aiDAPTIV Langchain User Guide

Overview

This guide explains how to set up, run, and use the Langchain-based aiDAPTIV integration. You will learn how to build the KV cache, upload documents, chat with the AI assistant, and perform accelerated inference.

Chapter 1: Installation and Setting

Dependencies

The python dependencies are listed as follows:

```
 langchain
langchain-community
langchain-openai
pypdf
```

Pre-requisites

1. Make sure that the aiDAPTIV server is hosted on `http://localhost:8000/v1` so that the application can work properly.

Installation Steps

Run (or double click) the `Langchain_Installer_1.0.0.exe` script to install and run the application (It may take some time to parse PDFs if there's any uploaded).

Chapter 2: How to Use?

Usage Workflow

1. **Initial Setup**

- Run (or double-click) the installer: `Langchain_Installer_1.0.0.exe`
- The installer will start the application automatically. If PDF files are already present, they may take some time to be parsed.

2. **Basic Operation**

- **Step-by-step Instructions**
    - Choose `1. Build KV Cache (for accelerated inference)` by typing `1` to enhance the inference speed when conducting Q&A with your documents.
    - Choose `2. Chat with AI Assistant` by typing `2` when you're ready to chat with your assistant.
    - After selecting Option 2, you will be navigated to select which document to be used as context, select 1 document, then you can start chatting to get insights from the document selected.
- **Using Example Files (Golden Workflow)**
    - Start by building KV Cache to enhance the inference speed. (This may take a few minutes to process large documents)

```
 Before you asks question, you may upload your PDF file
Currently accessed files:
1. Biology-2-2.pdf
2. Biology-2-3.pdf
3. Biology-2-4.pdf
4. Biology-2-1.pdf

Actions available:
1. Build KV Cache (for accelerated inference)
2. Chat with AI Assistant
3. Quit
Please select your action: 1
Building KV Cache for 0 tokens
Done processing KV Cache for text 1. Time taken: X secc
```

```
Building KV Cache for 0 tokens
Finished KV Cache building for all documents, navigatir
```

- You may now start chatting with your assistant with accelerated speed. First, select the document you want to know more about.

```
 Currently accessed files:
1. Biology-2-1.pdf
2. Biology-2-2.pdf

Please select a file to act as reference for your Q&A:
========================================================
You have entered the chat room, enter "Q" to quit this
========================================================
Assistant:
Hello, how can I help you?

User:
```

- Afterwards, you should start seeing the response being streamed after you type your question.

```
 Currently accessed files:
1. Biology-2-1.pdf
2. Biology-2-2.pdf

Please select a file to act as reference for your Q&A:
========================================================
You have entered the chat room, enter "Q" to quit this
========================================================
Assistant:
Hello, how can I help you?

User: What is this document about?
Assistant:
```

```
This document is about cellular respiration, specifical
========================================================
```

- You can ask more questions about your documents. Type `q` or `Q` in the chat to exit the application. If that doesn't work you may try pressing `CTRL + C` to terminate the application.

Chapter 3: Troubleshooting

Issue 1: Application Exits with Server Connection Error

- **Symptoms**:
  - The application prints the message: `Please ensure that the aiDAPTIV server is up and available on <OPENAI_BASE_URL>!`
  - The program waits for 15 seconds, then closes automatically.
  - Chat or KV cache features do not start.
  - **Cause**: This occurs when the application cannot reach the aiDAPTIV server.
- **Solution**:
  - Ensure that the aiDAPTIV server is running and reachable at the configured OPENAI_BASE_URL.
  - Confirm the server is listening on the correct port (e.g., `http://localhost:13141/v1`)
  - Restart the aiDAPTIV server, then relaunch the application.
- **Verification Steps**:
  - Restart the Langchain application.
  - The error message should no longer appear.
  - You should be able to build KV cache or enter chat mode normally.