



Optimierung von Large-Language-Model basierten Datenextraktionsprozessen zur Strukturierung interner E-Mail-Kommunikationsdaten

SPERRVERMERK

Zweite Projektarbeit

aus dem Studiengang Wirtschaftsinformatik Sales & Consulting an der Dualen
Hochschule Baden-Württemberg Mannheim

von

Julian Konz

Bearbeitungszeitraum: 17.02.2025 - 05.05.2025

Matrikelnummer, Kurs: 3468097, WWI23SCB

Studiengangsleiter: Prof. Dr. Clemens Martin

Ausbildungsfirma: SAP SE
Dietmar-Hopp-Allee 16
69190 Walldorf, Deutschland

Betreuer der Ausbildungsfirma: Felix Bartler
felix.bartler@sap.com
+496227750225

Wissenschaftliche Betreuerin: Prof. Dr. Sarah Detzler
sarah.detzler@dhw.de
+4962141051412

I. Eidestattliche Erklärung

Ich versichere hiermit, dass ich meine Projektarbeit mit dem Thema: „Titel“ selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort, Datum

Unterschrift

II. Sperrvermerk

Die nachfolgende Arbeit enthält vertrauliche Daten und Informationen der SAP SE, Dietmar-Hopp-Allee 16, 69190 Walldorf, Deutschland. Der Inhalt dieser Arbeit darf weder als Ganzes noch in Auszügen Personen außerhalb des Prüfungsprozesses und des Evaluationsverfahrens zugänglich gemacht werden. Veröffentlichungen oder Vervielfältigungen der Projektarbeit - auch auszugsweise - sind ohne ausdrückliche Genehmigung der SAP SE in einem unbegrenzten Zeitraum nicht gestattet. Über den Inhalt dieser Arbeit ist Stillschweigen zu wahren.

SAP und die SAP Logos sind eingetragene Warenzeichen der SAP SE. Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in dieser Arbeit berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedem benutzt werden dürfen.

III. Gleichbehandlung der Geschlechter

In dieser Praxisarbeit wird aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet. Weibliche und anderweitige Geschlechteridentitäten werden dabei ausdrücklich mitgemeint, soweit es für die Aussage erforderlich ist.

IV. Disclaimer

Ein Teil der Literatur, die für die Anfertigung dieser Arbeit genutzt wird, ist nur über die Plattformen o'Reilly, SAP Multi Experience Platform und MyEducator abrufbar. Bei diesen Ressourcen existieren keine Seitennummern, es wird bei Verweisen stattdessen die Kapitelnummer oder der Kapitelnamen angegeben.

Um den Lesefluss zu verbessern, werden Abbildungen, Prompts und Tabellen, die den Lesefluss stören, im Anhang platziert, auf den im Text zusätzlich verwiesen wird.

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.

Inhaltsverzeichnis

| | |
|--|-----------|
| 1. Einleitung | 1 |
| 1.1. Motivation | 1 |
| 1.2. Ziel und Gang | 2 |
| 2. Methodik | 3 |
| 3. Grundlagen | 5 |
| 3.1. AI Regional Implementation Group | 5 |
| 3.2. Cross Industry Standard Process for Data Mining | 7 |
| 3.3. Large Language Models | 10 |
| 3.3.1. Vorteile von Large Language Models | 11 |
| 3.3.2. Risiken und Herausforderungen | 12 |
| 3.3.3. Large Language Models in der Datenextraktion | 13 |
| 3.4. Services & Tools | 14 |
| 3.4.1. SAP Generative AI Hub | 14 |
| 3.4.2. SAP Multi Experience Platform | 16 |
| 3.5. Datenextraktionsoptimierung | 17 |
| 3.5.1. Modellwahl | 17 |
| 3.5.2. Prompt Engineering | 19 |
| 3.6. Evaluationsmetriken | 21 |
| 3.6.1. Precision, Recall und F1-Score | 21 |
| 3.6.2. Recall-Oriented Understudy for Gisting Evaluation | 23 |
| 3.6.3. Accuracy | 24 |
| 4. Praxis | 25 |
| 4.1. Business Understanding | 25 |
| 4.2. Data Understanding | 29 |
| 4.3. Data Preparation | 34 |
| 4.4. Modelling | 37 |
| 4.5. Evaluation | 42 |
| 4.5.1. Datenextraktionsanalyse | 42 |
| 4.5.2. Analyse der Opportunity-Zuordnung | 46 |
| 4.5.3. Gesamtbetrachtung & Konfigurationswahl | 48 |

| | |
|---|-----------|
| 4.5.4. Hypothesenvalidierung | 51 |
| 4.6. Deployment | 54 |
| 5. Schlussbetrachtung | 55 |
| 5.1. Zusammenfassung der Ergebnisse | 55 |
| 5.2. Einordnung der Ergebnisse | 56 |
| 5.3. Herausforderungen und Limitationen | 56 |
| 5.4. Ausblick | 58 |
| i. Literaturverzeichnis | i |
| ii. Anhang | xix |

Abbildungsverzeichnis

| | |
|---|------------|
| Abbildung 1: Abstrahierte Prozesse in der RIG AI in Anlehnung an [1] | 5 |
| Abbildung 2: Phasen des CRISP-DM Prozess [2] | 7 |
| Abbildung 3: Orchestration Service des SAP GenAI Hub [3] | 15 |
| Abbildung 4: Architektur einer MXP Applikation [4] | 16 |
| Abbildung 5: Attribute und Häufigkeit der Daten. Eigene Darstellung. | 29 |
| Abbildung 6: Anzahl der Kunden je E-Mail. Eigene Darstellung. | 30 |
| Abbildung 7: Anzahl an E-Mails je Sendergruppe. Eigene Darstellung. | 31 |
| Abbildung 8: Analyse des E-Mail-Betreffs. Eigene Darstellung. | 32 |
| Abbildung 9: Ergebnisse einer initialen Extraktion. Eigene Darstellung. | 33 |
| Abbildung 10: Anteil relevanter Daten. Eigene Darstellung. | 34 |
| Abbildung 11: Anteil identifizierten Opportunities je Qualität. Eigene Darstellung. .. | 36 |
| Abbildung 12: Extraktionspipeline des PoCs. Eigene Darstellung. | 39 |
| Abbildung 13: Durchschnittliche Metriken je Feld (aggregiert). Eigene Darstellung. . | 42 |
| Abbildung 14: Durchschnittlicher F1-Score je Feld. Eigene Darstellung. | 43 |
| Abbildung 15: Durchschnittliche Accuracy je Feld. Eigene Darstellung. | 44 |
| Abbildung 16: Durchschnittliche Metriken (Reportingfelder). Eigene Darstellung. . . | 46 |
| Abbildung 17: Identifikationsrate (beste Konfiguration). Eigene Darstellung. | 47 |
| Abbildung 18: Durchschnittliche Metriken je Prozess. Eigene Darstellung. | 49 |
| Abbildung 19: Durchschnittliche Metriken je Prozess. Eigene Darstellung. | 50 |
| Abbildung 20: Auswirkung von Prompt/LLM. Eigene Darstellung. | 52 |
| Abbildung 21: SAP AI Services - Aufgaben der RIG [5, S. 8] | xix |

Tabellenverzeichnis

| | |
|--|------|
| Tabelle 1: Confusion Matrix für eindimensionale Klassifizierung je Klasse [6, S. 3] . | 22 |
| Tabelle 2: Übersicht des Validierungsdatensatzes auf Basis von Abbildung 9 | 35 |
| Tabelle 3: Verwendete Python Libraries | xx |
| Tabelle 4: Zu extrahierende Datenfelder je Kunde | xx |
| Tabelle 5: Anzunehmende Statuswerte einer Aktivierung | xxi |
| Tabelle 6: Material ID's der SAP Business AI Produkte | xxii |

Promptverzeichnis

| | |
|---|--------------|
| Prompt 1: Struktur des verwendete Basisprompts in Anlehnung an [7, S. 2-3] | 40 |
| Prompt 2: Struktur des verwendeten Chain-of-Thought Prompts | xxii |
| Prompt 3: Struktur des verwendeten Self-consistency Prompts | xxiii |
| Prompt 4: Anhang je Beispiel bei Verwendung von One-/Few-Shot-Prompting | xxiii |

