AI Tutor Chatbot

Agentic Retrieval-Augmented Generation (RAG)

Agenda

- Project Overview & Key Objectives
- High-Level Architecture
- Core Features & Tools
- Performance & Evaluation
- Conclusion & Future Directions

Project Overview & Key Objectives

AI Tutor Chatbot

addresses classical RAG limits using **agentic capabilities**

Local CPU/GPU & Offline-Focused

avoids reliance on persistent online infrastructure

Core Components:

- **LLMs (e.g., Llama)** for language understanding & generation
- **Chroma + sentence-transformer embeddings** for semantic retrieval
- **LangChain & LangGraph** for prompt flow and tool management

vLLM

Llama 3

unsloth

LangChain

LangGraph

groq

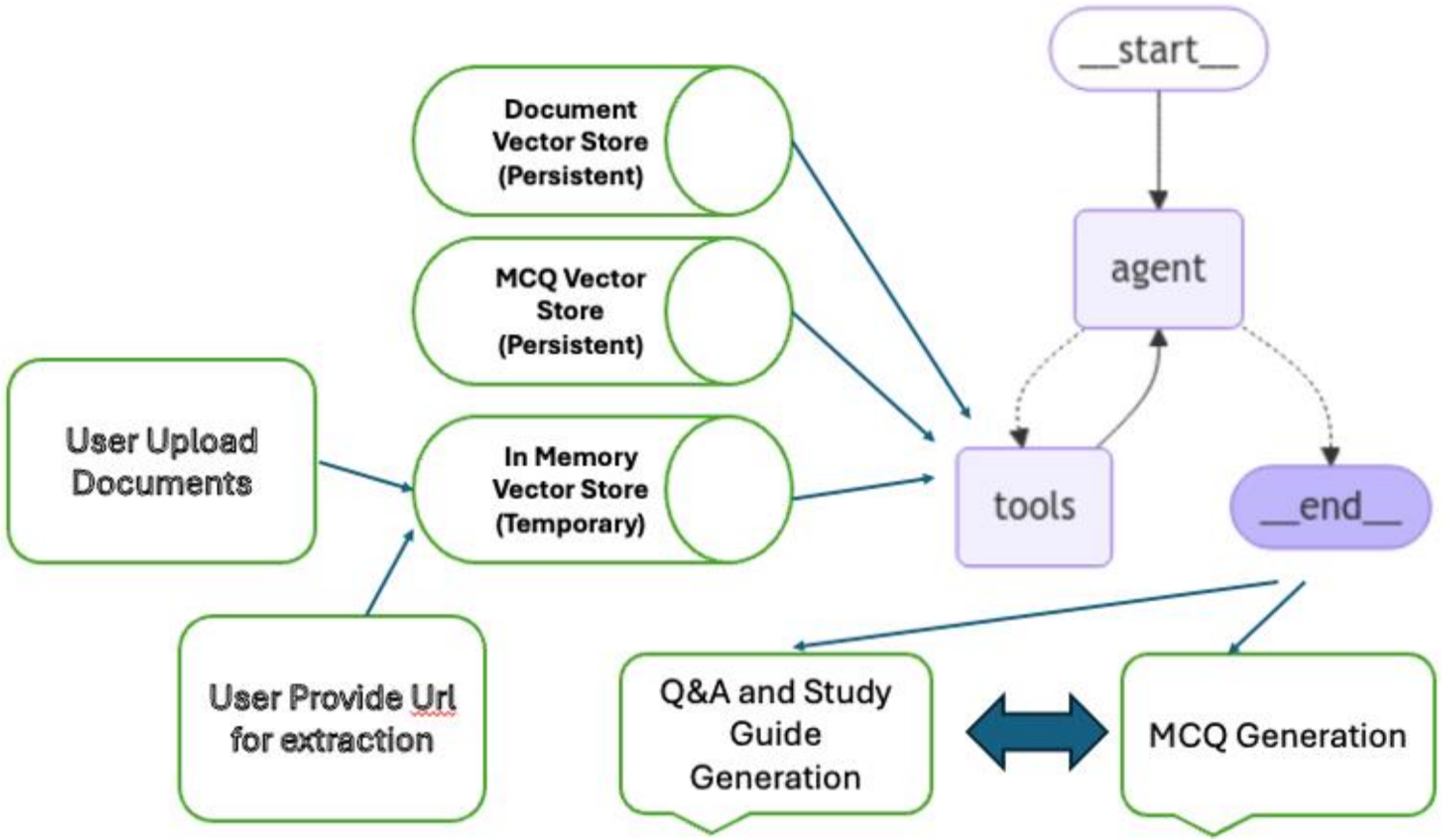
