DIALOKIA - Integration dialektischer Logik in KI-Architekturen

Professuren:	Intelligent Information Systems Prof. Dr. Benno Stein
	Fakultät Medien
Laufzeit:	1. Oktober 2024 bis 30. September 2027
Drittmittelgeber:	BMBF
Fördersumme:	386.200 Euro

Beschreibung:

Große Sprachmodelle (Large Language Models, LLMs) können einfache Schlüsse aus Texten ziehen. In dieser Fähigkeit liegt aber auch eine Gefahr bei der Verwendung, denn Sprachmodelle sind dahin gehend trainiert, Aussagen möglichst selbstsicher zu formulieren.

Gemeinsam mit Prof. Dr. Matthias Hagen (Universität Jena) widmet sich das Projekt DIALOKIA deshalb der Frage, wie man Sprachmodelle besser den Prinzipien der Logik folgen lassen kann. Im Fokus des Projektes steht dabei die dialektische Betrachtungsweise: so lässt sich eine Allaussage durch Angabe eines Gegenbeispiels anzweifeln, oder eine Schlussfolgerung auf Basis einer Expertenaussage lässt sich anzweifeln, indem die Autorität des Experten auf dessen Fachgebiet in Frage gestellt wird. Unter Verwendung des Prinzips der kritischen Fragen verfolgt das Projekt die Idee, valide Argumentationen durch die sog. Lehrer-Schüler-Überwachung zwischen zwei Sprachmodellen zu trainieren.

Summary:

Large language models (LLMs) can draw simple conclusions from texts. However, this ability also poses a risk when using them, as language models are trained to formulate statements as confidently as possible.

Together with Prof. Dr. Matthias Hagen (University of Jena), the DIALOKIA project is therefore dedicated to the question of how language models can better follow the principles of logic. The project focuses on the dialectical approach: a general statement can be challenged by providing a counterexample, or a conclusion based on an expert statement can be challenged by questioning the authority of the expert in their field. Using the principle of critical questions, the project pursues the idea of training valid argumentation through so-called teacher-student supervision between two language models.

Webseite der Professur: https://weimar.webis.de

Kontakt:
Bauhaus-Universität Weimar
Intelligent Information Systems
Prof. Dr. Benno Stein
benno.stein@uni-weimar.de

Bauhausstr.11
99423 Weimar
Tel. 03643 / 58 37 27