Extraktion von Vertragsinformationen mittels Large Language Models

Nils Hartmann

Referent: Prof. Dr. Markus Döhring | Korreferentin: Prof. Dr. Antje Jahn
Darmstadt University of Applied Sciences

Einleitung

Large Language Models (LLMs) erfahren einen außergewöhnlichen Aufwärtstrend und kommen bei Anwendungsfällen weit über der anfänglichen Chatbotinteraktion hinaus erfolgreich zum Einsatz, wie beispielsweise der maschinellen Übersetzung [1] oder Generierung von Bildern und Audiodaten [5]. Die Anwendung zur Extraktion von Fachdaten ist dabei jedoch wenig erprobt, speziell im Kontext von Rechtstexten, welche aufgrund ihrer Komplexität eine besondere Herausforderung darstellen. Unter diesen Rahmenbedingungen gestaltet sich die Implementierung einer Automatisierungslösung bisher kompliziert sowie zeit- und kostenintensiv, weshalb meist auf spezifische, extra für den Anwendungsfall konzipierte und feinjustierte Modelle zurückgegriffen wird. Da die aktuellen LLMs jedoch ein besseres Verständnis von natürlicher Sprache sowie den darin teils vielschichtigen Beziehungen aufweisen, und in vielen Betrieben bereits Modelle unterstützend als Chatbot bereitgestellt werden, könnten diese kostengünstig zur Unterstützung im manuellen Prozess beitragen. Vor allem die Erkenntnis, dass sich aktuelle Sprachmodelle mit passendem Prompt Engineering auch für Aufgaben außerhalb der konzipierten Domäne eignen [1], wirft die Frage nach den Grenzen hierbei auf.

Ziel dieser Arbeit ist es daher, die Nutzbarkeit von LLMs ohne spezifische Feinjustierung zur Anwendung der Informationsextraktion und Klauselidentifikation auf Rechtstexten zu untersuchen, und im Zuge dessen auf folgende Forschungsfragen einzugehen:

- Wie unterscheiden sich die Ergebnisse abhängig vom genutzten Modell?
- Welchen Einfluss besitzen Promptdesign, Kontextlänge und betrachtete Fachdaten? Lassen sich klare Empfehlungen ableiten?
- Wie konsistent wird die Vorgabe einer festen Antwortstruktur abhängig der Modelle, Promptdesigns und Kontextlängen eingehalten?

Entwickelte Prompts

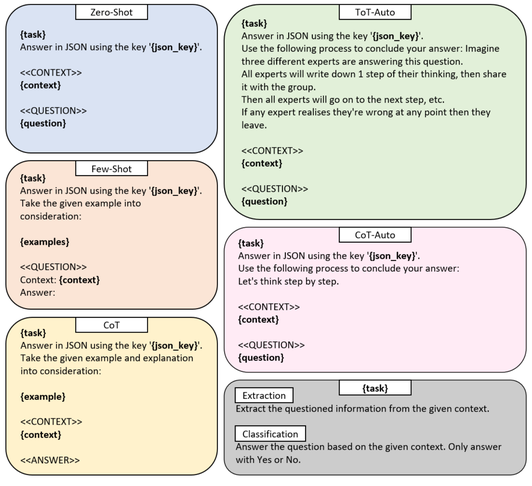

Figure 1. Übersicht der entwickelten Promptvorlagen.

Auf Grundlage der bekannten und bereits erprobten Designs Zero-Shot, Few-Shot, Chain-of-Thought (CoT) und Tree-of-Thoughts (ToT) wurden fünf Vorlagen erstellt, illustriert in Abbildung 1. Alle fettgedruckten Stellen in geschweiften Klammern entsprechen Platzhaltern für tatsächlichen Text.