Abkürzungsverzeichnis

| | |
|-----------------|---|
| AE | Account Executive |
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| CRISP-DM | Cross Industry Standard Process for Data Mining |
| CRM | Customer Relationship Management |
| CSV | Comma-separated values |
| CoT | Chain-of-Thought |
| GenAI | Generative Artificial Intelligence |
| ID | Identifier |
| JSON | JavaScript Object Notation |
| KI | Künstlicher Intelligenz |
| KPI | Key Performance Indicator |
| LLM | Large Language Model |
| MXP | Multi Experience Platform |
| NLP | Natural Language Processing |
| PLM | Pre-trained Language Model |
| PoC | Proof of Concept |
| RIG | Regional Implementation Group |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |

| | |
|-------------|-------------------------------------|
| TTFT | Time-To-First-Token |
| aCMS | Augmented Content Management System |
| pp | Prozentpunkt |

Variablenverzeichnis

| | |
|-----------|---|
| B | Menge aller n-gramme im vorhergesagten Text |
| D | Menge aller n-gramme im Referenztext |
| FN | False Negative |
| FP | False Positive |
| H | Menge an tatsächlichen Klassen |
| L | Menge an Klassen |
| N | Anzahl an Iterationen |
| P | Menge an vorhergesagten Klassen |
| SE | Standardfehler |
| TN | True Negative |
| TP | True Positive |
| b | vorhergesagter Text |
| c | Klasse |
| d | Referenztext |
| k | Metrik-Gewichtung |
| λ | Textlänge |
| α | Signifikanzniveau |
| μ | Allgemeine Mittelwert |
| σ | Standardabweichung |

ω

Wort

1. Einleitung

1.1. Motivation

Auf Basis des Fundaments von C. E. Shannon [8] und der Forschung von A. Vaswani *u. a.* [9] zu neuronalen Netzen, welche kontextabhängige Informationen aus komplexen Eingaben nutzen können, um korrekte Modellausgaben zu generieren, ist ein neues Zeitalter von Künstlicher Intelligenz (KI), auf Englisch als Artificial Intelligence (AI) bekannt, eingetreten. Diese Entwicklungen führten zu erheblichen Fortschritten in der Integration von AI-Technologien in vielfältige Anwendungsfelder, insbesondere in betriebswirtschaftliche Prozesse [10, S. 8-9]. In der Unternehmenswelt eröffnet der Einsatz von AI signifikante Potenziale zur Effizienzsteigerung und Automatisierung ehemals manueller Aufgaben [11].

SAP SE nutzt AI in der SAP Business AI Produktreihe, welche KI-gestützte Lösungen bereitstellt, die sich nahtlos in Geschäftsprozesse einfügen [12]. Ein Anwendungsfeld stellt die Datenextraktion dar, die aus unstrukturierten Datenquellen strukturierte Daten gewinnt. Diese können zur Automatisierung manueller Prozesse genutzt werden [13, S. 263-264]. SAP arbeitet daran, aufwendige interne Prozesse mittels strukturierter Datenextraktion zu automatisieren [14].

Der Verkauf sowie die Integration, auch als Aktivierung bezeichnet, der Business AI-Produkte bei Kunden gehören zu den Hauptzielen der SAP im Jahr 2025 [15]. Zur Messung der Erreichung dieses Ziels sind Reportingdaten notwendig, welche die Anzahl der erfolgreichen Aktivierungen dokumentieren. Durch das steigende Interesse von Kunden an SAP Business AI-Produkten arbeitet die AI Regional Implementation Group (RIG), die für die Aktivierung von Business AI zuständige Beratungsgruppe der SAP, unter erhöhtem Druck [11]. Um sich auf wertschöp-

fende Aktivitäten für die Kunden zu fokussieren, soll ein Ansatz entwickelt werden, mittels der Anwendung von AI aus E-Mails strukturierte Reportingdaten auszulesen und diese automatisiert in das Reporting der RIG einzupflegen.

1.2. Ziel und Gang

Ziel dieser Arbeit ist eine Automatisierung des RIG Reportingprozesses. Hierzu wird ein Ansatz der Automatisierung entwickelt, als Proof of Concept (PoC) implementiert und hinsichtlich seiner Qualität evaluiert.

Der aktuelle Reportingprozess basiert auf der manuellen Extraktion von Reportingdaten aus E-Mails sowie deren Übertragung in eine interne Reporting-Anwendung. Aufgrund seiner repetitiven Natur soll ein automatisierter Ansatz erprobt werden, welcher strukturierte Reportingdaten mittels AI-gestützter Datenextraktion erfasst, und automatisiert in das Reporting überführt. Dies sieht sowohl eine Entlastung der AI RIG als auch eine verbesserte Informationsgrundlage für strategische Entscheidungen der SAP vor. Folgende Forschungsfragen an die Automatisierung des Reportingprozesses der RIG werden gestellt:

- **Forschungsfrage 1:** Wie kann der Reportingprozess der RIG mithilfe von Automatisierung durch AI verbessert werden?
- **Forschungsfrage 2:** Welche Modellwahl und Prompting-Techniken haben einen positiven Einfluss auf die Genauigkeit von E-Mail-Extraktionsaufgaben?

Die methodische Grundlage bildet das Cross Industry Standard Process for Data Mining (CRISP-DM)-Modell. Nach einer Einführung in die Prozesse der RIG, eingesetzte Tools sowie relevante theoretische Konzepte in Abschnitt 3 werden in Abschnitt 4 die Phasen von CRISP-DM zur Entwicklung des PoCs durchlaufen. Eine abschließende Betrachtung erfolgt in Abschnitt 5.

2. Methodik

Um den Reportingprozess der RIG durch strukturierte E-Mail-Kommunikationsdatenextraktion zu verbessern und deren Genauigkeit zu steigern, wird in dieser Arbeit der CRISP-DM-Prozess als methodischer Rahmen verwendet. CRISP-DM hat sich seit seiner Einführung als De-facto-Standard für Data-Mining-Projekte etabliert und bietet einen bewährten, strukturierten Ansatz für die Durchführung von Datenextraktionsprojekten [16, S. 529-532], [17, S. 10], [18, S. 4-5]. Da es sich bei dieser Arbeit um die Entwicklung eines PoC handelt, werden Kriterien festgelegt, anhand derer entschieden wird, ob eine Produktivnahme des PoC aussteht oder welche Schritte für eine produktive Implementierung ausstehen.

Zur Datenextraktion werden Pre-trained Language Models (PLMs) verwendet. PLMs, Teilgruppe der Large Language Models (LLMs), sind vortrainierte Modelle, welche auf großen Datenmengen vortrainiert sind und hinsichtlich Datenextraktionsaufgaben in anderen Domänen herkömmliche regelbasierte und statistische Natural Language Processing (NLP)-Ansätze in Genauigkeit übertreffen [19, 5167-5171]. Darüber hinaus kann mittels Optimierungsverfahren ihre Extraktionsgenauigkeit weiter gesteigert werden kann [20, S. 1-4]. In der Medizin findet PLM-basierte Datenextraktion aus Freitexten bereits Anwendung und weist hohe Genauigkeiten auf [10, S. 6-9], [21, S. 5-7]. Durch ihre hohe Genauigkeit der Datenextraktion in vergleichbaren Domänen werden in dieser Arbeit PLMs eingesetzt [22, S. 11-14].

Im Rahmen des PoC wird ein Experiment durchgeführt, um verschiedene Modellierungsansätze zu testen und deren Auswirkung auf die Datenqualität zu messen und zu validieren [23, S. 1-3]. Hierzu werden vorab auf Literatur basierte Hypothesen getroffen und diese anhand von Evaluationsmetriken als „angenommen“,

„teilweise angenommen“ und „abgelehnt“ klassifiziert. Durch gezielte Variation von LLMs und Prompting-Techniken können die Auswirkungen auf die Qualität der Datenstrukturierung systematisch untersucht werden und die für den Anwendungsfall besten Konfigurationen ausgewählt werden [24, S. 1014-1016].

Zur Bewertung der Ergebnisse werden die Metriken Precision, Recall, F1-Score sowie Recall-Oriented Understudy for Gisting Evaluation (ROUGE) eingesetzt, da sie sowohl exakte als auch Teilübereinstimmungen mit einer Referenzquelle quantitativ erfassen [25, S. 6-12]. Die Auswahl dieser Metriken ergibt sich durch die Varianz der zu extrahierenden Felder, die Klassifikationen und Freitextanteile enthalten, wodurch eine ganzheitliche Evaluierung der Datenextraktionsqualität ermöglicht wird [26, S. 8], [27, S. 1-2]. Als Referenzquelle werden manuell extrahierte Daten auf Basis eines reduzierten Datensatzes verwendet. Sämtliche genannte Metriken sind in der Literatur etabliert und finden Anwendung bei LLM-gestützten Datenextraktionen in vergleichbaren Anwendungsbereichen [28, S. 5], [29, S. 2-3].

