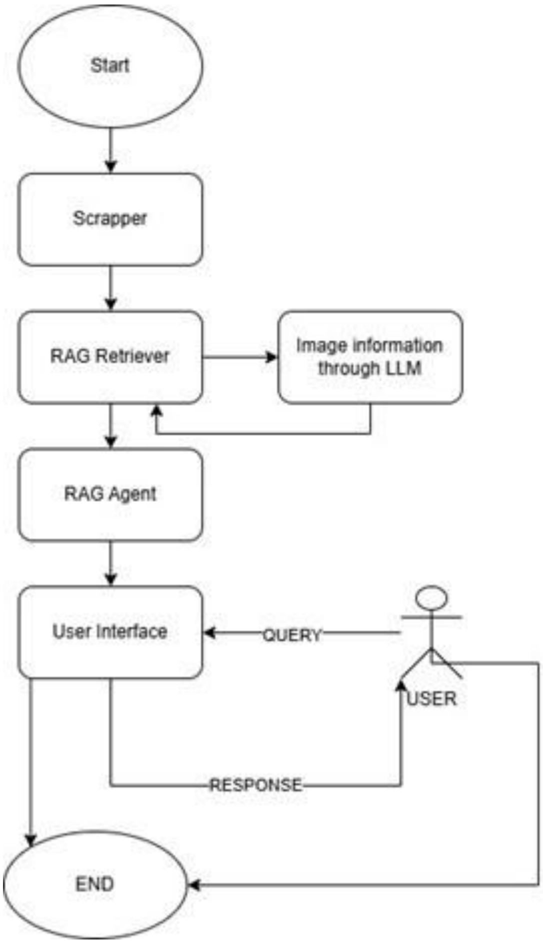
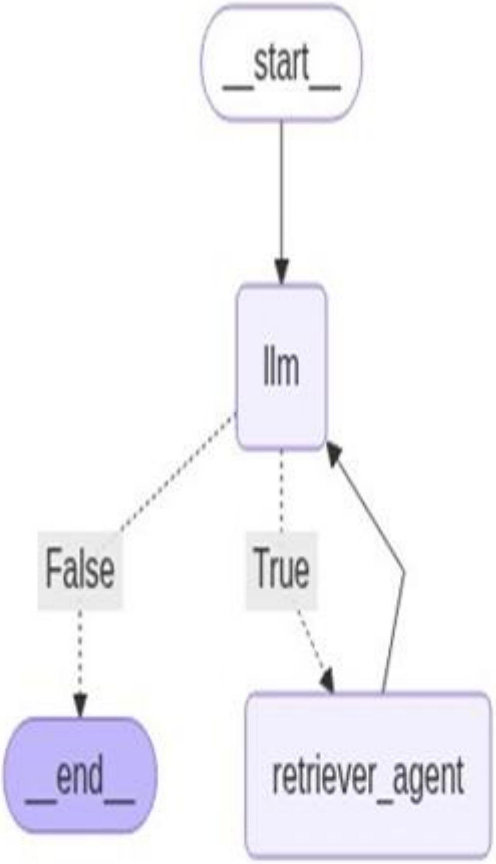
RAG Research Assistant for AI Literature Analysis

🏗️ System Architecture & Workflow





End-to-End Process Flow:

1. START → SCRAPER → Automated content extraction from research papers
2. RAG RETRIEVER → Document processing, chunking, and vector embedding
3. IMAGE INFORMATION → Multi-modal analysis via vision-language models
4. RAG AGENT → Intelligent query processing and response generation
5. USER INTERFACE → Interactive chat experience with streaming responses
6. END → Complete research analysis cycle

🤖 AI Models & Core Technologies

Primary AI Models

Component	Model/Service	Purpose
Language Model	Azure OpenAI GPT-4o	Primary reasoning, response generation, and complex query understanding
Embeddings	text-embedding-3-small	Document vectorization and semantic similarity matching
Vision Model	GPT-4o Vision	Image analysis and description generation for multi-modal content

Core Framework Stack

Layer	Technology	Function
Agent Framework	LangGraph	Multi-step reasoning, state management, and tool orchestration
LLM Integration	LangChain	Model abstraction, prompt engineering, and tool calling
Vector Database	ChromaDB	High-performance similarity search and document storage
Web Interface	Streamlit	Interactive chat interface with real-time streaming
Web Scraping	FireCrawl	Professional content extraction from research websites
Document Processing	Unstructured	Intelligent markdown parsing and text chunking

How would you scale this for 1M documents/images?

- **Batch Processing and async**:
  - To process various documents and queries by the multiple users.
- **Caching**:
  - Storing the context of most often used queries inorder to reduce the usage and cost of LLM services.