ZUSAMMENFASSUNG

Large Language Models sind im Trend und finden Anwendung in den verschiedensten Themenbereichen. Dabei zeigte sich, dass die Modelle auch ohne Anpassung das Potential zur Nutzung in neuen, beim Training nicht vorgesehenen Domänen besitzen. Die Eignung zur Informationsextraktion, insbesondere im Kontext von Rechtsdokumenten wie Verträgen, ist dabei jedoch wenig untersucht. Vor allem allgemeine, für die spezifische Anwendung nicht vorgesehenen Chat- oder Instruction-Modellen fanden aufgrund der höheren Komplexität des Bereichs wenig Betrachtung. Modelle wie diese werden allerdings in vielen Fällen bereits von Unternehmen zur allgemeinen Unterstützung der Mitarbeitenden genutzt, und liegen somit bereits häufig vor.

Das Ziel der vorliegenden Arbeit ist es daher, solche Modelle im Kontext der Informationsextraktion auf juristischen Dokumenten zu untersuchen. Hierbei wird die exakte Extraktion bestimmter Fachdaten als auch die Identifikation bzw. Klassifikation von Vertragsklauseln betrachtet. Ebenso wird der Einfluss verschiedener Prompttechniken und unterschiedlicher Kontextlängen auf die Modellergebnisse erforscht. Zuletzt wird die Eignung allgemeiner Large Language Models in Abhängigkeit der Prompttypen für derartig komplexe und spezifische Anwendungsfälle bewertet.

Dazu wurden sechs öffentliche Modelle, basierend auf Falcon und Llama2, sowie OpenAIs populäres GPT-3.5-Turbo ausgewählt und auf einem annotierten Vertragsdatensatz, dem CUAD, zur Extraktion und Klassifikation ausgewertet. Hierfür wurden fünf verschiedene Promptvorlagen auf Basis der populären Techniken Zero-Shot, Few-Shot, Chain-Of-Thought, Chain-Of-Thought-Automatic und Tree-Of-Thoughts-Automatic entwickelt und verglichen. Zusätzlich wurden drei unterschiedliche Kontextgrößen betrachtet, bestehend aus 1, 5 und 10 Sätzen.

Die Ergebnisse zeigen dabei, dass die Anwendbarkeit auch bei komplexen Themengebieten gegeben ist, weshalb eine Feinjustierung der Modelle für neue Aufgaben nicht erforderlich erscheint. Insbesondere GPT-3.5-Turbo demonstriert beeindruckende Werte mit einer Gesamtaccuracy von 80.93 bei Nutzung von Few-Shot. Diese Prompttechnik offenbart dabei eindeutig die besten Resultate. Limitierungen lassen sich bei den Falcon-Modellen sowie den Prompttechniken Zero-Shot, Chain-Of-Thought-Automatic und Tree-Of-Thoughts-Automatic erkennen. Außerdem empfiehlt es sich, die Kontextgröße so gering wie möglich zu halten, da mit ansteigender Größe die Ergebnisgüte der Modelle fällt.