3. Grundlagen

Im Folgenden werden theoretische Grundlagen erläutert, die für ein tieferes Verständnis in den Folgekapiteln vorausgesetzt werden.

3.1. AI Regional Implementation Group

Die SAP AI Regional Implementation Group (RIG) ist eine Beratungsgruppe der SAP, die Kunden bei der Aktivierung ihrer SAP Business AI unterstützt. Ihre Aufgaben sind dabei die Analyse der Kundenlandschaft, die Identifizierung von AI-Szenarien und die Begleitung während der Aktivierung [5, S. 5]. Wie aus Abbildung 21 ersichtlich unterstützt die RIG in der „Discover“ bis hin zur „Realize“ Phase Kollegen aus dem Sales und dem Support.

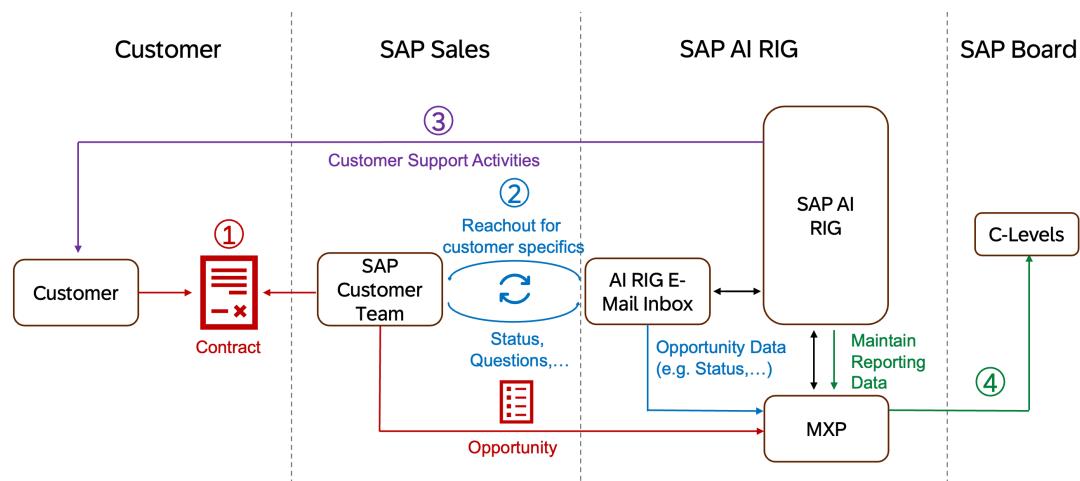


Abbildung 1: Abstrahierte Prozesse in der RIG AI in Anlehnung an [1]

Jeder Kunde, der Interesse am Zukauf von SAP Produkten hat, wird in SAP Systemen als „Opportunity“ abgebildet und von einem zugewiesenen Betreuer aus dem SAP Customer Team, meist einem Account Executive (AE), betreut. Eine Opportunity enthält Kundendaten sowie das Produkt, an dem der Kunde Interesse hat. Hat dieses AI-Bezug, wird über die Multi Experience Platform

(MXP) die Opportunity der RIG zur Verfügung gestellt. Der technische Prozess wird in Abschnitt 3.4.2 vertieft. [1, S. 9-14] [Abbildung 1, Prozess 1]

Sobald der Kunde den Kaufvertrag unterschrieben hat, tritt die RIG mittels einer Kundenmailkampagne mit dem zuständigen AE in Kontakt und ergänzt für die Verarbeitung notwendige Daten im RIG-MXP. Sobald diese in der Opportunity ergänzt wurden, wird der Kunde bei der Aktivierung der gekauften AI-Produkte unterstützt und der Prozess der Aktivierung im MXP aktuell gehalten. [5, S. 8] [Abbildung 1, Prozess 2-3]

Nach dem erfolgreichen Abschluss der Aktivierung werden fehlende Daten in einer Reporting-Anwendung, gehostet über die MXP, ergänzt und der Status der Opportunity auf „Activated“ gesetzt. Anschließend werden die aggregierten Daten aller abgeschlossener Opportunities dem Management zur Verfügung gestellt [1, S. 9-14] [Abbildung 1, Prozess 4].

3.2. Cross Industry Standard Process for Data Mining

CRISP-DM ist ein De-facto-Standard Vorgehensmodell, um Data-Science- und Data-Mining-Projekte strukturiert zu planen und durchzuführen [16, S. 529-532], [30, S. 2]. Der CRISP-DM-Prozess ist industrie- und technologieunabhängig konzipiert und in sechs Phasen unterteilt, die von der geschäftlichen Zieldefinition bis zur finalen Implementierung reichen [31, S. 9-10]. Das Vorgehensmodell gliedert sich in die in Abbildung 2 dargestellten Phasen und wird im Folgenden erläutert:

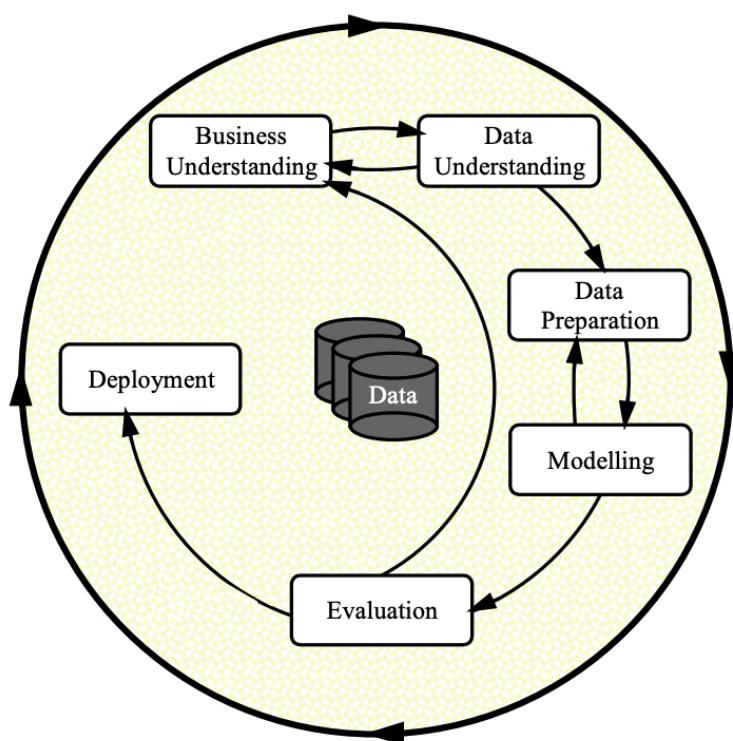


Abbildung 2: Phasen des CRISP-DM Prozess [2]

- **Business Understanding:** In der ersten Phase werden Ziele aus Unternehmensperspektive definiert und die Anforderungen an das Data-Science-Projekt präzisiert [31, S. 16-19]. Dazu werden ein oder mehrere Erfolgskriterien zur Bewertung der Zielerfüllung ausgewählt [16, S. 527].

- **Data Understanding:** In dieser Phase werden zur Verfügung stehende Datenquellen identifiziert, Daten gesammelt und erste Analysen durchgeführt, um einen Überblick über Datenqualität, Datenumfang sowie Hinweise zu Datenanalysemethoden zu erlangen. [31, S. 20-22]
- **Data Preparation:** In Phase Drei wird aus dem rohen Datensatz ein zur Modellierung geeigneter Datensatz geschaffen. Dazu gehört die Selektierung, Säuberung, Ergänzung von Attributen und Einträgen, Transformation und Konvertierung in ein geeignetes Format. [31, S. 23-26]
- **Modelling:** In der vierten Phase werden für den Geschäftskontext geeignete Modellierungstechniken ausgewählt, angewendet und für die Evaluation der Ergebnisse relevante Metriken bestimmt. [31, S. 27-29]
- **Evaluation:** In dieser Phase wird das Modelling und die Ergebnisse anhand der Evaluationsmetriken bewertet und evaluiert, inwiefern die in Business Understanding definierten Ziele erreicht sind. [31, S. 30-31].
- **Deployment:** Nach der erfolgreichen Erfüllung aller Ziele erfolgt die Planung und Durchführung der Produktivnahme (Deployment), Projektabschluss und Dokumentation. Hierbei werden die Verantwortungen nach Ende des Projektes hinsichtlich Wartung und Support festgehalten. [31, S. 32-34]

Die Verwendung von CRISP-DM bringt einige Vorteile mit sich. CRISP-DM erlaubt eine Iteration in vorherige Phasen [Abbildung 2], beispielsweise durch von den Geschäftszielen abweichende Datengrundlage [32, S. 9], Notwendigkeit weitere Datenvorverarbeitung oder Nichterfüllung der Geschäftsziele in der Evaluation [33, Abs. 2.1]. Erkenntnisse aus den vorherigen Iterationen fließen in den Prozess mit ein und verbessern die Datenqualität und Modellleistung nachhaltig [31, S. 9-10].