chroma

DeepSeek

Into the unknown

docker

High-Level Architecture

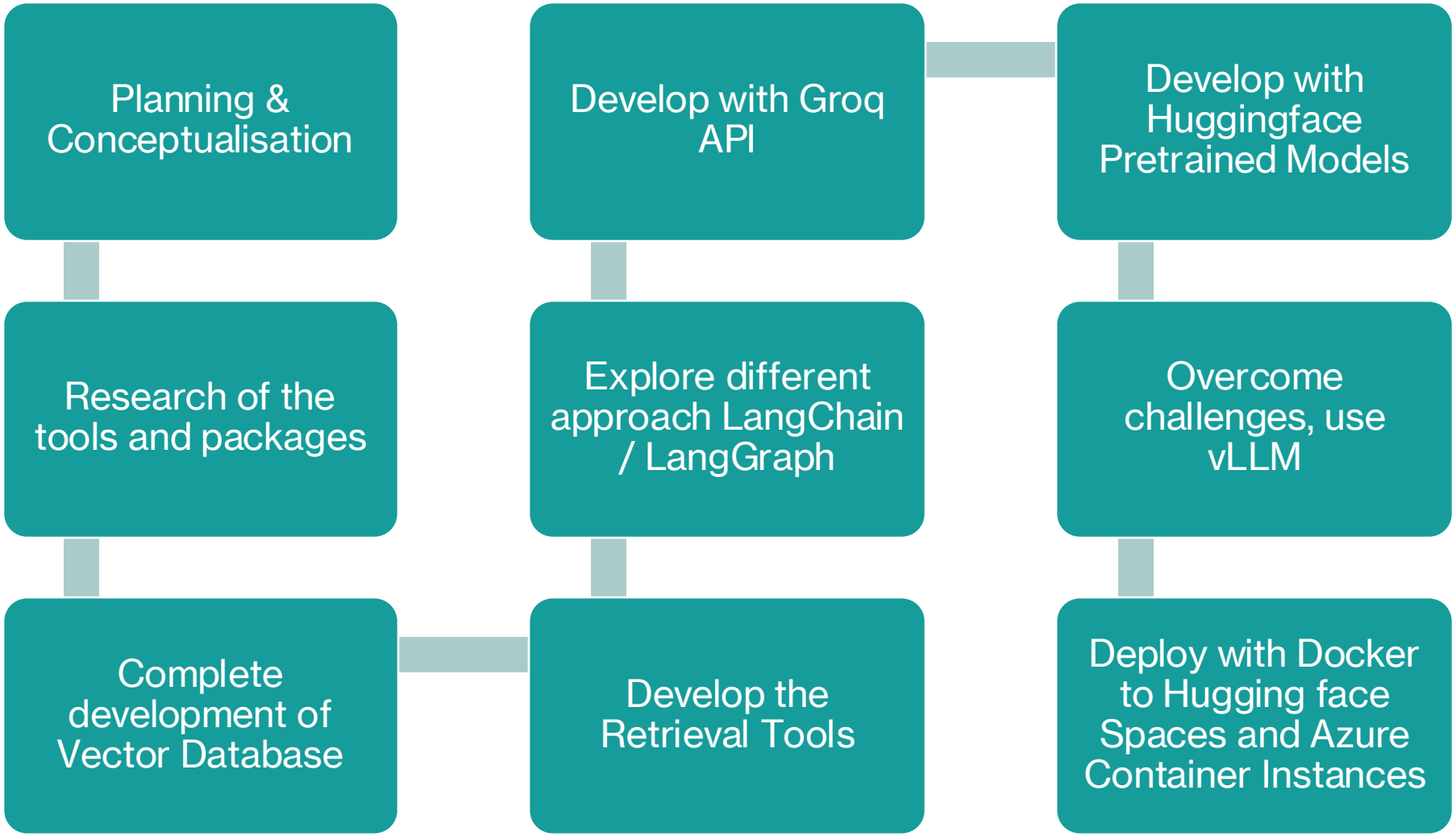


- **Nodes**
  - LLM Agent
  - Tool Invocations
  - Memory Summaries
- **Vector Stores (Chroma)**
  - general_vs
  - mcq_vs
  - in_memory_vs
- **LLM Integration**
  - vLLM
  - ChatGroq with OpenAI-compatible API
- **Gradio UI**

Core Features & Tools

JSON-based calls (StructuredTool classes)	Retrieval Methods	Session Memory & Summaries
• Utilizes structured JSON for tool interactions	• Similarity Search • Ensemble methods for improved accuracy • MCQ (Multiple Choice Questions) retrieval	• “Summary Tool” for finalizing context • Short-term memory (in-memory) • Long-term memory (persistent stores)

