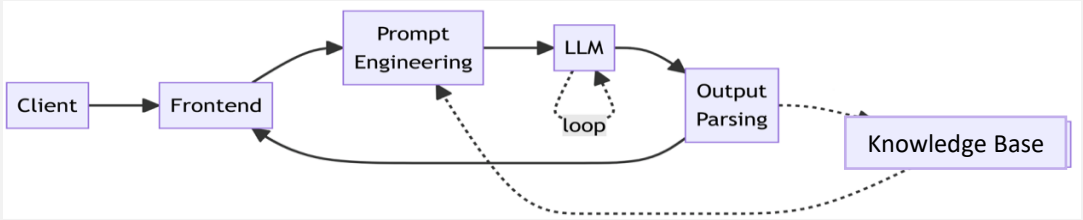
Carnegie Mellon University

Lecture 3:
LLM Application Design

SPRING 2025

MOHAMED FARAG

FARAG@CMU.EDU



Agenda

- Overcoming LLM Limitations
- LLM Application Design using FlowiseAI
- Design your First LLM Application

Reminder: LLM Limitations

LLMs are not perfect! Keep in mind the following limitations when using LLMs:

- **Knowledge Cutoff:** LLMs are limited to the training date.
- **Accuracy**: Generative AI doesn't work well with tabular data and mathematical calculations (Use supervised learning instead).
- **Inability to take action:** Large language models are not equipped to execute interactive tasks such as searches, calculations, or data retrievals, which considerably restricts their capabilities.
- **Bias and Toxicity**

LLM Limitations – Example

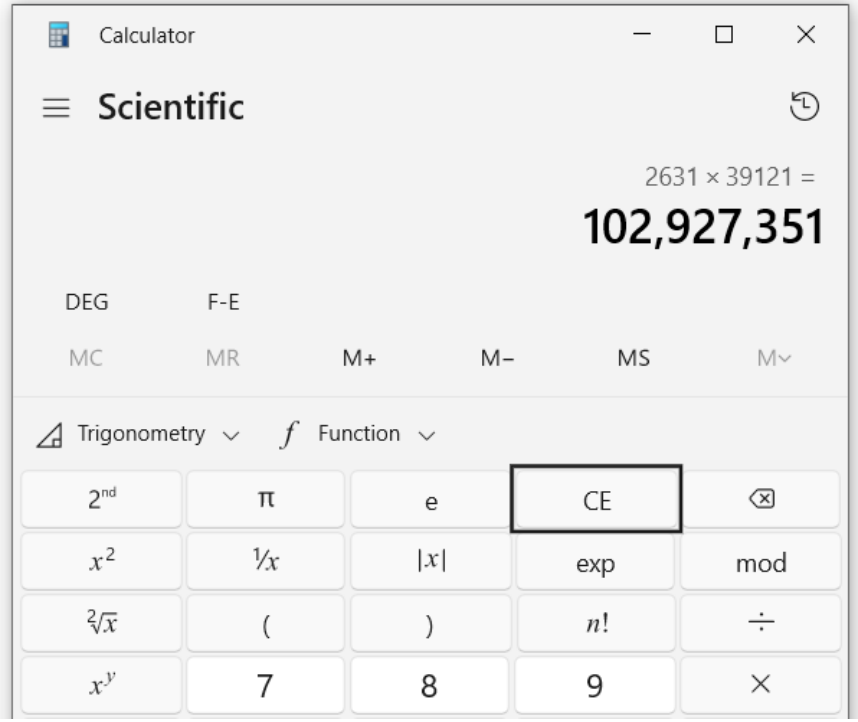
ChatGPT 3.5

You
What is the result of 26 multiplied by 25?

ChatGPT
26 multiplied by 25 equals 650.

You
What is the result of 2631 multiplied by 39121?

ChatGPT
The result of 2631 multiplied by 39121 is 103,034,151.



Overcoming LLM Limitations

1. **Retrieval augmentation:** This method supplements a language model's outdated training data with information from external knowledge bases, providing additional context and minimizing the chances of generating incorrect or imaginary content.
2. **Chaining:** This approach combines tasks such as searching and performing calculations within a workflow
3. **Prompt engineering:** This requires meticulously designing prompts to include essential context, guiding the language model towards more accurate and relevant responses.