Zudem bietet CRISP-DM eine Struktur, die sich sowohl zur alleinigen Arbeit als auch in einem Team eignet. Durch ihren Status als De-facto-Standard Vorgehensmodell für Data Mining Projekte ist sie vielen Data Scientists bekannt [16, S.] und unterstützt diese dabei, Kommunikationsaufwand zu reduzieren und Verantwortlichkeiten zu klären [16, S. 529]. Die frühzeitige Einbindung von Geschäftszielen stellt sicher, dass das Projekt relevant ist.

Allerdings birgt CRISP-DM Risiken bei der Anwendung. Die zuvor beschriebene Iterativität wird in der Praxis oft nicht gelebt und Phasen werden linear durchlaufen [2, S. 7-8], [34, S. 32]. Gründe dafür sind fehlende Richtlinien, wann eine Iteration in eine vorherige Phase notwendig ist [34, S. 32-33]. Zudem setzt CRISP-DM keinen Fokus auf Projektmanagement, wodurch es in der Praxis zu Abweichung von Phasen- und Zielen kommen kann und Fristen nicht eingehalten werden können [2, S. 7-8], [34, S. 32-33].

CRISP-DM setzt zudem einen klaren Fokus auf Datenanalyse und Modellentwicklung, wobei nur eine geringe Anzahl der Modelle am Ende des CRISP-DM-Zyklus produktiv entwickelt werden. Gründe dafür sind fehlende Standards zu Wartung & Kontrolle nach Evaluation des Modells. Viele der Modelle werden im Anschluss mittels einer anderen Methodik produktiv implementiert. [16, S. 532-533]

In dieser Arbeit wird CRISP-DM als methodischer Rahmen verwendet, da der Fokus auf Datenanalyse und Modellentwicklung den Anforderungen des Projektes entspricht und Projektmanagementaspekte aufgrund der Einzelentwicklung des PoC zweitrangig sind. Zudem ist die Methodik als De-facto-Standard im Team der AI RIG etabliert, wodurch eine spätere Übergabe zur Produktivnahme des PoC effizient erfolgen kann, wobei diesen das Risiko eines fehlenden Projektmanagement bekannt ist.

3.3. Large Language Models

Large Language Models (LLMs) sind generative Sprachmodelle, die aus einer Modelleingabe (Prompt) anhand von Wahrscheinlichkeiten eine natürliche Ausgabe produzieren [35, S. 2]. Sie entspringen der Domäne des NLP und basieren in der Regel auf neuronalen Netzen, die auf großen Datensätzen trainiert werden [36, S. 3]. Ein neuronales Netz verbindet Eingabe- und Ausgabedaten über künstliche Neuronen, welche die Aufgabe haben, aus mehreren Eingaben, anhand von Wahrscheinlichkeiten (Bias) aus vorherigen Iterationen, eine Ausgabe zu berechnen. Die Ausgaben werden anschließend von einer Softmax-Funktion, welcher einer Eingabe eine Ausgabewahrscheinlichkeit zuweist [37, S. 1], bewertet und die Antwort (Prediction) mit der höchsten Wahrscheinlichkeit ausgibt. [38, S. 442-443].

Eine Vielzahl von LLMs basieren auf Transformer-Architektur, die mithilfe von mehrschichtigen neuronalen Netzen eine Ausgabe auf Basis des Encoder-Decoder-Prinzip produzieren [9, S. 1-6], [39, S. 2]. Das Encoder-Decoder-Prinzip trennt die Architektur eines Modells in Encoder und Decoder. Der Encoder wandelt eine Wortsequenz in eine Vektorrepräsentation (Token) um, während der Decoder die durch die Softmax-Funktion wahrscheinlichsten Token transformiert und in natürlicher Sprache ausgibt. [40, S. 4-5], [41, S. 31-32]. Zur Generierung des Ergebnisses werden Relationen zwischen Tokens mittels Selbstaufmerksamkeit (Self-attention) extrahiert [41, S. 33]. Self-attention bestimmt für jeden Token einen relativen Query-, Key- und Value-Vektor, welche verschiedene Anteile des Tokens repräsentieren [9, S. 6-7]. [40, S. 6] Mit Hilfe dieser Vektoren können auf der Basis von Vektorenaddition und Vektorenabständen kontextuelle Beziehungen zwischen allen Token der Eingabesequenz, sowie gespeicherten Parametertokens abgebildet und auf Basis dieser begründete Voraussagen ge-

troffen werden, welche hinsichtlich Genauigkeit die von regelbasierten Verfahren in diversen Domänen übersteigen [41, S. 34], [42, S. 15572-15573].

LLMs grenzen sich gegenüber anderen Sprachmodellen besonders durch ihre Modellgröße und umfangreiches Training ab, wodurch sie eine hohe Leistung in diversen Anwendungen aufweisen [35, S. 2-3]. Es wird differenziert zwischen LLMs und PLMs. PLMs bilden eine Untergruppe von LLMs und zeichnen sich durch ein bereits erfolgtes Training des Modells aus, wobei andere Modelle vor der Nutzung trainiert werden müssen, um korrekte Ergebnisse zu produzieren. [43, Abs. 1]. In der Literatur werden diese Begriffe häufig synonym genutzt, weswegen in dieser Arbeit die allgemeine Bezeichnung LLM verwendet wird [43, Abs. 1]. [44, S. 1-3].

3.3.1. Vorteile von Large Language Models

LLMs haben in den letzten Jahren erhebliche Fortschritte in der natürlichen Sprachverarbeitung erzielt und bieten eine Vielzahl von Vorteilen in diversen Domänen. [11]

Im Gegensatz zu traditionellen Modellen zeichnen sich LLMs durch ihr natürliches Sprachverständnis aus [45, S. 1-2]. Aufgrund ihrer Transformer-Architektur ist es LLMs möglich, die Relationen zwischen Sätzen aufzufassen, wodurch sie in der Lage sind, komplexe Satzstrukturen zu produzieren und eine korrekte Ausgabe zu produzieren [46, S. 10-11].

LLMs zeichnen sich durch ihre Adaptierbarkeit aus. Durch gezielte Eingaben lassen sich LLMs mit wenigen oder ohne Beispieldaten flexibel auf Anwendungsfälle einstellen [47, S. 1]. Im Gegensatz zu regelbasierten NLP-Verfahren können

LLMs Daten zum kontinuierlichen Lernen verwenden, wodurch ihre Antwortqualität kontinuierlich steigt [48, S. 3-4].

Ein weiterer Vorteil ist die Genauigkeit von LLMs. In Extraktions- und Evaluationsaufgaben übertreffen LLMs regelbasierte Verfahren sowohl in Genauigkeit als auch Fehlerrate. [39, S. 5-8], [49, S. 8]. Im medizinischen Umfeld werden Modelle bereits seit langem eingesetzt und übertreffen Mediziner in diversen Domänen, beispielsweise in der Krebserkennung oder der Datenextraktion [50, S. 4-8], [51, S. 8-15].

Ebenso erlauben LLMs die Automatisierung diverser Prozesse, speziell in Unternehmen und der Forschung in Aufgaben wie Textgenerierung oder Zusammenfassung von Informationen [52, S. 1-4].

