**AI and Agentic Systems**

**Hands-on Lab #3**

**RAG Chabot**

A cartoon of a robot pointing at vinyl records

AI-generated content may be incorrect.

Key topics: retrieval-augmented generation, batch loading of embeddings, recommender chatbot, rapid prototyping with Gradio

**Dependencies**

Ollama is installed locally, with a few LLM models pulled (pick a couple): llama3 (4.7 GB), mistral (4.4 GB), phi3 (2.2 GB), gemma3:270m (291 MB), qwen:0.5b (394 MB). Individual models can be pulled down with the following command:

ollama pull [model\_name]

pip install gradio langchain\_chroma

**Context**

In this hands-on session, you'll build a chatbot designed to help music lovers, DJs, and curious listeners discover new albums based on the vibe they're craving. Whether someone types in “chill beats for a rainy afternoon” or “high-energy tracks for a late-night set,” your chatbot will interpret the mood, match it to relevant albums, and deliver personalized recommendations. You’ll explore prompt engineering, semantic search, and context-aware retrieval—all while tuning into the creative potential of AI in music discovery. By the end, you’ll have a working prototype that feels more like a musical companion than a search engine. Let’s get in tune with the tech.

**Lab-specific activities**

Step 1: Embedding music documents in ChromaDB via chunking

Step 2: Building a CLI Music Recommender Chatbot

Step 3: Rapid prototyping with Gradio

Step 4: Initial evaluation using different models, prompts, temperatures, top-k in your retriever, and size of ChromaDB

**Post-lab follow-up**

By tomorrow night, complete the question(s) associated with this lab In LearningSuite.