Overcoming LLM Limitations

4. **Monitoring, filtering, and reviews:**

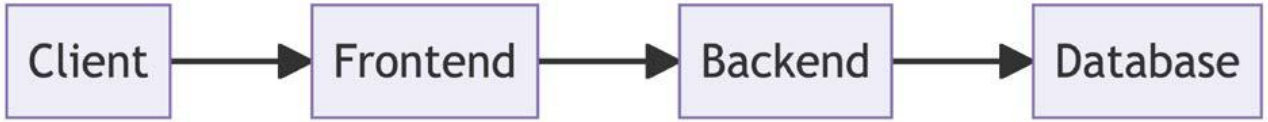
   - **Filters:** like block lists, sensitivity classifiers, and banned word filters.
   - **Constitutional principles:** monitor and filter unethical or inappropriate content
   - **Human Reviews:** to provide continuous feedback to the LLM

Building LLM Applications

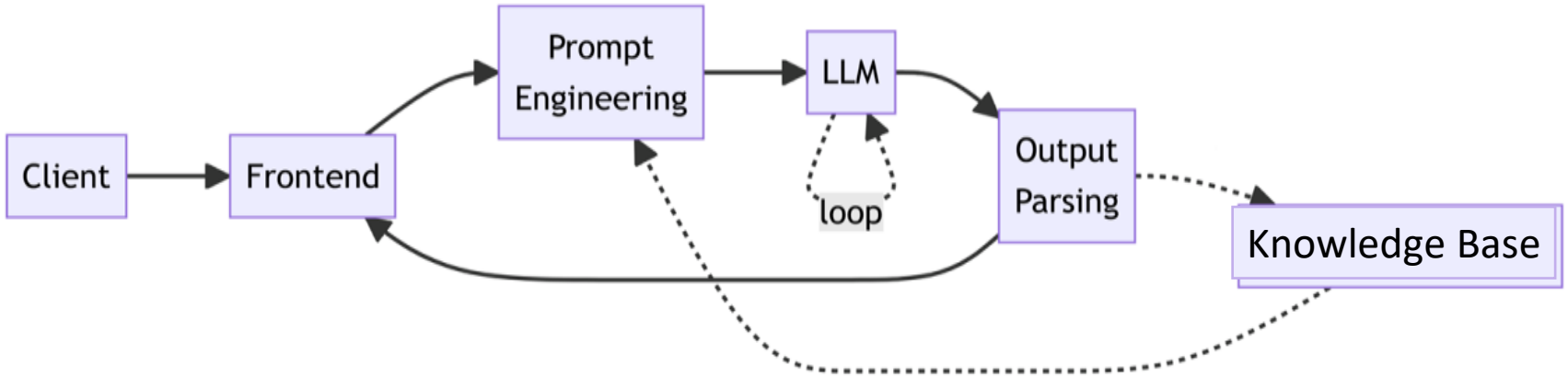
- To leverage LLM benefits and overcome LLM challenges, new kind of Apps is needed
- This new type is called **LLM Apps.** LLM Apps offer several benefits. For example:
    - Human-like language processing without rigid rules.
    - Personalized responses using previous interactions.
    - Advanced algorithms for complex, multi-step reasoning.
    - Dynamic responses using real-time information or LLM.

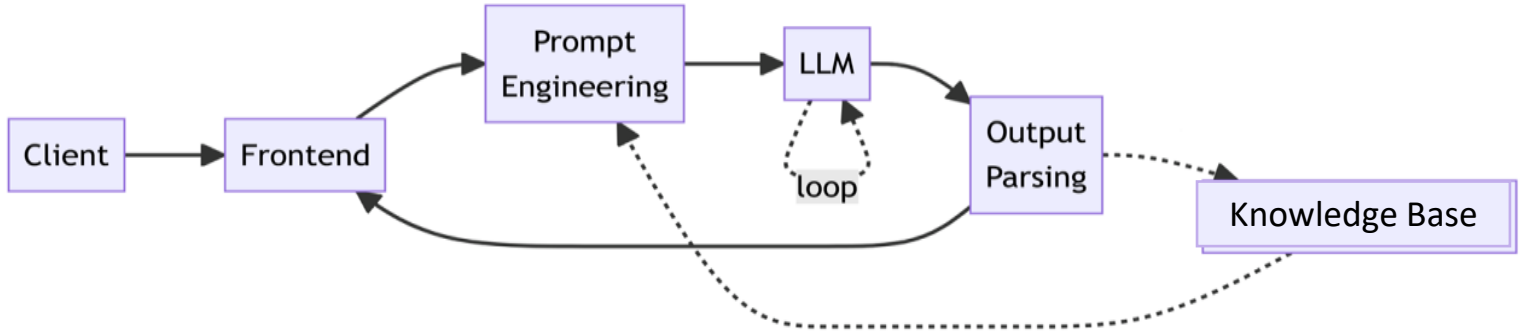
Designing LLM Applications – The General Idea

