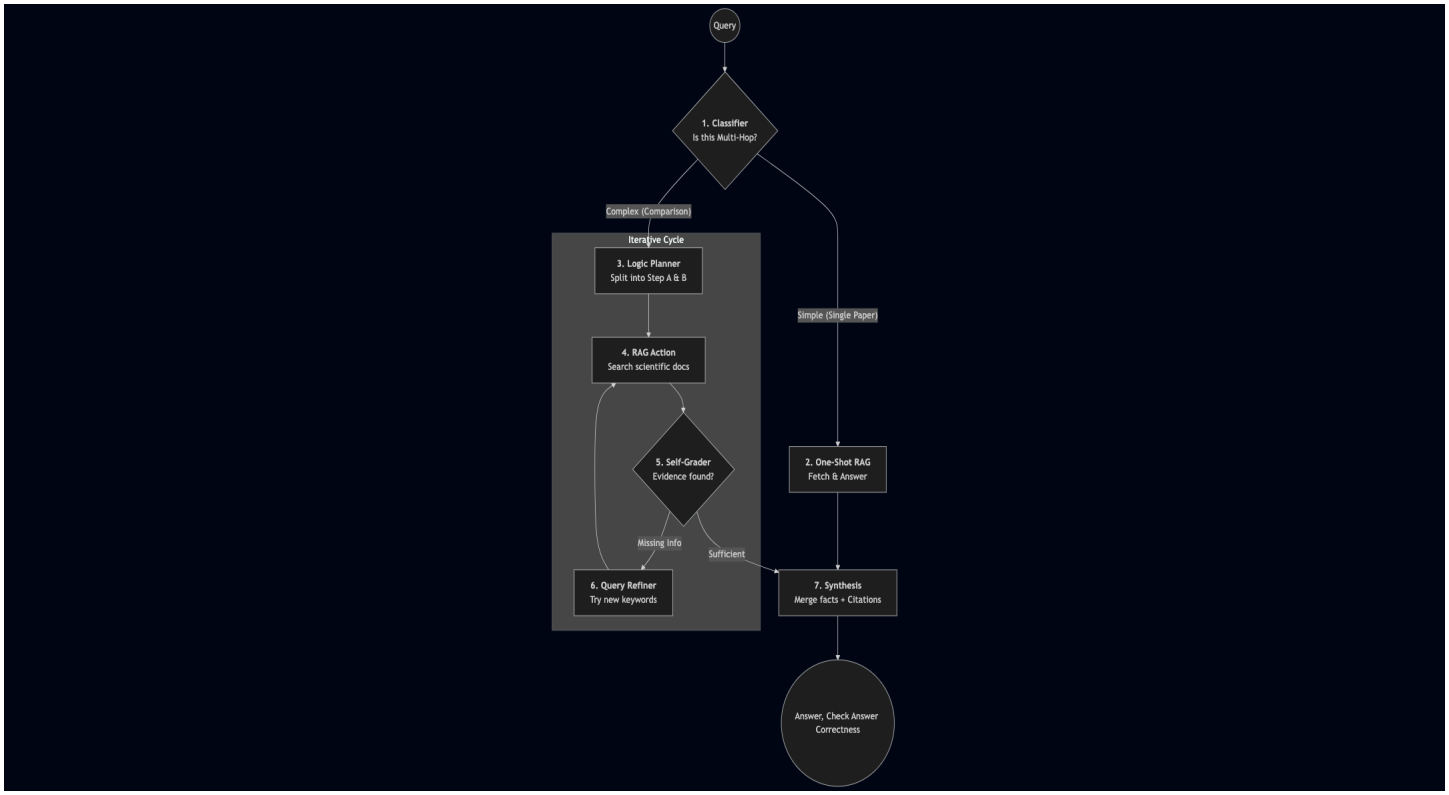
- **Scale:** 108 expert-curated question-answer pairs.
- **Components:** Each entry includes a **Question**, a **Ground Truth Answer**, a **Source_IDX (**The paper index**),** and a **Gold_REF** (the exact paragraph in the scientific paper containing the answer).
- **Depth:** Questions are categorized by complexity levels (1–4), testing both simple fact retrieval and complex multi-paper synthesis.

2. Evaluation Architecture

2.1.1 RAG Engineering Loop



2.1.2 Agent Engineering Loop



2.2 Configuration Comparison Table

This table establishes our benchmarks. The **Ceiling** is GPT-4o with direct access to the correct information (no search errors).

Configuration	Retrieval & Reasoning Strategy	**Model**	**Purpose**
Baseline	Vector Search only (k=5), Fixed-size chunks(512).	Llama-3.1 8B	Establish starting point.
Phase 1: Opti-RAG	Hybrid Search (Vector + BM25) + Reranking + Semantic Chunking.	Llama-3.1 8B	Minimize search errors.
Phase 2: Agent	ReAct loop with autonomous query decomposition and self-correction.	Llama-3.1 8B	Solve multi-paper reasoning.
The Ceiling	**Direct Gold_REF Injection**	**GPT-4o**	The theoretical maximum.

3. Metrics & Targets

Metric	**Measurement Tool**	**Target**	**Definition**
Context Recall	**RAGAS**	**> 0.92**	Ratio of ground-truth claims found within retrieved contexts.
Faithfulness	**RAGAS**	**> 0.98**	Percentage of claims in the answer supported by retrieved context.
Answer Correctness	**RAGAS**	**> 85% of Oracle**	Weighted factual overlap and semantic similarity to the Oracle.
Trajectory Success	**LLM-as-a-Judge**	**> 0.80**	Assessment of logical efficiency and valid tool use in the reasoning path.

4. Diagnostic Framework

Metric	The Problem	The Optimization
Context Recall	Lexical or semantic mismatch between query and index.	Hybrid Search (Vector + BM25) + Semantic Chunking + Cross-Encoder Reranker for high-precision ranking.
Faithfulness	The model is hallucinating.	Chain-of-Verification (CoVe) prompt + Source-Enforcement (requiring the model to quote the specific chunk ID first).
Answer Correctness	Agent misses the link between disparate facts.	Few-Shot Chain-of-Thought (CoT) + **Query Decomposition** to bridge cross-paper context.
Trajectory Success	Loop cycles, tool misuse, or loss of state.	Stateful Memory Management + Self-Correction Grader that forces a "Plan Pivot" if a search returns 0 results.

Final Verification Goal

The system is considered **successful** when the **Phase 2 Answer Correctness** is within **15%** of the GPT-4o Ceiling, while maintaining **> 0.98 Faithfulness**. This proves the 8B model is a reliable researcher.

External Dependencies And Tech Stack

LLM Runtime: Ollama

AI Model: Llama-3.1 8B (4-bit)

Embeddings: HuggingFaceBGE (BGE-Small)

Vector DB: ChromaDB

DB: SQLite

Desktop App: Electron + React

Backend: Python/Flask

Metrics Engine: RAGAS

Orchestration: LlamaIndex (Core + **Workflows**)

Hybrid Search: Rank-BM25

Reranker: FlashRank (Lightweight/Local)