V. Abstract

Titel:	Optimierung von Large-Language-Model basierten Datenextraktionsprozessen zur Strukturierung interner E-Mail-Kommunikationsdaten
Verfasser:	Julian Konz
Kurs:	WWI 23 SCB
Ausbildungsbetrieb:	SAP SE

Im Rahmen dieser Arbeit wird ein Proof of Concept zur automatisierten Extraktion strukturierter Reportingdaten aus E-Mails der SAP AI Regional Implementation Group durch Large Language Models entwickelt und optimiert. Dies erfolgte entlang des Cross Industry Standard Process for Data Mining und evaluierte den Einfluss der Modellauswahl und Promptingstrategie auf die Datenqualität.

In einem experimentellen Aufbau werden drei Modelle (GPT-4o, o3-mini, o1) mit neun Prompting-Strategien (Baseline, Chain-of-Thought und Self-consistency, jeweils als Zero-/One- und Few-Shot-Prompt) verglichen. Die Analyse zeigte einen signifikanten Einfluss der Modellwahl auf die Extraktionsgenauigkeit. Das Modell o1 erzielte dabei die höchsten Ergebnisse gegenüber den anderen Modellen. Die beste Konfiguration erreichte ein Self-Consistency-One-Shot-Prompt des Modells OpenAI o1 mit einer Datenextraktionsgenauigkeit von 74% und einer Identifikationsgenauigkeit des zugehörigen Reporting-Eintrags von 95%.

Die entwickelte Modellierung erfüllt die von der SAP AI Regional Implementation Group definierten Qualitätskriterien für die automatisierte Datenextraktion. Auf Basis dieser Ergebnisse wird nach Validierung auf einem erweiterten Datensatz eine schrittweise Produktivnahme empfohlen.