Kurz-Exposé für Bachelorarbeit

Name: Joshua Tobias Hoffmann
Studiengang: B.Sc. Informatik, 5. Fachsemester
Fachbereich: Mathematik und Informatik, Philipps-Universität Marburg
Vorläufiger Titel: Entwicklung eines RAG-basierten Chatbots für Studienordnungen der Philipps-Universität Marburg

---

Thema

Diese Arbeit entwickelt ein Retrieval-Augmented Generation (RAG) System zur Beantwortung studienrelevanter Anfragen auf Basis offizieller universitärer Dokumente. Studierende müssen komplexe Prüfungsordnungen, Modulhandbücher und Veranstaltungskalender mit über 50 Seiten juristisch-formaler Sprache navigieren. Fehlinformationen können zu Studienzeitverlängerungen oder ungültigen Modulwahlen führen, während traditionelle Beratungsangebote überlastet sind. Der zu entwickelnde Chatbot soll 24/7-Verfügbarkeit mit faktenbasierten, quellenattribuierten Antworten kombinieren und dabei die kritische Herausforderung von LLM-Halluzinationen in diesem sicherheitskritischen Kontext adressieren.

Fragestellung

Die zentrale Forschungsfrage lautet: **Wie kann ein RAG-System gestaltet werden, um Faktentreue bei kritischen Informationen in strukturierten akademischen Dokumenten zu gewährleisten, ohne die Nutzbarkeit zu beeinträchtigen?**

Diese Hauptfrage gliedert sich in vier operative Teilfragen:

1. **Preprocessing:** Wie können heterogene akademische Dokumente (PDFs mit Text, Tabellen, Bildern) in ein konsistentes, maschinenlesbares Format überführt werden, das semantische Struktur bewahrt?

2. **Chunking:** Wie sollten vorverarbeitete Dokumente in semantisch kohärente Chunks segmentiert werden, die Kontexterhaltung mit Recheneffizienz balancieren?
3. **Retrieval:** Wie können hybride Retrieval-Strategien sowohl Precision als auch Recall für Studierendenanfragen maximieren?
4. **Generation:** Wie kann LLM-Generierung eingeschränkt werden, um Halluzinationen zu verhindern bei Erhaltung natürlicher, hilfreicher Antworten?

Die Relevanz dieser Fragestellung ergibt sich aus der existenziellen Bedeutung korrekter Informationen für Studienverläufe. Ein einzelner halluzinierter ECTS-Wert kann Studienabschlüsse verzögern, weshalb absolute Faktentreue bei kritischen Informationen wie Leistungspunkten, Fristen und Zulassungsvoraussetzungen erforderlich ist.

Theorie

Die Arbeit verortet sich im Schnittpunkt von Natural Language Processing, Information Retrieval und Human-Computer Interaction. Als theoretisches Fundament dient das RAG-Paradigma (Lewis et al., 2020), das parametrisches Wissen von LLMs mit nicht-parametrischem Zugriff auf externe Wissensbasen kombiniert. Im Gegensatz zu rein generativen Ansätzen ermöglicht RAG Quellenattribution und reduziert Halluzinationen durch Grounding in abgerufenen Dokumenten.

Zentral für strukturierte Dokumente ist die Forschung zu Document Understanding (Borchmann et al., 2021), die zeigt, dass traditionelle textbasierte Embeddings bei tabellarischen Daten versagen. Hier werden Ansätze wie Layout-aware Parsing und multimodale Embeddings relevant, die räumliche Beziehungen in PDFs erfassen.

Für die Evaluation stützt sich die Arbeit auf das RAGAS-Framework (Es et al., 2023), das RAG-Systeme entlang dreier Dimensionen bewertet: (1) Context Relevance (Wurden die richtigen Chunks abgerufen?), (2) Answer Faithfulness (Bleibt die Antwort den Quellen treu?), und (3) Answer Relevance (Adressiert die Antwort die Anfrage?). Diese Metriken erlauben eine objektive Quantifizierung der Systemleistung jenseits subjektiver Beurteilungen.

Ergänzend werden Konzepte aus der Uncertain AI relevant, insbesondere Abstaining-Mechanismen, die LLMs befähigen, bei unzureichender Informationsgrundlage keine Antwort zu geben statt zu halluzinieren.

Methodik und Untersuchungsgegenstand

Datengrundlage: Die Arbeit nutzt offizielle Dokumente eines konkreten Studiengangs der Philipps-Universität Marburg (Prüfungsordnung, Modulhandbuch, Veranstaltungskalender). Diese werden exemplarisch gewählt, da sie repräsentative Herausforderungen (Tabellen, Querverweise, juristische Sprache) aufweisen.

