Bachelor- / Masterarbeit

Wenn KI widerspricht: Entwicklung einer LLM-Architektur mit Feedback-Mechanismen

Aufgabenstellung und Vorgehensweise

Große Sprachmodelle (LLMs) wie ChatGPT oder Llama liefern häufig überzeugende Antworten, hinterfragen jedoch selten die Aussagen der Nutzenden. In sensiblen oder wissensintensiven Anwendungsszenarien (z. B. Beratung, Bildung oder Informationssysteme) ist es jedoch wichtig, dass ein System unbelegte, widersprüchliche oder unklare Aussagen erkennt und angemessen reagiert.

Ziel dieser Arbeit ist die Entwicklung einer LLM-basierten Architektur, die Aussagen von Nutzenden mithilfe externer Wissensquellen überprüft und situationsabhängig „zurückspricht“. Dies kann beispielsweise durch Nachfragen nach Belegen, das Aufzeigen möglicher Widersprüche oder das Signalisieren von Unsicherheit erfolgen.

Hierfür soll zunächst eine Architektur entworfen werden, die ein bestehendes LLM mit einer strukturierten Wissensquelle (z. B. Textsammlung oder Wissensgraph) kombiniert. Der Schwerpunkt liegt auf der Entwicklung eines Mechanismus zur Erkennung unstimmiger Aussagen sowie auf der Implementierung geeigneter Feedback-Strategien. Abschließend wird untersucht, in welchen Fällen das System sinnvoll reagiert und wo Grenzen bestehen.

Die Arbeit gliedert sich in folgende Schritte:

- Analyse bestehender Ansätze zu wissensbasiertem Feedback
- Konzeption einer Architektur zur Integration externer Wissensquellen
- Implementierung eines Prototyps zur Erkennung unstimmiger oder unbelegter Aussagen
- Entwicklung und Integration von Feedback-Strategien
- Evaluation anhand definierter Testszenarien

Anforderungen

- Programmiererfahrung in Python
- Grundkenntnisse zu LLMs und Retrieval-Verfahren
- Interesse an dialogfähigen KI-Systemen

Art der Arbeit

Bachelor-/ Masterarbeit

Ansprechperson

Vanessa Frohn | **E-Mail:** vfrohn@uni-wuppertal.de

Bachelor / Master thesis

When AI disagrees: Development of an LLM architecture with feedback mechanisms

Task and approach

Large language models (LLMs) such as ChatGPT or Llama often provide convincing answers – but rarely question the statements made by users. In sensitive or knowledge-intensive application scenarios (e.g., consulting, education, or information systems), however, it is important that a system recognizes unsubstantiated, contradictory, or unclear statements and responds appropriately.

The goal of this work is to develop an LLM-based architecture that verifies user statements using external knowledge sources and "responds" depending on the situation. This can be done, for example, by asking for evidence, pointing out possible contradictions, or signaling uncertainty.

To this end, an architecture will first be designed that combines an existing LLM with a structured knowledge source (e.g., text collection or knowledge graph). The focus is on developing a mechanism for recognizing inconsistent statements and implementing appropriate feedback strategies. Finally, we will examine in which cases the system responds appropriately and where its limitations lie.

The thesis is divided into the following steps:

- Analysis of existing approaches to knowledge-based feedback
- Design of an architecture for integrating external knowledge sources
- Implementation of a prototype for detecting inconsistent or unsubstantiated statements
- Development and integration of feedback strategies
- Evaluation based on defined test scenarios

Requirements

- Programming experience in Python
- Basic knowledge of LLMs and retrieval methods
- Interest in dialog-enabled AI systems

Type of work

Bachelor / Master thesis

Contact Person

Vanessa Frohn | **E-Mail:** vfrohn@uni-wuppertal.de