3.3.2. Risiken und Herausforderungen

Dennoch birgt der Einsatz von LLMs Risiken bei ihrer Verwendung gegenüber regelbasierten Verfahren. Ein großes Risiko bilden sogenannte „Halluzinationen“. Als Halluzination wird die Generierung von Inhalten, die faktisch falsch sind und keine Grundlage in den Trainingsdaten haben [53, S. 4], [54, S. 1-3]. Diese können durch diverse Faktoren auftreten. Darunter gehören Fehlberechnungen statistischer Annahmen, unvollständiges Modelltraining oder die Hinzugabe von zu geringem Kontext in das Modell [53, S. 5]. Sie können mit diversen Methoden minimiert, aber nicht vollständig vermieden werden [54, S. 11-12]

Einschränkungen treten bei LLMs hinsichtlich Kosten, Zeit und Daten auf [55, S. 1]. Lokale LLMs mit Milliarden Parametern erfordern enorme Rechenleistung und spezialisierte Hardware für deren Betrieb, zuzüglich Trainings- und Wartungskosten. Ein lokales Deployment wird bei vielen Unternehmen aus Kostengründen

ausgeschlossen. Aus diesem Grund bieten viele Modellhersteller eine Nutzung des Modells über Application Programming Interfaces (APIs)-Anfragen an, bei der das Modell vom Anbieter gehostet wird und Kosten auf Basis der Nutzung kalkuliert werden [55, S. 1-3], [56]. Durch die Verwendung von API-Schnittstellen zwischen Modellanbieter und Anwender kommt es zu Latenzen zwischen Eingabe und Modellausgabe [57, S. 15-16]. [58, S. 2].

Die Verwendung von LLMs bringt Datenschutzrisiken mit sich. Sensitive Daten wie Namen, Adressen o.ä. werden bei Eingabe in ein LLM zu Trainingszwecken beim LLM-Anbieter verwendet [59, S. 5167], wodurch dem LLM-Nutzer bei fehlender Zustimmung durch Herausgabe an Externe Schaden drohen kann [60].

3.3.3. Large Language Models in der Datenextraktion

LLMs haben in der automatisierten Datenextraktion für Wirtschaft und Forschung deutliche Fortschritte ermöglicht. Im Vergleich zu klassischen NLP-Verfahren erreichen LLMs häufig eine höhere Genauigkeit bei der Extraktion strukturierter Informationen aus unstrukturierten Textdaten [39, S. 5-8], [61, S. 4-5].

In der Medizin werden bereits LLMs zur Extraktion von Daten aus unstrukturierten Daten genutzt [19, S. 5167-5174], [26, S. 7-12], [62, S. 4-7] und übertreffen bei der Extraktion klinischer Informationen aus Freitext klassische NLP-Verfahren [39, S. 5-8]. In der Wirtschaft steigern LLMs in einer Studie zur Extraktion von Finanzdokumenten die Leistung gegenüber manuellen Verfahren um bis zu 29% [61, S. 4].

Dennoch bringt der Einsatz von LLMs in der Datenextraktion auch Herausforderungen mit sich. Das Problem der Halluzinationen von LLMs tritt ebenfalls in Extraktionsaufgaben auf, etwa bei der Extraktion nicht vorhandene Beziehungen zwischen Entitäten. Dadurch wird es in der Extraktion sensibler Geschäftsdaten

als zentrales Problem angesehen. [61, S. 7-11], [63, S. 3]. Darüber hinaus ist die Verwendung von LLMs gegenüber regelbasierten NLP-Extraktionsverfahren mit erhöhtem Zeit- und Ressourcenaufwand verbunden, weswegen sie im geschäftlichen Kontext oft nicht in Betracht gezogen werden [55, S. 1], [57, S. 15-16]

3.4. Services & Tools

In dieser Arbeit werden diverse Tools und Services zur Realisierung genutzt, die im Folgenden weiter erläutert werden.

3.4.1. SAP Generative AI Hub

Der Generative Artificial Intelligence (GenAI) Hub ist eine zentrale Plattform der SAP, die den Zugang zu generativen AI-Modellen über den SAP AI Core ermöglicht [64]. Der SAP AI Core Service fungiert als zentrale AI Laufzeit-Umgebung innerhalb der SAP Business Technology Platform, die zugrunde liegende Cloud-Plattform zur Entwicklung, Integration und Erweiterung von SAP Anwendungen. [65, S. 59-61]. Der SAP GenAI Hub bietet mittels AI scenarios eine einheitliche Schnittstelle für diverse AI-Modellanbieter, wodurch ein einfaches und ressourcenoptimiertes Wechseln zwischen Modellen möglich ist [64], [65, S. 59-60]. Innerhalb der SAP fallen für die Verwendung des GenAI Hub keine Kosten an, Kunden zahlen Lizenzkosten zur Nutzung der Plattform.

Zur Realisierung einer einheitlichen Schnittstelle bietet der SAP GenAI Hub den Orchestration Service unter dem AI Scenario „orchestration“ an. Dieser erlaubt das Nutzen von ausgewählten AI-Modellen diverser Anbieter mithilfe einer einheitlichen API-Schnittstelle. [3].



Abbildung 3: Orchestration Service des SAP GenAI Hub [3]

Der Orchestration Service bietet die in Abbildung 3 dargestellten Funktionen, welche in der Anwendung im Geschäftskontext wichtig sind. Prompt Templating erlaubt das dynamische Einfügen von Attributen in Prompts und Data Masking ermöglicht die Kodierung von vordefinierten Daten bei der Eingabe in ein Modell. [65, S. 97]. Die Modellausgabe kann nach Verarbeitung wieder encodiert werden. Darüber hinaus stellt der GenAI Hub eine einheitliche Implementierung für strukturierte Modellausgaben (bspw. als JavaScript Object Notation (JSON)-Format) unter der Funktion „Structured Output“ zur Verfügung, welche die Ausgabe in einer vordefinierten Datenstruktur garantiert, die zur Weiterverarbeitung der Daten notwendig ist [66], [67].

Alle im GenAI Hub verfügbaren Modelle entsprechen den Datenschutz- und Sicherheitsanforderungen der SAP SE [68]. Im Laufe des Jahres 2025 sollen weitere Modelle zum SAP GenAI Hub hinzugefügt werden und weitere Funktionen wie Hyperparametertuning ermöglicht werden [69]. Damit bietet der SAP GenAI Hub eine solide Basis für über diese Arbeit hinausgehende Forschung.

3.4.2. SAP Multi Experience Platform

Die SAP Multi Experience Platform (MXP) ist eine Plattform, die Datenverwaltung mit einer Low-Code basierten Entwicklung von Applikationen verbindet [4, Abs. Introduction]. Sie stellt eine Erweiterung zur SAP Analytics Cloud dar und erlaubt eine Integration mehrerer Datenquellen, deren Anreicherung mit weiteren Daten und anschließende Darstellung der Daten in Form einer Webanwendung [4, Abs. Introduction], [70]. Im Gegensatz zu anderer Anwendung, wie der SAP Datasphere, fokussiert sich die MXP auf die Low-Code basierte Entwicklung von Webanwendungen zur Verwaltung und Manipulation aggregierter Daten, wobei eine Analyse der aggregierten Daten zweitrangig ist [71, S. 2-4]. MXP Applikationen eignen sich für datenintensive Anwendungen, beispielsweise zur Visualisierung von Kennzahlen oder zum Reporting [4, Abs. Experience Building & Designing].

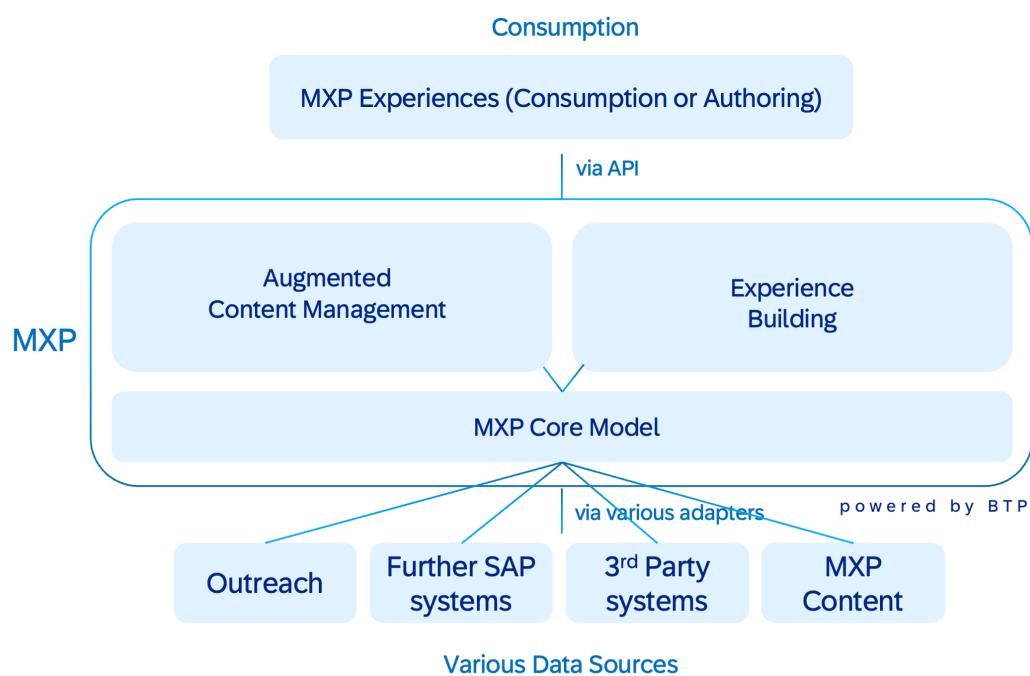


Abbildung 4: Architektur einer MXP Applikation [4]

