**Choosing and Setting Up LLM Backends for RAG Applications**

This project integrates multiple Large Language Models (LLMs) for Retrieval-Augmented Generation (RAG). Different backends can be used depending on hardware capacity, cost, and performance requirements. Below is a breakdown of the most common options and how to set them up.

**🔹 1. Ollama (Local LLMs)**

Ollama allows you to run models **completely offline** on your own machine. This is great for avoiding API costs or rate limits.

**✅ Pros**

* 100% offline, no API keys required.
* No cost per request.
* Data never leaves your machine.

**⚠️ Cons**

* Limited by your local CPU/GPU and available RAM.
* Large models (7B–13B parameters) may not run on laptops.

**📥 Setup**

1. Install Ollama.
2. Pull a model:
3. ollama pull llama3:3b # Lightest, runs on most machines
4. ollama pull mistral:7b # Better quality, needs ~8GB RAM
5. ollama pull llama3:8b # High quality, needs ~16GB RAM
6. In the app’s dropdown, select the pulled model.

**🔹 2. OpenAI GPT (Cloud)**

OpenAI models (like GPT-4o mini, GPT-4, GPT-3.5) run on OpenAI’s servers. They are **powerful and reliable**, but require an API key and billing account.

**✅ Pros**

* State-of-the-art reasoning and accuracy.
* No local hardware limitations.
* Simple integration with LangChain.

**⚠️ Cons**

* Requires API key + billing (free tier is limited).
* API quotas and rate limits may apply.
* Data is sent to OpenAI’s servers.

**📥 Setup**

1. Create an account at [platform.openai.com](https://platform.openai.com).
2. Generate an API key.
3. Save it as an environment variable:
4. setx OPENAI\_API\_KEY "sk-your-key-here"
5. Restart terminal and select GPT-4o mini in the app.

**🔹 3. Anthropic Claude (Cloud)**

Claude (Claude 3 Sonnet, Opus, Haiku) is another high-performance model hosted by Anthropic.

**✅ Pros**

* Excellent reasoning and summarization.
* Often more efficient than GPT in long-document RAG tasks.
* Easy LangChain integration.

**⚠️ Cons**

* Requires Anthropic API key + billing.
* API quotas.
* Data leaves your machine.

**📥 Setup**

1. Sign up at console.anthropic.com.
2. Create an API key.
3. Save it as:
4. setx ANTHROPIC\_API\_KEY "sk-ant-your-key-here"
5. Restart terminal and select Claude in the app.

**🔹 4. Hugging Face Hub (Cloud)**

The Hugging Face Hub provides thousands of open-source models (Falcon, Mistral, etc.). You can run them via API or download them locally.

**✅ Pros**

* Wide variety of models (open-source).
* Free tiers available with Hugging Face tokens.
* Easy to experiment with new models.

**⚠️ Cons**

* Free tier has rate limits.
* Performance depends on chosen model.
* Hosted models may be slower than OpenAI/Anthropic.

**📥 Setup**

1. Create a free account at huggingface.co.
2. Get your **Access Token**.
3. Save it as:
4. setx HF\_TOKEN "hf-your-token-here"
5. Use langchain-huggingface in your project to call Hugging Face models.

**⚖️ Recommendation**

* **For laptops with limited RAM:** Start with llama3:3b (Ollama).
* **For maximum accuracy:** Use OpenAI GPT-4o mini (requires billing).
* **For strong summarization:** Try Anthropic Claude.
* **For flexibility & open-source:** Hugging Face Hub models.

This flexibility means the project can run in **three modes**:

* 💻 **Local-first (Ollama)** → zero cost, offline.
* ☁️ **Cloud-first (OpenAI/Claude)** → best accuracy.
* 🔄 **Hybrid** → FAISS + embeddings locally, with a fallback to cloud LLMs.

flowchart TD

A[Start: User selects LLM] --> B{Where to run the model?}

B --> |Local (offline)| C[Ollama]

B --> |Cloud (online)| D[API-based LLMs]

C --> C1[Llama 3:3B → Lightest, runs on most laptops]

C --> C2[Llama 2:7B / Mistral 7B → More accurate, needs ~8GB RAM]

C --> C3[Llama 3:8B → High accuracy, needs ~16GB RAM]

D --> D1[OpenAI GPT-4o mini → Best reasoning, requires billing + OPENAI\_API\_KEY]

D --> D2[Anthropic Claude 3 → Great summarization, requires ANTHROPIC\_API\_KEY]

D --> D3[Hugging Face Hub → Wide model choice, requires HF\_TOKEN]

style A fill:#2c3e50,stroke:#fff,stroke-width:2px,color:#fff

style B fill:#34495e,stroke:#fff,stroke-width:2px,color:#fff

style C fill:#27ae60,stroke:#fff,stroke-width:1px,color:#fff

style D fill:#2980b9,stroke:#fff,stroke-width:1px,color:#fff

style C1 fill:#2ecc71,stroke:#fff,color:#000

style C2 fill:#2ecc71,stroke:#fff,color:#000

style C3 fill:#2ecc71,stroke:#fff,color:#000

style D1 fill:#3498db,stroke:#fff,color:#000

style D2 fill:#3498db,stroke:#fff,color:#000

style D3 fill:#3498db,stroke:#fff,color:#000