- **Traditional Apps:**



- **LLM Apps:**





- A **frontend layer** to collect user input as text queries or decisions.
- A **prompt engineering** layer to construct prompts that guide the LLM. Prompt engineering may include prompt parsing as well.
- An **LLM backend** to analyze prompts and produce relevant responses.
- An **output parsing** layer to interpret LLM responses for the application interface.
- Optional integration with **knowledge bases (internal or external)** via APIs, data stores, and reasoning algorithms to augment the LLM's capabilities.

LLM Application Examples

1. Chatbots and virtual assistants
2. Intelligent search engines
3. Automated content creation
4. Question answering
5. Sentiment analysis
6. Text summarization
7. Data analysis & Code Generation

LLM App Builders: FlowiseAI

- FlowiseAI is a low-code/no-code drag & drop tool with the goal to make it easy for people to visualize and build LLM apps.
- FlowiseAI is used to design and build Proof-of-Concept LLM Applications.
- Get started by checking their documentation: https://docs.flowiseai.com/getting-started

FlowiseAI Deployment Options

SaaS	Self-Hosted on the Cloud	Local Machine
~$35/user	Pay for cloud use	Free

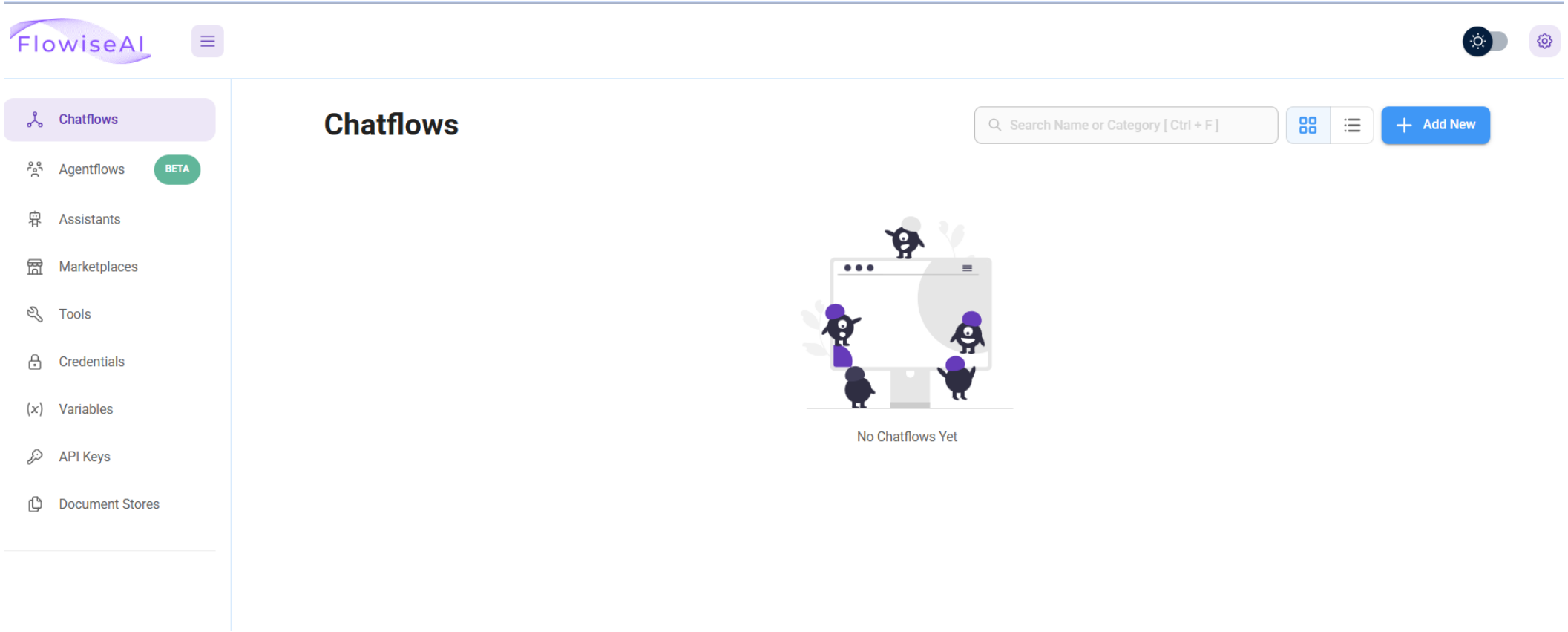
FlowiseAI – Local Installation – Option 1: Docker