Methodisches Vorgehen: Die Entwicklung folgt einem iterativen Pipeline-Ansatz mit vier Hauptphasen:

1. **Preprocessing:** Evaluation verschiedener PDF-Parsing-Bibliotheken (PyMuPDF, pdfplumber, Tabula) zur Extraktion von Text und Tabellen. Vergleich ihrer Genauigkeit bei OCR-Text und komplexen Layouts.
2. **Chunking:** Empirischer Vergleich zwischen Sliding Window (mit 50-Token-Überlappung) und hierarchischer Segmentierung (entlang von Paragrafen). Bestimmung optimaler Chunk-Größen durch Variation zwischen 256-1024 Tokens.
3. **Retrieval:** Implementation eines hybriden Systems, das dense Retrieval (FAISS mit sentence-transformers) und sparse Retrieval (BM25) kombiniert. Evaluation anhand synthetisierter Studierendenanfragen mit Ground-Truth-Annotationen durch Studienberater.
4. **Generation:** Prompt-Engineering zur Halluzinationsvermeidung mit strikten Instruktionen ("Antworte NUR mit bereitgestellten Quellen"). Post-Processing-Validierung zur automatischen Detektion nicht-attributierter Aussagen.

Evaluation: Die Bewertung erfolgt dreistufig: (1) Inhaltliche Abdeckung relevanter Informationen in Chunks, (2) Retrieval-Genauigkeit (Precision/Recall der abgerufenen Chunks), (3) Generierungsqualität (Faithfulness, Halluzination Rate, Quellenattribution). Als Testdatensatz dienen 100 synthetisierte

Anfragen über fünf Kategorien: ECTS-Lookups, Voraussetzungsketten, Fristenfragen, Prüfungsformate und Out-of-Scope-Anfragen zum Testen von Abstaining-Verhalten.

Potenzielle Probleme: Die Arbeit antizipiert Herausforderungen bei (1) widersprüchlichen Informationen zwischen Dokumenten, die Priorisierungsregeln erfordern, (2) temporaler Gültigkeit verschiedener Ordnungsversionen, und (3) der Balancierung zwischen strikter Quellenanhaftung und natürlicher Sprachgenerierung.

Gliederung

1. **Einleitung**
   - Motivation und Problemstellung
   - Forschungsfragen und Zielsetzung
   - Aufbau der Arbeit
2. **Grundlagen und Related Work**
   - RAG-Systeme und Faktentreue
   - Verständnis strukturierter Dokumente
   - Abstaining und Unsicherheitsschätzung
   - Domänenspezifische Chatbots
   - Evaluation von RAG-Systemen
3. **Konzeption der RAG-Pipeline**
   - Systemarchitektur
   - Anforderungsanalyse für Studienordnungen
   - Design Decisions für jede Pipeline-Komponente
4. **Implementation**

- Data Preprocessing (PDF-Parsing, Tabellenextraktion)
- Data Chunking (Segmentierungsstrategien)
- Retrieval (Hybride Suche, Embedding-Modelle)
- Augmented Generation (Prompt-Engineering, vLLM-Integration)
- Webapplikation (Frontend/Backend, Authentifizierung)

5. **Evaluation**
   - Testdatensatz und Annotationsprozess
   - Quantitative Metriken (RAGAS, Precision/Recall)
   - Qualitative Analyse (Fallstudien, Fehlerklassifikation)
   - Vergleich mit Baseline-Systemen
6. **Diskussion**
   - Interpretation der Ergebnisse
   - Limitationen und Generalisierbarkeit
   - Implikationen für universitäre Informationssysteme
7. **Fazit und Ausblick**

Literatur

1. **Lewis, P. et al. (2020).** "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Proceedings of NeurIPS 2020.* – Grundlegendes Paper zum RAG-Paradigma.
2. **Es, S. et al. (2023).** "RAGAS: Automated Evaluation of Retrieval Augmented Generation." *arXiv preprint.* – Framework zur objektiven RAG-Evaluation.

3. **Borchmann, Ł. et al. (2021).** "DUE: End-to-End Document Understanding Benchmark." *Proceedings of NeurIPS 2021.* – Methoden für strukturierte Dokumentenverarbeitung.

4. **Gao, L. et al. (2023).** "Retrieval-Augmented Generation for Large Language Models: A Survey." *arXiv preprint.* – Systematischer Überblick über RAG-Architekturen und Best Practices.

5. **Zhang, Y. et al. (2023).** "R-Tuning: Teaching Large Language Models to Say 'I Don't Know'." *arXiv preprint.* – Abstaining-Mechanismen zur Halluzinationsvermeidung.

6. **Khattab, O. & Zaharia, M. (2020).** "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT." *Proceedings of SIGIR 2020.* – Effiziente Retrieval-Architektur.

7. **Lin, J. et al. (2021).** "Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations." *Proceedings of SIGIR 2021.* – Implementierungsgrundlage für hybride Retrieval-Systeme.

8. **Ji, Z. et al. (2023).** "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys.* – Umfassende Analyse von Halluzinationsphänomenen in LLMs.

9. **Asai, A. et al. (2023).** "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection." *arXiv preprint.* – Selbstreflexive Mechanismen zur Qualitätskontrolle.

10. **Wang, Y. et al. (2022).** "Document Understanding Dataset and Evaluation." *arXiv preprint.* – Benchmark-Datensätze für akademische Dokumente