Die MXP nutzt Augmented Content Management Systems (aCMSs), um Aufwand bei der Integration mehrerer Datenquellen zur Applikationsentwicklung zu mindern. Das aCMS ist ein Datenverwaltungssystem, welches mittels einer

Zwischenschicht, dem MXP Core Model, Daten aus mehreren Pipelines zusammensetzt und die Anreicherung mit zusätzlichen Daten ermöglicht [Abbildung 4]. Diese angereicherten Datensets werden als „Entity“ in einer „Worksphere“, anwendungsspezifischen Datastores, gespeichert. Auf Entities kann mittels einer API zugegriffen werden, welche es erlaubt, die Daten auszulesen und zu bearbeiten. [4, Abs. Content Management]

Eine Experience bildet ein interaktives User Interface einer MXP Applikation, welches als Webanwendung zur Laufzeit Daten aus dem aCMS konsumiert. Sie ruft die Daten einer oder mehrerer Entities per API-Call auf und ermöglicht Nutzern die Manipulation der Daten aus dem aCMS [Abbildung 4].

3.5. Datenextraktionsoptimierung

Das Ziel der Datenextraktionsoptimierung ist die Verbesserung der Datenextraktionsergebnisse hinsichtlich gesetzter Evaluationsmetriken gegenüber einer Referenzlösung [13, S. 279-281]. In anderen Domänen eingesetzte Verfahren umfassen die Anpassung der LLM-Eingabe anhand von Prompt Engineering sowie der Einsatz mehrerer Modelle, welche im Folgenden näher betrachtet werden [7, S. 1-2], [19, S. 5165-5166].

3.5.1. Modellwahl

Die Wahl des LLMs ist maßgeblich für die Qualität der Datenextraktion [19, S. 5165-5166]. LLMs unterscheiden sich in Kriterien wie Parametergröße und Robustheit gegenüber Noise voneinander, wodurch ihre Performance stark an den Anwendungsfall gekoppelt ist [72, S. 39:6-39:12]. Dabei beschreibt Noise fehlerhafte, irrelevante oder zufällige Datenanteile, die analytische Auswertun-

gen verfälschen und daher bereinigt werden müssen [73, S. 1756]. Anhand diverser Kriterien wird eine Vorentscheidung für einsetzbare LLMs getroffen, die anschließend anhand von Performance-Metriken auf einem Testdatensatz evaluiert werden, um das beste LLM für den Anwendungszweck zu ermitteln [19, S. 5167-5171]. Diese Kriterien umfassen:

- **Modellgröße:** Modelle mit geringem Parameterumfang sind meist schneller und günstiger als große Modelle, wobei große Modelle komplexe Aufgaben ohne Vortraining in diversen Anwendungsbereichen besser absolvieren können [74, S. 12-14].
- **Robustheit:** Robuste Modelle sind in der Datenextraktion performanter bei erhöhter Komplexität der Anfrage, wobei sie meist durch die Anwendung von schlussgefolgertem Denken (Reasoning) ressourcenintensiv arbeiten [75, S. 4-6].
- **Verfügbarkeit:** Durch externe Faktoren wie Unternehmensrichtlinien, Lizzenzen oder zu verwendende Tools (bspw. OpenAIs Structured -Outputs [67]) ist die Wahl des Modells eingeschränkt, wodurch nur eine limitierte Auswahl an Modellen zur Verfügung steht [76, S. 1].

Anhand dieser oder weiterer anforderungsspezifischer Kriterien werden Modelle ausgewählt, die sich hinsichtlich der definierten Kriterien zueinander unterscheiden, um den Einfluss jedes Kriteriums auf die Qualität der Modellausgabe zu evaluieren und eine fundierte Modellwahl zu ermöglichen [19, S. 5167-5171]

3.5.2. Prompt Engineering

Prompt Engineering bezeichnet die gezielte Formulierung von Prompts, um die Qualität von LLM-Ausgaben zu steigern [77, S. 1]. Dies kann durch die Anwendung von Prompting-Templates und -Techniken umgesetzt werden [77, S. 2-10]. Prompt Engineering zeichnet sich durch seine simplifizierte Umsetzbarkeit und Effektivität bei der Optimierung von PLMs-Outputs aus [77, S. 1], [78, S. 7-11].

Ein Prompt enthält Textsegmente wie Aufgabenstellungen, Kontext oder Beispiele, die das LLM zur Generierung einer Antwort benötigt [79, S. 3-4]. Gute Prompts weisen eine präzise Zieldefinition, eine klare Struktur und kontextbezogene Informationen auf [80, S. 1-2]. Um die Gewichtung dieser durch das Modell zu kontrollieren, kann einem Textsegment eine Rolle zugewiesen werden [81]. Hierbei wird zwischen folgenden Rollen unterschieden:

- **System Message:** Spezifiziert die Rahmenbedingungen oder Verhaltensregeln des Modells und wird von Modellen stärker gewichtet als die User Message [82].
- **User Message:** Enthält die zu bewältigende Aufgabe [81].

Prompt Engineering nutzt diese Eigenschaften aus, um dem Modell gezielte Vorgaben zu geben, wie eine Aufgabe zu bewältigen ist [83]. Im Folgenden werden Prompting-Techniken vorgestellt, welche ihn vergleichbaren Anwendungsfällen diverser Domänen zu einer Steigerung der Modellausgabequalität geführt haben [47, S. 20-21], [77, S. 2-9]:

- **Zero-Shot-Prompting:** Das LLM erhält ausschließlich eine Aufgabe, ohne die Hinzugabe weiterer Daten [77, S. 3]. Dies erlaubt es, LLMs unmittelbar einzusetzen, bedingt jedoch oftmals eine geringere Leistung bei komplexen Fragestellungen [78, S. 7-11].

- **One-/Few-Shot-Prompting:** Das LLM erhält eine Aufgabe mit Hinzugabe von einem oder mehreren Beispiele, bestehend aus Aufgabenstellung, Kontext und idealer LLM-Ausgabe. One/Few-Shot-Prompting übertrifft die Genauigkeit von Zero-Shot-Prompting in mehreren Bereichen, benötigt aber mehr Eingabetokens sowie vordefinierte Beispiele mit manuell validierten Ergebnissen [77, S. 2]. Mit steigender Beispielanzahl wächst die Genauigkeit der Extraktion [47, S. 20-21], wobei der Zuwachs der Genauigkeit ab vier bis acht Beispielen abflacht und nur geringfügig eine Verbesserung erzielt [84]
- **Chain-of-Thought (CoT)-Prompting:** CoT-Prompting fordert das Modell zur Step-by-Step Lösung der Aufgabe auf. Durch den größeren Kontext liefert CoT-Prompting bei komplexen Aufgaben bessere LLM-Ausgaben, erfordert aber erhöhten manuellen Prompting-Aufwand durch exakte Kontextbeschreibung [79, S. 7].[77, S. 2-7]
- **Self-consistency-Prompting:** Ähnlich zu CoT-Prompting erhält das LLM neben der Aufgabe die Aufforderung, eigene Zwischenergebnisse zu produzieren und diese kritisch zu hinterfragen [79, S. 8-9]. Dadurch validiert sich das Modell selbst, wodurch Halluzinationen reduziert und die Qualität der Modellausgabe gegenüber anderen Prompting-Techniken in Datenextraktionsaufgaben verbessert werden kann [85, S. 5-8].

