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 Generation (RAG) framework designed for transparency, control, and efficiency.

- It performs **question-to-question retrieval**, **CrossEncoder reranking**, and **answer generation with minimal LLM use**.
- During corpus preparation, LLMs generate concise summaries and guiding questions for each chunk.
- At query time, user questions are embedded and matched against those guiding questions for fast, interpretable retrieval.
- EcoSearch also introduces **oracle-spevaluation**, measuring semantic and structural overlap between retrieved and ideal oracle sentences — making retrieval performance quantifiable and transparent.

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
- **Transparent** — retrieval quality evaluated via **oracle-span overlap.**

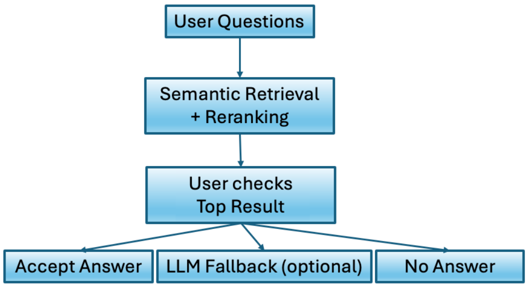
How It Works

1. **Preparation:** Documents are split into overlapping chunks, summarised and paired with a guiding question.
2. **Indexing:** Summaries and guiding questions are embedded (SentenceTransformers) and stored in FAISS.

Contact for Demo or More Info mail linkedin

3. **Retrieval:** User questions are embedded and compared to guiding questions for semantic matching.
4. **Reranking:** A CrossEncoder refines top candidates for precision.
5. **Decision:** User can accept the retrieved chunk, trigger a focused LLM fallback, or declare “no answer.”

Visual Flow



Platform Extensions

- A **web-based prototype** (Python, FastAPI, Streamlit) provides interactive retrieval and evaluation.
- A **mobile companion** (React Native, Expo) is being finalised, bringing OCR capture, multi-page PDF assembly, and local query capabilities — extending EcoSearch into real-world, on-device contexts.

Future Work

- **Agentic RAG** with autonomous query planning and multi-step reasoning.
- **User-in-the-loop prompt refinement** for adaptive guiding-question generation.
- **Local & private LLMs** (Ollama / LM Studio) for on-premises retrieval and QA.
- **Continued optimisation of chunking** and sentence alignment algorithms.