Abschlussprojekt: Entwicklung eines RAG-Systems

Überblick

In dieser Aufgabe entwickeln Sie ein **Retrieval-Augmented Generation (RAG)**-System, das externe Dokumente verwendet, um Benutzerfragen zu beantworten. Ziel ist es zu zeigen, wie große Sprachmodelle (LLMs) mit aktuellem, domänenspezifischem Wissen verknüpft werden können.

Das verwendete Modell hat einen Wissensstand mit **Stichtag August 2024**. Ihr System muss daher Informationen aus **externer Dokumentenrecherche** verwenden, um Fragen zu **Ereignissen nach diesem Datum** korrekt zu beantworten.

Projektanforderungen

Dokumentenauswahl und -verarbeitung

Wählen Sie ein Dokument über ein Ereignis nach August 2024. Speichern und indexieren Sie es mit **ChromaDB** und aktivierter Persistenz. Verwenden Sie eine geeignete Textaufteilung, um mindestens **50 sinnvolle Chunks** zu erzeugen.

Systemarchitektur

Nutzen Sie das Modell `gemini-2.0-flash`. Die Pipeline ist mit **LangChain** oder **LlamaIndex** umzusetzen. Zur Nachvollziehbarkeit und Beobachtbarkeit verwenden Sie **LangSmith** oder **LangFuse**.

Der Einsatz von **vorgefertigten Agenten ist nicht erlaubt**. Die zentralen Komponenten müssen selbst entwickelt werden.

Das System muss folgende Funktionen unterstützen:

- **Dialogführung** – mehrstufige Benutzerdialoge.
- **Kontextverwaltung** – Frühere Benutzereingaben sollen gespeichert und so verwaltet werden, dass sie auch in späteren Sitzungen oder an anderen Tagen erneut abgerufen und für die Beantwortung neuer Fragen genutzt werden können.

Wirksamkeitstest

Erstellen Sie mindestens **fünf Fragen**, die nur mit Hilfe des Dokuments korrekt beantwortet werden können. Das Modell soll diese Fragen **nicht ohne Dokumentretrieval** beantworten können.

Testen Sie Ihr System mit und ohne Retrieval und dokumentieren Sie die Unterschiede. Variieren Sie außerdem die **System-Prompts** und analysieren Sie deren Auswirkungen.

Codequalität

Ihr Repository muss gute Entwicklungsstandards einhalten:

- Keine großen Dateien im Git-Verlauf.
- Keine geheimen Token in den Commits.
- Gut strukturierter und dokumentierter Code.

Abgabehinweise

Abgabefrist: 11.05 um 23:59 Uhr

Jede:r Studierende hat einen eigenen Branch im Repository. Öffnen Sie einen **Pull Request (PR)** von Ihrem Arbeits-Branch auf den Ihnen zugewiesenen Branch.

Das Repository befindet sich unter folgendem Link: `github.com/hussamalafandi/GenerativeAI-II-Project`

Ihr PR muss Folgendes enthalten:

- Die vollständige Implementierung Ihres RAG-Systems.
- Ein Jupyter-Notebook oder Skript mit Demonstration der:
  - Dokumentenindexierung
  - Dokumentenabfrage (Retrieval)
  - Fragebeantwortung
  - Prompt-Variationen
- Einen Link zu Ihrem Projekt in **LangSmith** oder **LangFuse**.

Bonus (Zusatzpunkte)

Für Zusatzpunkte muss Ihr System beide der folgenden Funktionen enthalten:

1. **Metadatenfilterung** – Eingrenzung der Dokumentensuche nach Attributen.
2. **Multi-Query-Retrieval** – Nutzung mehrerer oder umformulierter Anfragen zur Verbesserung der Antwortqualität.

Abschließender Hinweis

Ihr System soll nachweisen, dass es nur mit der Hilfe von Dokumentretrieval zu korrekten Antworten kommt. Der Mehrwert Ihres RAG-Systems zeigt sich darin, dass das Sprachmodell allein nicht ausreicht.