Maurizio Scibilia Software Engineer | Semantic Search & Retrieval-Augmented Generation (RAG) | NLP | Cost-Efficient LLM Pipelines

EcoSearch:
Smarter Question Answering for Modern Enterprises

Efficient, transparent, and cost-effective QA—retrieval-first, LLM only when needed.

The Problem

Companies own vast internal knowledge—manuals, policies, archives—but struggle to make it accessible at scale. LLM-based QA systems are:

- Expensive (API calls for every question)
- Slow (cloud roundtrips)
- Not always private or compliant

Our Solution: EcoSearch

EcoSearch is a next-generation retrieval-augmented system.

- **LLMs are used only once, during initial corpus preparation,** to create concise summaries and guiding questions for each document chunk.
- **At query time, user data never leaves your environment unless the user requests a fallback LLM answer.**
- Most questions are answered instantly and locally, without additional LLM calls.

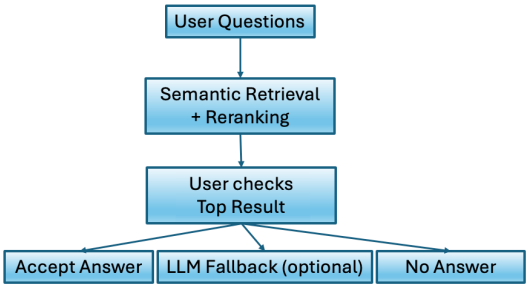
Key Advantages

- **Cost savings** on LLM usage (calls only for preprocessing and rare fallback)
- **Low latency** (answers in under 2 seconds for most queries)
- **User control:** You see exactly which data is retrieved and can choose if/when to use LLM fallback
- **Flexible:** Can run local or cloud, supports any LLM provider

How It Works

1. **One-time setup:**
   - Your documents are split, summarised, and guiding questions are generated with an LLM.
2. **Query time:**
   - User submits a question
   - System retrieves and reranks the best context from your data
   - User decides to accept the answer or trigger an LLM fallback

Visual Flow



Future Work

- **Agentic RAG** with autonomous query planning and multi-step reasoning.
- **User-in-the-loop prompt refinement** to iteratively improve answer quality.
- **Local & private LLMs** (Ollama / LM Studio) for on-premise retrieval and QA.

Contact for Demo or More Info mail linkedin