Development Phases



Performance & Evaluation

Testing Aspects

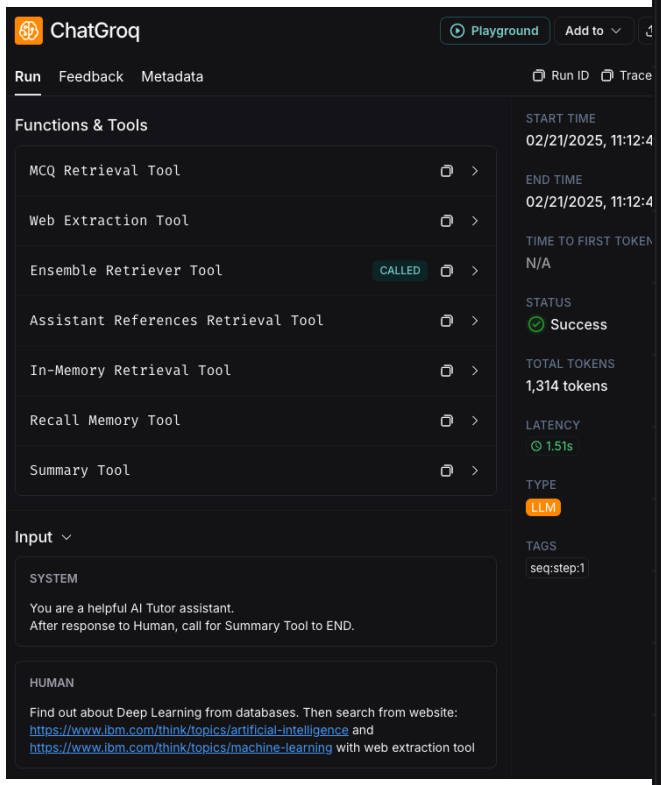
- Relevance, completeness, and correctness of answers

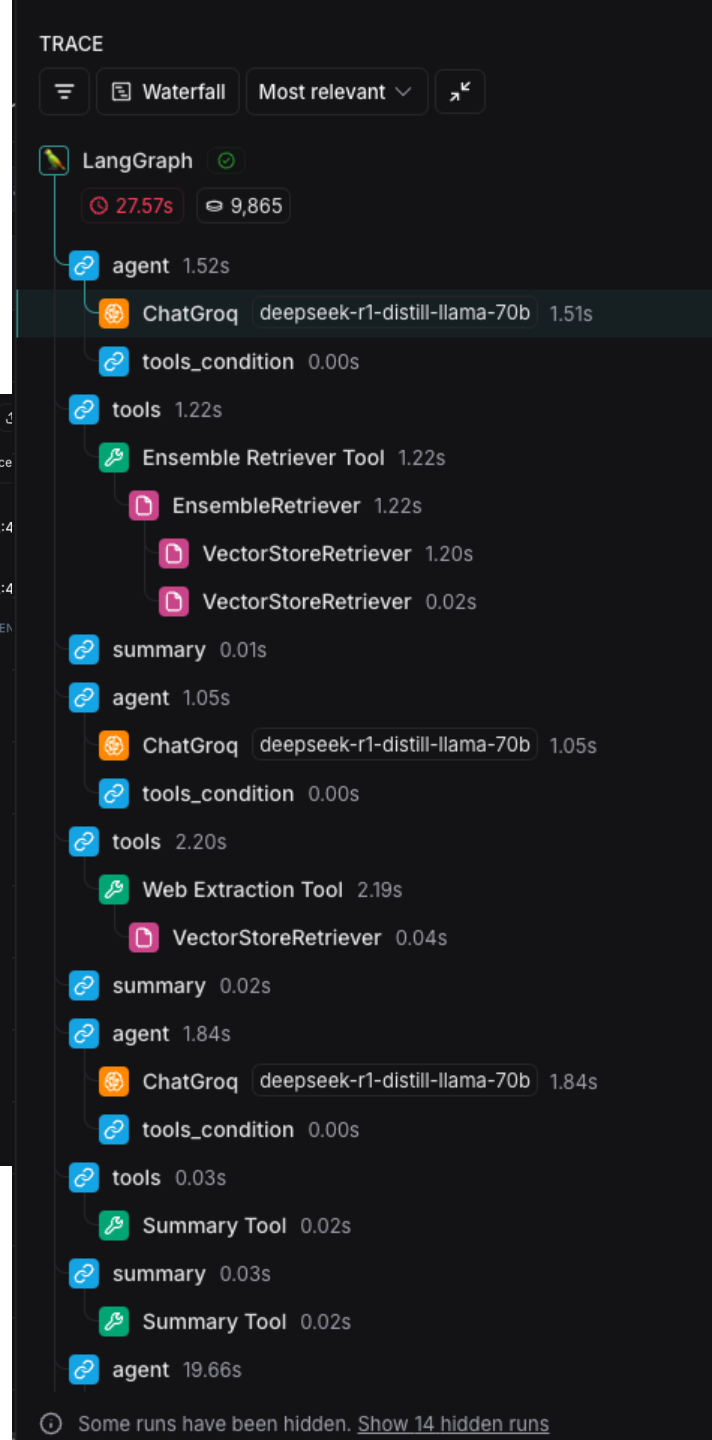
LangSmith Observability Features

- Logs each step and tool call
- Helps refine prompts and measure hallucinations

Key Challenges

- Smaller/quantized models struggle with correct JSON outputs
- Docker and resource constraints for deployment





Conclusion & Future Directions

Achievements

- Agentic RAG with dynamic memory and advanced tool usage
- Education-focused features including MCQ retrieval and quiz checks

Future Enhancements

- Domain-specific fine-tuning
- Quantization to reduce resource requirement