1. Download Docker: https://www.docker.com/products/docker-desktop/
   Note: if you using Apple chip, enable Rosetta emulation from Docker Desktop settings
2. Run FlowiseAI container using the following commands:
   - docker pull flowiseai/flowise
   - docker run -d --name flowise -p 3000:3000 flowiseai/flowise
3. Navigate to http://localhost:3000/

FlowiseAI – Local Installation – Option 2: NodeJS

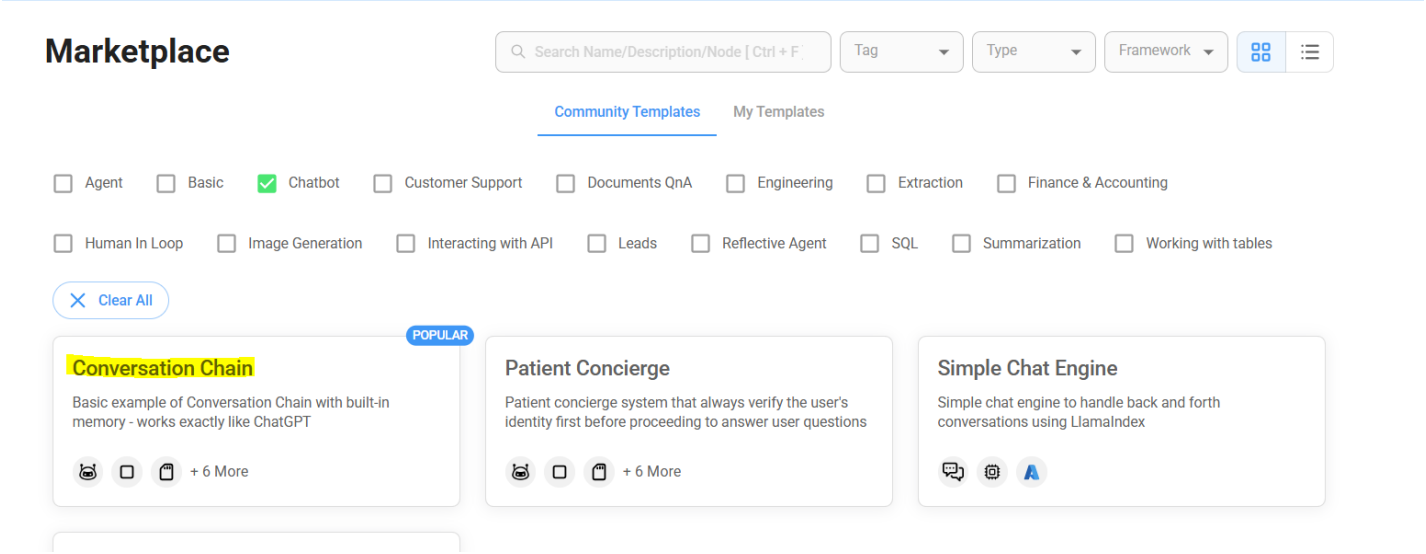
1. Download NodeJS: https://nodejs.org/en/download
2. Install and start FlowiseAI in your terminal
    - npm install -g flowise
    - npx flowise start
3. Navigate to http://localhost:3000/

