CEditor – Der LLM-Lügendetektor

Cedric Mack (17)
79639 Grenzach-Wyhlen, Hans-Thoma-Gymnasium, Lörrach

Large Language Models (LLMs), wie sie etwa hinter ChatGPT stehen, beeindrucken durch ihre Fähigkeit, authentische Texte zu generieren. Doch diese Modelle haben eine kritische Schwäche: Sie neigen dazu, sogenannte Halluzinationen – Aussagen, die überzeugend klingen, aber faktisch falsch sind, zu generieren. Diese Halluzinationen stellen deutliche Risiken dar, vor allem in sensiblen Bereichen wie der Medizin oder dem Rechtswesen. Mit zunehmender Modellgröße sinkt zwar die Wahrscheinlichkeit einer Halluzination, gleichzeitig steigen aber auch die Betriebskosten der LLMs. Dieses Projekt untersucht erfolgreich Möglichkeiten zur Erkennung dieser Halluzinationen und entwickelt einen neuen Ansatz, der durch Kombination verschiedener LLMs mit verschiedener Größen und Fähigkeiten Halluzinationen vermindert und dabei die Betriebskosten gering hält. Der neue Ansatz wird in der im Projekt entwickelten Software „CEditor“ implementiert und mit einer Benutzeroberfläche zugänglich gemacht.

SPARTE:
Jugend forscht

ERARBEITUNGSORT:
phaenovum Schülerforschungszentrum Lörrach-Dreiländereck

BETREUUNG:
Marcel Neidinger
Pirmin Gohn