Datensatz

Als Grundlage diente der Contract Understanding Atticus Dataset (CUAD) [2], welcher aus 510 kommerziellen Verträgen mit über 13000 manuellen Annotationen für insgesamt 41 verschiedene Fachdaten und Klauseln besteht. Daraus wurden 18 Fachdaten ausgewählt und pro Fachdatum 100 Samples aus dem annotierten Pool entnommen, in drei verschiedenen Kontextgrößen (1, 5 und 10 Sätzen), wodurch der finale Datensatz somit 5400 Einträge umfasste.

Ergebnisse der Einhaltung der JSON Vorgabe

Die Ergebnisse zur Einhaltung der JSON-Vorgabe aus Tabelle 1 zeigen deutlich, dass Few-Shot und CoT generell zu signifikant höheren JSON-Anteilen in den Modellantworten führten als die restlichen Prompttypen. Bei Auslassen der Besonderheit von Llama2-13B bei Few-Shot, bei welchem eine Art „Schutzmechanismus“ vor Falschaussagen im juristischen Kontext ausgelöst wurde, erreichte der durchschnittliche JSON-Anteil 97.08 %. Auf Modellebene konnte GPT-3.5-Turbo gar einen 100 % JSON-Anteil mit Few-Shot erreichen. Da die Inklusion einer beispielhaften Antwortstruktur Few-Shot und CoT von den restlichen Designs abgrenzt und zugleich diese beiden Typen zu den maßgeblich besten Ergebnissen führten, ist davon auszugehen, dass eben diese Angabe eines Beispiels für eine konsistente Einhaltung einer vorgegeben Struktur vonnöten ist.

Prompt	JSON	JSON-Anteil in %
Few-Shot	33153	88.66
CoT	26609	70.39
Zero-Shot	8476	22.42
ToT-Auto	7652	20.24
CoT-Auto	7520	19.89

Table 1. Einhaltung der JSON-Vorgabe aggregiert pro Prompttyp. Gesamtanzahl der möglichen JSON entspricht 37800.

Ebenso zeigte sich Few-Shot am robustesten im Bezug zur Kontextgröße. Hierbei schwankten die JSON-Anteile pro Modell um weniger als 4 Prozentpunkte, während diese Abweichungen bei steigender Kontextgröße bei den restlichen Promptdesigns über 10 Prozentpunkten lag, teils bis über 25.

Ergebnisse der Extraktionsaufgabe

Bei der Extraktion der Fachdaten wurden starke modellabhängige Unterschiede erkennbar (siehe Tabelle 2), ebenso wie der Einfluss der Prompttechnik. So zeigte erneut Few-Shot die besten Ergebnisse, gefolgt von CoT. Im Modellvergleich erreichte GPT-3.5-Turbo wiederholt die performantesten Resultate mit einer Accuracy >80, gefolgt von den beiden Open-Source Llama2 Varianten Nous-Hermes und WizardLM mit Accuracys von etwa 71 bzw. 70, wobei all diese Werte durch Few-Shot erzielt wurden.

Modell	Prompttyp	Korrekte Ergebnisse	Accuracy	JSON-bedingte Accuracy
GPT-3.5-Turbo	Few-Shot	4‘370	80.93	80.93
Nous-Hermes	Few-Shot	3‘840	71.11	71.81
WizardLM	Few-Shot	3‘799	70.35	71.61
Llama2-7B	Few-Shot	3‘455	63.98	65.46
Falcon-40B	Few-Shot	3‘005	55.65	59.47
Llama2-13B	CoT	2‘701	50.02	61.10
Falcon-7B	Few-Shot	2‘211	40.95	43.17

Table 2. Höchste Performance jedes Modells nach Accuracy. Gesamtanzahl der möglichen korrekten Ergebnisse: 5‘400.