FlowiseAI – Home Page



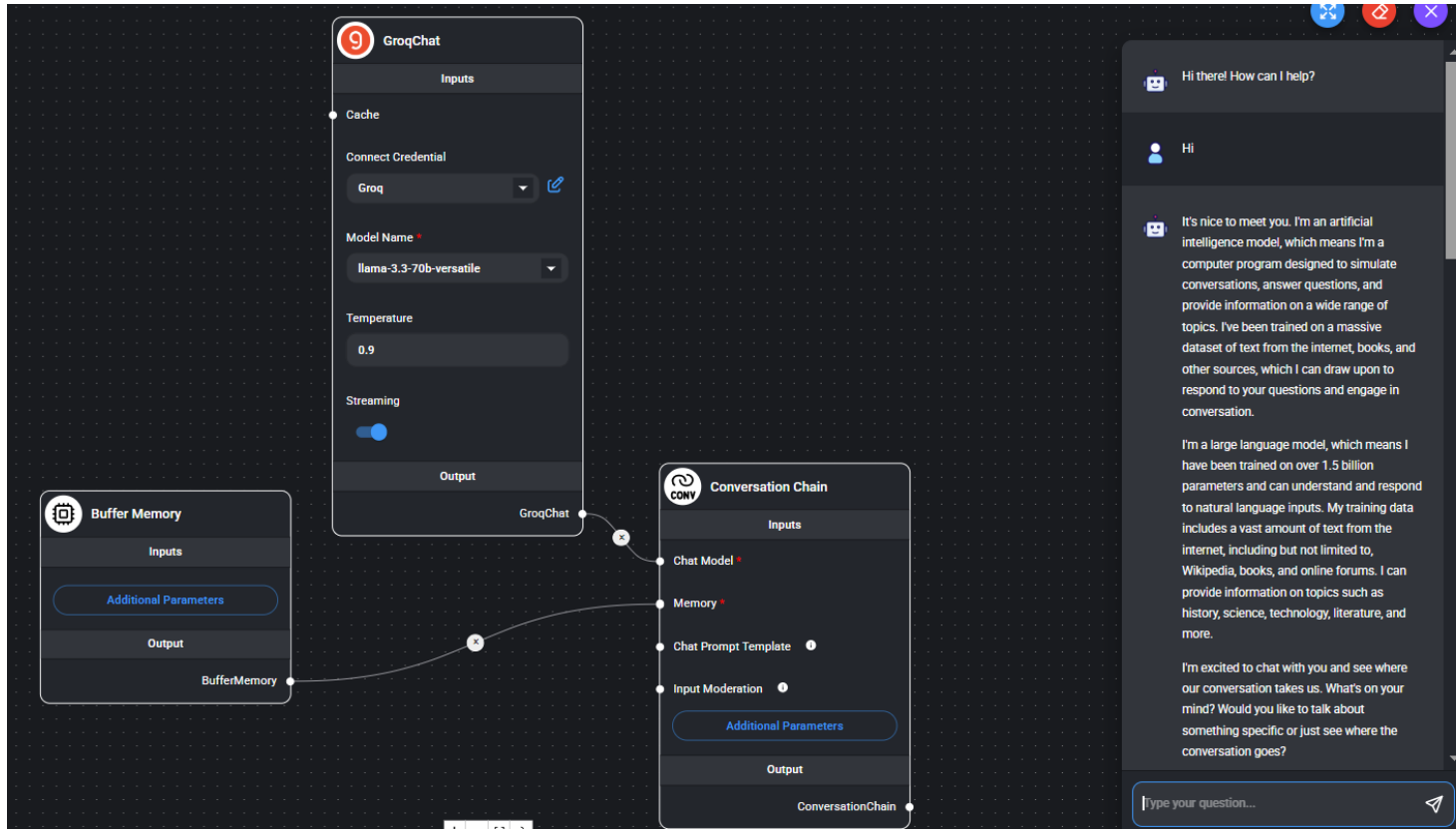
Design Your First LLM Application

1. Create New Credentials for your LLM Application.
    - You may use OpenAI or Groq (will use Vertex AI next lec.)
    - Create New Conversation Chain from the Marketplace



