Retrieval-Augmented Generation (RAG) und Agentensysteme

Assignment
DHBW Stuttgart
Modul: Anwendungsaspekte des Machine Learning
Studiengruppe: WWI2022A/B
Abgabe: 31. August 2025, 23:59 Uhr
Einzelarbeit, Umfang ca. 5–7 Seiten + Code
Abgabeformat: frei wählbar, z.B. PDF-Dokument + lauffähiger Code (GitHub-Link oder ZIP), oder alles als ein Dokument
Selbstständigkeitserklärung: Bitte fügen Sie am Ende des Dokuments eine unterschriebene Erklärung hinzu, dass Sie die Arbeit selbstständig verfasst haben.
Sofern wissenschaftliche Quellen verwendet werden, diese bitte zitieren, da es ansonsten als Plagiat gilt!

Zielsetzung

Dieses Assignment vertieft zwei zentrale Konzepte moderner KI-Systeme:

1. **Ein Retrieval-Augmented Generation (RAG)-System** mit einer selbstgewählten Wissensquelle
2. **Ein Agentensystem** mit erweiterten Fähigkeiten wie Tool-Nutzung, Routing und Orchestrierung

Sie verbinden eigenständige Recherche, theoretische Einordnung, praktische Implementierung und optional eine kritische Reflexion.
Im Fokus steht, dass Sie nicht nur funktionierende Prototypen entwickeln, sondern auch die dahinterliegenden Konzepte verstehen. Besonders bei Agentensystemen ist eine vorgelagerte Recherchephase vorgesehen.

Aufgabenübersicht

Sie bearbeiten **zwei verpflichtende Teile (dritter Teil optional!)**:

1. **RAG & Agentensysteme: Theoretische Recherche & Einordnung**
2. **RAG & Agentensysteme: Implementierung**
3. **(optional) Transferanalyse und Reflexion**

Teil 1: Theoretische Recherche & Einordnung

1.1 RAG-Systeme (ca. 2 Seiten)

Recherchieren und beschreiben Sie:

- Die Architektur eines typischen RAG-Systems
- Unterschiede zu reinen generativen Modellen und klassischen QA-Systemen
- Die Rolle von:
  - **Embeddings & Vektordatenbanken**
  - **Prompting & Kontextkonstruktion**
  - **Modellwahl und Kostenaspekten**
- Typische Herausforderungen bei der Umsetzung (z. B. Kontextfenster, Latenz, Retrieval-Qualität)
- **Optional:** Bewertung und Vergleich zweier Open-Source-RAG-Frameworks (z. B. LangChain, LlamaIndex, Haystack)

1.2 Agentensysteme (ca. 3 Seiten)

Recherchieren Sie aktuelle Ansätze zur Entwicklung von Agentensystemen mit LLMs.

Beschreiben Sie:

- Was ist ein Agent in diesem Kontext?
  Was unterscheidet ihn von einer einfachen Prompt-Chain? (z. B. Entscheidungslogik, Tool-Integration, Memory)
- Zentrale Fähigkeiten von Agentensystemen:
  - Tool-Nutzung (z. B. API-Zugriff, Dateizugriff, Rechner)
  - Orchestrierung (Planung/Koordination mehrerer Schritte)
  - Routing (Aufgabenverteilung an spezialisierte Subagenten)
  - Memory, Planning, Reaktion auf Nutzerkontext
- Typische Frameworks (z. B. LangChain Agents, AutoGen, CrewAI)

Vergleichen Sie zwei dieser Frameworks anhand von:

- Architektur
- Modularität
- Typischer Use Cases
- Technischer Einstiegshürden (z. B. Installation, Lernkurve, Dokumentation)

Teil 2: Implementierung eines KI-Systems mit RAG + Agentenkomponenten

2.1 RAG-System

Erstellen Sie ein funktionierendes Retrieval-Augmented Generation-System.

- **Wissensquelle:**
  - PDF- oder Word-Datei (max. 10 Seiten) **oder**
  - YouTube-Transkript (max. 15 Minuten)
- **Pipeline:**
  - Chunking → Embedding → semantisches Retrieval → LLM-Antwort
- **Tools:** frei wählbar (z. B. LangChain, Haystack, LlamaIndex oder eigene Implementierung)
- **Dokumentation:**
  - Architektur (inkl. **Diagramm**)
  - Modellkonfiguration (LLM, Embedding-Modell, Tokenlimits etc.)
  - Optional: Besonderheiten bei Datenvorverarbeitung

Achten Sie auf gute Codekommentare. Alle Abstrahierungen über Frameworks müssen erklärt werden, sprich:
Was wurde wegabstrahiert und wie funktioniert das?

2.2 Agentensystem

Bauen Sie Agentensysteme mit mindestens **drei der sechs Komponenten**:

- Structured Output (z. B. Ausgabe als Tabelle oder JSON)
- Tool-Nutzung (z. B. Rechner, Wetter-API, Dateizugriff)
- Prompt-Chaining (mehrstufige Aufgaben)
- Routing (verschiedene Aufgaben → verschiedene Subagenten)
- Parallelisierung (z. B. gleichzeitige Abfragen)
- Orchestrierung (Koordination von Teilaufgaben)

Empfehlenswert ist aus Aufwandssicht vor allem die Bearbeitung der ersten beiden Themen.

Die Agentenkomponenten **dürfen mit dem RAG-System interagieren** (z. B. Tool-Nutzung basierend auf abgerufenen Informationen).

(optional)
Teil 3: Transferanalyse und Reflexion (ca. 1-2 Seiten)

Wählen Sie eine konkrete Anwendung Ihrer Wahl, z. B.:

- Nachhilfe- oder Lern-Chatbot
- Internes Wissenssystem in einer Firma
- Beratungstool (z. B. Recht, Finanzen, Medizin)

Reflektieren Sie:

- **Wie müsste Ihr System angepasst werden**, um in dieser Domäne produktiv eingesetzt zu werden?
  (z. B. hinsichtlich Datenbasis, Antwortgeschwindigkeit, Nutzerführung)
- **Welche ethischen und rechtlichen Fragestellungen** ergeben sich?
  (z. B. Datenschutz, Halluzinationen, Verantwortung für Handlungen)
- **Welche technischen oder konzeptionellen Grenzen** sehen Sie beim Einsatz von RAG + Agenten in Ihrer gewählten Domäne?