Bachelor- / Masterarbeit

Wenn KI Begriffe falsch versteht: Messung und Bewertung von konzeptionellem Fehlverhalten in LLM-Antworten

Aufgabenstellung und Vorgehensweise

Große Sprachmodelle können überzeugende Antworten generieren, zeigen jedoch häufig Schwächen im Umgang mit zentralen Konzepten, etwa durch inkonsistente Definitionen, falsche Zuordnungen oder fehlerhafte Abgrenzungen. Dieses sogenannte konzeptionelle Fehlverhalten stellt eine Herausforderung für die zuverlässige Nutzung solcher Systeme dar.

Ziel dieser Arbeit ist die Entwicklung eines Ansatzes zur systematischen Messung und Bewertung von konzeptionellem Fehlverhalten in Antworten von Sprachmodellen. Hierbei soll untersucht werden, wie sich solche Fehler operationalisieren und durch geeignete Kriterien erfassen lassen.

Zunächst werden mögliche Indikatoren für konzeptionelles Fehlverhalten identifiziert und in ein Bewertungsmodell überführt. Darauf aufbauend können geeignete Beispiele oder Testszenarien entwickelt werden, anhand derer Antworten analysiert und klassifiziert werden. Optional kann ein prototypisches System zur (teil-)automatisierten Analyse implementiert werden.

Die Arbeit legt den Schwerpunkt auf die Entwicklung nachvollziehbarer Bewertungskriterien sowie auf die strukturierte Analyse von Modellantworten.

Die Arbeit gliedert sich in folgende Schritte:

- Analyse von Formen konzeptionellen Fehlverhaltens in KI-Antworten
- Entwicklung von Kriterien zur Identifikation und Bewertung
- Erstellung eines Testdatensatzes oder Beispielszenarien
- Durchführung einer systematischen Analyse von Modellantworten

Anforderungen

- Programmiererfahrung in Python von Vorteil
- Grundkenntnisse in NLP oder maschinellem Lernen
- Interesse an der Analyse und Bewertung von KI-Systemen

Art der Arbeit

Bachelor-/ Masterarbeit

Ansprechperson

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

Bachelor / Master thesis

When AI Misunderstands Concepts: Measuring and Evaluating Conceptual Errors in LLM Responses

Task and approach

Large language models (LLMs) can generate convincing responses, but they often exhibit weaknesses in handling key concepts, such as inconsistent definitions, incorrect classifications, or flawed distinctions. This so-called conceptual error poses a challenge to the reliable use of such systems.

The goal of this work is to develop an approach for the systematic measurement and evaluation of conceptual errors in responses from language models. The aim is to investigate how such errors can be operationalized and captured using appropriate criteria.

First, potential indicators of conceptual errors are identified and incorporated into an evaluation model. Building on this, suitable examples or test scenarios can be developed to analyze and classify responses. Optionally, a prototype system for automated analysis can be implemented.

The thesis focuses on the development of transparent evaluation criteria as well as on the structured analysis of model responses.

The thesis is divided into the following steps:

- Analysis of forms of conceptual errors in AI responses
- Development of criteria for identification and evaluation
- Creation of a test dataset or example scenarios
- Conducting a systematic analysis of model responses

Requirements

- Programming experience in Python is a plus
- Basic knowledge of NLP or machine learning
- Interest in the analysis and evaluation of AI systems

Type of work

Bachelor / Master thesis

Contact Person

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