Bei Betrachtung der Fachdaten erwies sich Renewal Term als augenscheinlich komplexer. Die Accuracy betrug maximal 45.67 (Nous-Hermes). Durch manuelle Auswertung der Modellantworten konnte eine mögliche Erklärung idenzifiziert werden: Die Modellantworten enthielten häufig Zusätze zu der angefragten Information wie die maximale Anzahl an Renewal Terms, welche auch bei manueller Extraktion in den Ermessenspielraum fallen würden. Bei Einbeziehen dieser Antworten würde sich die Accuracy bei Nous Hermes auf 68.66 steigern. Ebenfalls zeigte sich eine Abnahme der Antwortgüte bei steigender Kontextgröße (durchschnittlich um -28.70%). Die größeren Modelle GPT-3.5-Turbo und Falcon-40B konnten dabei durch ihre höhere Anzahl an Parametern ein besseres Verständnis bei größerem Kontext bewahren, allerdings ließ sich auch bei diesen Modellen ein Abfall um etwa -13% beobachten.

Fazit

Die Untersuchungen konnten erfolgreich zeigen, dass aktuelle, nicht feinjustierte LLMs auch für spezifische Anwendungsfälle auf komplexen juristischen Dokumenten anwendbar sind. Eine eigene zeit- und kostenintensive Feinjustierung für jede neue Aufgabe ist somit mit voriger Promptentwicklung nicht zwangsweise erforderlich. Dabei zeigte sich Few-Shot eindeutig als die beste Prompttechnik, sowohl zur konsistenten Einhaltung einer JSON-Struktur in der Ausgabe, als auch für die eigentliche Extraktionsaufgabe. Da die wenigen bisherigen Forschungen zur genauen Extraktion verschiedener Fachdaten CoT benutzten, könnte Few-Shot somit eine Verbesserung darstellen. Auf Modellebene zeigte sich wenig überraschend GPT-3.5-Turbo als am besten geeignet, gefolgt von Nous-Hermes und WizardLM als kleinere Open-Source Alternativen.

Grenzen zeigen sich klar bei Zero-Shot, CoT-Auto und ToT-Auto auf. Diese Prompttypen führten bei der JSON-Vorgabe zu der schlechtesten Einhaltung und wiesen modellabhängige Inkonsistenzen bei der Nutzung zur Extraktion oder Klassifikation auf. Ebenso konnten die Falcon-Modelle generell nicht überzeugen. Zusätzlich führte ein Zunahme der Kontextgröße zu durchweg schlechteren Ergebnissen, weshalb ein möglichst granulares Retrieval der relevanten Textstellen für bestmögliche Ergebnisse genutzt werden sollte.

Ausblick

Bereits während der Anfertigung dieser Arbeit zeigten erste Publikationen zu GPT-4 eine deutliche Steigerung der Robustheit gegenüber steigenden Kontext und Einfluss der Promptstruktur bei zugleich höhere Antwortgüte [3] [4]. Somit ist davon auszugehen, dass die dargelegten Faktoren zukünftig eine fortlaufend geringere Rolle bei LLMs darstellen.

Im kommerziellen Kontext lässt sich aber auf Grundlage der Ergebnisse festhalten, dass Chatmodelle auch jetzt schon mit Few-Shot ein hohes Potential zur Nutzung in komplexen und fachspezifischen Anwendungsfällen aufweisen, und mit geringem Kosten- und Implementierungsaufwand als Unterstützung in vorhandenen Prozessen dienen können.

Referenzen

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, and Arvind Neelakantan et al.
Language Models are Few-Shot Learners.
Advances in Neural Information Processing Systems, 33:1877–1901, 2020.

[2] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball.
CUAD: an expert-annotated NLP dataset for legal contract review.
In *Proceedings of the Neural Information Processing Systems, December 2021*, 2021.

[3] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo.
Gpt-4 passes the bar exam.
Available at SSRN 4389233, 2023.

[4] Mosh Levy, Alon Jacoby, and Yoav Goldberg.
Same task, more tokens: the impact of input length on the reasoning performance of large language models.
CoRR, abs/2402.14848, 2024.

[5] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.
Zero-shot text-to-image generation.
In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2021.