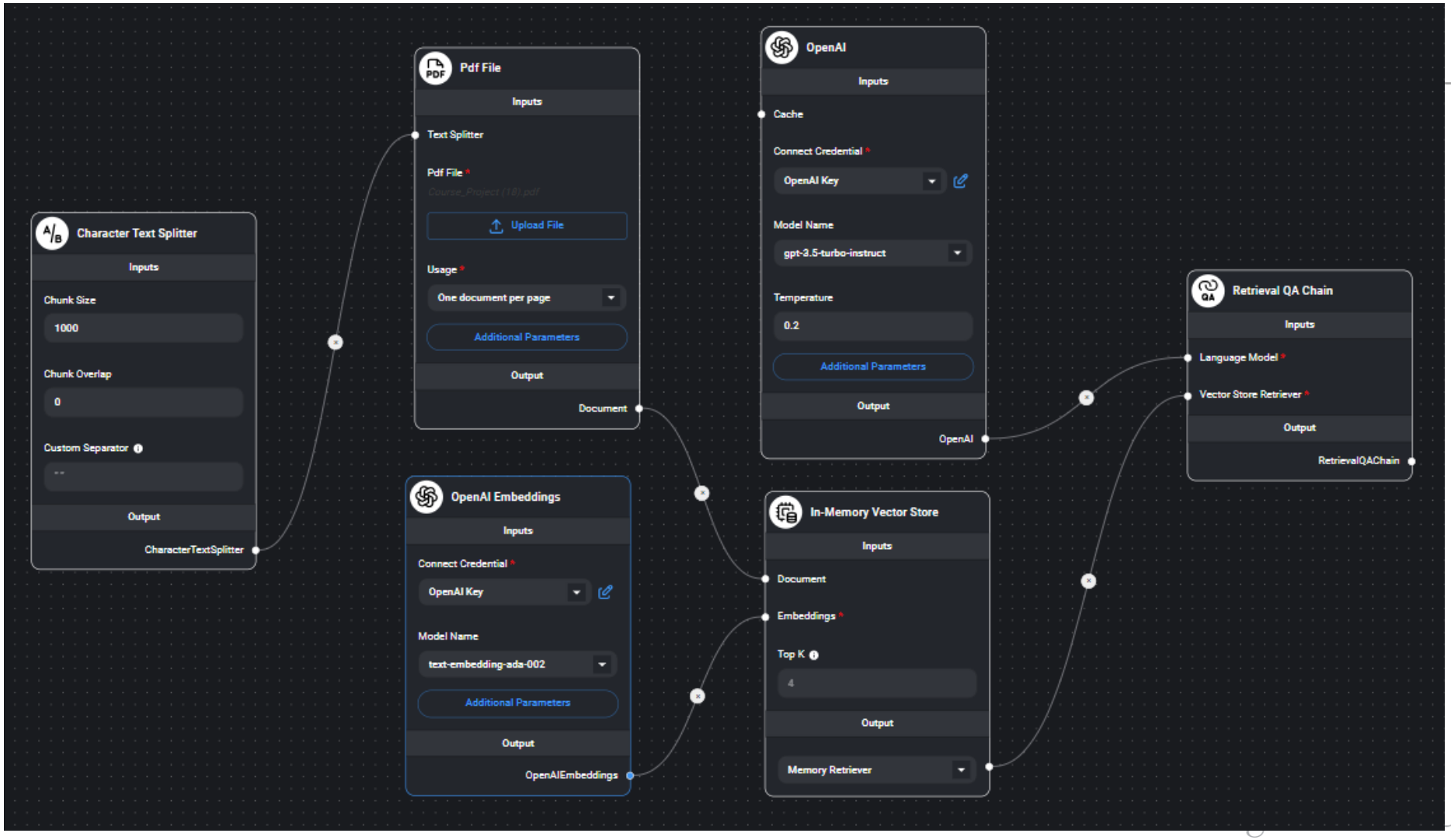
Design Your First LLM Application – Cont'd

1. Design your LLM app as shown in the image.
2. Save your design.
3. Open the conversation icon and chat with your model.



Try at Home: LLM App on FlowiseAI: Chat with Your PDFs

Note: Get OpenAI/Groq Key and Add Balance To Your Account



Asking Questions on Our Course Project PDF