Die praktische Umsetzung von vergangenen Anwendungsfällen zeigt, dass durch die Kombination mehrerer Prompting-Techniken, die Qualität der Modellausgabe erheblich gesteigert werden kann [77, S. 2], [79, S. 14-23]. Ebenso übertreffen LLMs generierte Prompts manuell erstellte Modelleingaben in Leistung, weswegen die in dieser Arbeit eingesetzten Prompts mithilfe von AI verbessert werden [86, S. 8-10]. Inwiefern die Datenextraktion mittels Prompt Engineering optimiert werden kann, wird in Abschnitt 4.4 untersucht.

3.6. Evaluationsmetriken

Um den Erfolg der LLM-gestützten Datenextraktion messbar zu machen, werden geeignete Evaluationsmetriken bestimmt. Diese spalten sich auf in Klassifikationsmetriken und Token-Similarity-Metriken. [25, S. 6-7]

Bei der Klassifizierung werden Daten in eine oder mehrere Klassen zugeordnet und anhand einer Referenzklassifizierung evaluiert [25, S. 6]. Hierbei wird zwischen Einfach- und Mehrfach-Klassifizierung unterscheiden. Bei Einfach-Klassifizierung wird einem Datensatz eine Klasse zugeordnet, bei Mehrfach-Klassifizierung können einem Datensatz mehrere Klassen zugeordnet werden.

Token-Similarity Metriken hingegen evaluieren Wortsequenzen. Dafür werden Metriken wie Recall-Oriented Understudy for Gisting Evaluation (ROUGE) eingesetzt, die eine Wortsequenz gegenüber einer Referenzquelle evaluiert und anhand der Überschneidung einen Wert bestimmt [27, S. 1-2].

Als Referenzquellen werden Quellen verwendet, zu denen bereits eine ideale Extraktion besteht. Zur Schaffung einer Evaluationsgrundlage werden i.d.R. Daten manuell evaluiert [87, Abs. 1].

3.6.1. Precision, Recall und F1-Score

Precision, Recall und F1-Score sind drei weit verbreitete Metriken zur Evaluation von Klassifikationen durch ein Modell gegenüber einer Referenzquelle [28, S. 5]. Sie geben Aufschluss, inwiefern eine Modellklassifikation der Referenzklassifikation entspricht [10, S. 7-9], [21, S. 6], [26, S. 8].

Sei c eine Klasse, L die Menge an Klassen $\{c_1, c_2, \dots, c_L\}$ inklusive der Null-Klasse c_\emptyset bei fehlender Label-Zuweisung, $|L|$ die Kardinalität der Menge L , $P \subseteq$

L die Menge an vorhergesagten Klassen und $H \subseteq L$ die Menge an tatsächlichen Klassen. Ist $|L| > 1$ und einer Extraktion kann genau eine Klasse zugeordnet werden, so spricht man von einem Multi-Label-Klassifikation mit eindeutiger Label-Zuordnung. Dafür wird der One-vs-All Ansatz verwendet, bei dem je Klasse $c \in L$ eine binäre Klassifikation vorgenommen wird, indem alle Exemplare der Klasse c als positiv und alle übrigen als negativ betrachtet werden, wodurch das ursprüngliche Mehrklassenproblem auf $|L|$ unabhängige Binäraufgaben reduziert wird [88, S. 1-2]. Dies wird für P und H durchgeführt und die Menge der überschneidenden Klassen den in Tabelle 1 dargestellten Mengen je Klasse zugewiesen. Wird einem Eintrag sowohl in P als auch in H der Klasse c_0 zugeordnet, wird dieser Eintrag als korrekt klassifiziert identifiziert und der Menge TP zugeordnet. [25, S. 7-9]

| | Tatsächlich positiv | Tatsächlich negativ |
|---------------------------|---------------------|---------------------|
| Als positiv klassifiziert | True Positive (TP) | True Negative (TN) |
| Als negativ klassifiziert | False Positive (FP) | False Negative (FN) |

Tabelle 1: Confusion Matrix für eindimensionale Klassifizierung je Klasse [6, S. 3]

Precision gibt an, welcher Anteil der als positiv identifizierten Ergebnisse tatsächlich positiv ist. Sie misst die Fähigkeit eines LLMs, negative Instanzen zu filtern. Die Formel der Precision ist im Folgenden dargestellt. Hierbei entspricht $|\text{Menge}|$ der Kardinalität einer Menge: [25, S. 8]

$$\text{Precision} = \frac{\text{korrekt als positiv klassifizierte Einträge}}{\text{als positiv klassifizierte Einträge}} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \quad (1)$$

Recall, auch bekannt als „True Positive Rate“ oder „Sensitivity“, beschreibt, wie vollständig relevante Ergebnisse erkannt wurden. Sie bestimmt, inwiefern ein Modell positive Ergebnisse identifiziert. Die Formel für Recall lautet: [25, S. 7-8]

$$\text{Recall} = \frac{\text{korrekt als positiv klassifizierte Einträge}}{\text{tatsächlich positive Einträge}} = \frac{|TP|}{|TP| + |FN|} \quad (2)$$

Der F1-Score bildet das harmonische Mittel zwischen Precision und Recall. Die Formel für den F1-Score lautet: [25, S. 8]

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot |TP|}{2 \cdot |TP| + |FP| + |FN|} \quad (3)$$

3.6.2. Recall-Oriented Understudy for Gisting Evaluation

ROUGE ist eine Evaluationsmetrik des NLP, welche zur Evaluation von generierten Texten gegenüber einer Referenzquelle angewendet wird [25, S. 11-12]. Sie quantifiziert die Ähnlichkeit zweier Texte in einer Metrik und ermöglicht gegenüber Klassifikationsmetriken die Evaluation von Auswirkungen der Modellierung auf Freitexte [25, S. 14]. ROUGE-n, eine Umsetzung von ROUGE, misst den Überlappungsgrad zwischen einem generierten Text zu einer Referenz mittels N-Grammen, welche Wortfolgen der Länge N repräsentieren [27, S. 1-2]. Formal entspricht ROUGE-n dem in Abschnitt 3.6.1 definierten F1-Score, wobei anstatt absoluter Zuordnung die Übereinstimmung von N-Grammen als True Positive herangezogen werden [25, S. 14]. Dadurch lassen sich die Metriken Precision, Recall und F1-Score zur Evaluation hinweg über verschiedene Arten an Extraktionsfeldern, aggregieren und berechnen [27, S. 1-2]

Sei ω ein Wort, $b = (\omega_1, \omega_2, \dots, \omega_{\Lambda_b})$ ein vorhergesagter Text der Länge Λ_b als Wortsequenz von ω_1 bis ω_{Λ_b} und $d = (\omega_1, \omega_2, \dots, \omega_{\Lambda_d})$ ein Referenztext der Länge Λ_d als Wortsequenz von ω_1 bis ω_{Λ_d} . Damit ist B die Menge aller n-gramme im vorhergesagten Text, D die Menge aller n-gramme im Referenztext und $D \cap B$ die Menge an überschneidenden n-Grammen, welche der Menge TP auf Basis von N-Grammen entsprechen. Die im generierten Text vorkommenden, aber

nicht in der Referenz enthaltenen N-Gramme $B \setminus D$ entsprechen der Menge FP, in der Referenz, aber nicht im generierten Text enthaltene N-Gramme $D \setminus B$ der Menge FN und weder in B noch in D vorkommende N-Gramme der Menge TN. Auf Basis dieser N-Gramm basierten Mengen lassen sich Precision_n , Recall_n sowie ROUGE_n als N-Gramm basierter F1-Score mit den folgenden Formeln berechnen: [25, S. 14]:

$$\text{Precision}_n = \frac{|D \cap B|}{|B|} \quad (4)$$

$$\text{Recall}_n = \frac{|D \cap B|}{|D|} \quad (5)$$

$$\text{ROUGE-n} = 2 \cdot \frac{\text{Precision}_n \cdot \text{Recall}_n}{\text{Precision}_n + \text{Recall}_n} \quad (6)$$

3.6.3. Accuracy

Accuracy, oft übersetzt als Genauigkeit, ist ein Maß zur Bewertung von Klassifikationsaufgaben. Sie misst den Anteil der korrekt klassifizierten Instanzen von allen Instanzen eines Datensatzes. Gegenüber F1 erweitert Accuracy die betrachteten Mengen um die Menge TN [25, S. 7]

$$\text{Accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} = \frac{|TP| + |TN|}{|L|} \quad (7)$$

Accuracy kann, analog zu Abschnitt 3.6.1, sowohl als Exact-Match berechnet werden, wobei nur vollständig identische Vorhersagen als korrekt identifiziert klassifiziert werden, als auch n-gramm-basiert, wie in Abschnitt 3.6.2, bei welcher jedes N-Gramm als Klassifikationseinheit betrachtet wird [25, S. 12-14]. Die zugehörigen Formeln ergeben sich aus den in den genannten Kapiteln dargestellten TP-, TN-, FP- und FN-Definitionen.

4. Praxis

Im Folgenden wird die praktische Umsetzung der Optimierung von strukturierter Datenextraktion beschrieben. Dabei werden die im Abbildung 2 dargestellten Phasen von CRISP-DM nacheinander durchlaufen.

4.1. Business Understanding

AI und ihre Aktivierung bei Kunden ist eines der Hauptziele der SAP SE für das Jahr 2025 [15]. Um diesen Fortschritt messbar zu machen, nutzt der Vorstand Reportingdaten aus der RIG. Diese werden aus den Opportunities des MXP-Tools extrahiert und enthalten Erfolgs-/Misserfolgsdaten je Kunde/Produkt. Dafür ist es notwendig, dass die E-Mails der RIG-Inbox bearbeitet und die Daten je Opportunity von der RIG gepflegt werden.

Mit der Erstellung einer Opportunity bei Kaufinteresse eines Kunden wird diese mit Daten aus darauffolgenden Kundeninteraktionen mittels aCMS der MXP von Team zu Team ergänzt [4, Abs. Content Management]. Der Status Quo ist, dass diese Daten manuell über eine Experience der RIG-MXP je Opportunity eingepflegt werden [4, Abs. Model Content]. Diese Arbeit wird durch das Backoffice-Subteam der RIG durchgeführt, deren Aufgabe es ist, anhand von Daten aus der RIG-E-Mail-Inbox Reportingdaten in die zugehörige Opportunity auf der MXP einzupflegen und Kundenmailkampagnen an ausstehende Opportunities zu senden. Der Vorstand nutzt die aggregierten Daten aller Opportunities der Periode, um fundierte Entscheidungen hinsichtlich der SAP AI-Strategie zu fällen [89].

Die RIG-E-Mail-Inbox ist die zentrale Anlaufstelle für die Kontaktaufnahme mit der RIG. Neben internen Fragen zu Business AI nutzt das SAP Kundenteam

die Inbox zur Koordinierung der AI-Aktivierung bei Kunden. Sobald eine E-Mail in der RIG-Inbox ankommt, ist das Ziel des Backoffice-Teams, die zugehörige Opportunity auf der MXP zu ermitteln. Um diese eindeutig zu identifizieren, wird eine Kombination aus den drei folgenden Identifiers (IDs) benötigt [89]:

- **CRM-Account-ID:** Die ID eines Kunden bzw. Customer Relationship Management (CRM)-Accounts im internen SAP System.
- **Opportunity-ID:** Die ID, welche eine Verkaufsmöglichkeit von Produkten gegenüber einem Kunden im CRM abbildet.
- **Material-ID:** Die im System hinterlegte Produktnummer des gekauften SAP Produktes. Sie bildet den CRM seitigen Vertragsbestandteil ab, weswegen mehrere Material-IDs für ein Produkt existieren.

Sind diese Daten in der E-Mail nicht gesetzt, wird versucht, über einen Sekundärschlüssel den zugehörigen MXP-Eintrag zu identifizieren. Zu diesen Feldern gehören:

- **Kundenname:** Der Name des vom AE betreuten Kunden.
- **Produktnname:** Der Name des verkauften Produktes.

Im Falle einer erfolgreichen Identifikation der Opportunity pflegt das Backoffice-Team folgende Daten ein:

- **Status:** Der aktuelle Status der Aktivierung. Dieser kann einen der in SAP RIG AI [1, S. 9-14] definierten Status annehmen.
- **Analysis:** Eine kurze Zusammenfassung der Arbeit der RIG mit Zeitstempel.
- **RIG-Kontakt:** Der Ansprechpartner für den AE aus der RIG. Dieser wird anhand der Interaktionshäufigkeit mit dem AE zugeordnet.
- **On-Hold-Date:** Das Datum, bis zu welchem eine Aktivierung beim Kunden temporär ausgesetzt ist. Wird gesetzt, wenn der Status auf ‚On Hold‘ gesetzt wird.

Das Hauptproblem des Backoffice-Teams ist eine Zuordnung der E-Mail zum zugehörigen MXP-Eintrag, da diese sowohl Duplikate als auch fehlende Einträge aufweisen. Anhand der in einer E-Mail ersichtlichen Daten ist oft eine eindeutige Identifikation der zugehörigen MXP-Opportunity nicht oder nur unter erhöhtem manuellen Suchaufwand möglich. Darüber hinaus muss die gesamte E-Mail mitsamt Verlauf gelesen werden, um die Felder auf der MXP zu setzen. Dieser Prozess wird für jede E-Mail der RIG-Inbox wiederholt und kostet das Backoffice-Team viel Zeit.

Das Ziel dieser Arbeit ist es, das Setzen von Reportingdaten in Teilen zu automatisieren und damit die RIG hinsichtlich der Bearbeitung der Inbox zu entlasten. Dafür soll ein Proof of Concept (PoC) entwickelt werden, welcher E-Mails verarbeiten und an die zugehörige MXP-Opportunity reporten kann. Notwendig dafür ist die Extraktion von Reporting- und Zuordnungsdaten in einer ausreichend hohen Qualität, dass eine Identifikation des zugehörigen MXP-Eintrags möglich ist. Dabei beschränkt sich der PoC auf die Identifikation eindeutiger Opportunities, welche keine Dopplungen auf der MXP aufweisen und somit eindeutig identifizierbar sind.

Zur Produktivnahme innerhalb der AI RIG hat diese in ihren Jahreszielen 2025 konkrete Key Performance Indicator (KPI) definiert, welche eine AI-gestützte Version erfüllen muss, um implementiert zu werden [1, S. 13]. Diese umfassen eine Extraktionsgenauigkeit über alle Felder von mindestens 70% sowie eine Identifikationsratenabweichung zur manuellen Zuordnung von maximal 10%. Darüber hinaus darf die falsch-positiv Rate an identifizierten MXP-Opportunities nicht 5% übersteigen, da das Setzen falscher MXP-Daten den Beratungsprozess der RIG direkt beeinflusst und langfristig der Kundenbindung schaden kann.

Um dieses Ziel zu erreichen, werden im Rahmen eines Experiments die in Abschnitt 3.5 definierten Ansätze der Extraktionsverbesserung angewendet und mittels der in Abschnitt 3.6 beschriebenen Metriken evaluiert. In dieser Arbeit wird keine formale Optimierung durchgeführt, sondern eine explorative Untersuchung zur Steigerung der Extraktionsergebnisqualität aufgrund zu geringer Datengrundlage [90, S. 1138-1139], [91, S. 6-9]. Diese bietet eine Grundlage für weitere empirische Forschung sowie Optimierungsverfahren.

Auf Basis von Literatur zu Datenextraktionsoptimierung diverser Domänen wurden folgende Hypothesen definiert, welche im Rahmen des Experiments auf die Domäne der E-Mail-Kommunikation validiert werden sollen. Die entsprechende Literatur ist jeder Hypothese angehängt:

- **Hypothese 1:** Der Einsatz von One-Shot-Prompting verbessert die Datenextraktionsgenauigkeit gegenüber Zero-Shot-Prompting [84].
- **Hypothese 2:** Few-Shot-Prompting führt zu höherer Datenextraktionsgenauigkeit gegenüber Zero- und One-Shot-Prompting [47, S. 20-21].
- **Hypothese 3:** Die Anwendung Chain-of-Thought-Prompting extrahiert qualitativ bessere Daten als ein Basisprompt [79, S. 7].
- **Hypothese 4:** Self-consistency-Prompting extrahiert Daten mit einer höheren Datenextraktionsgenauigkeit als ein Basisprompt [85, S. 5-8].
- **Hypothese 5:** Reasoning-Modelle wie OpenAIs „o1“ und „o3-mini“ weisen höhere Extraktionsgenauigkeit und geringere Halluzinationen auf gegenüber Non-Reasoning Modellen [75, S. 4-6].
- **Hypothese 6:** Modelle mit großem Parameterumfang weisen bessere Datenextraktionsergebnisse als kleine Modelle auf [74, S. 12-14].

Bei der Evaluation des PoCs werden die in Abschnitt 3.6 vorgestellten Metriken Precision, Recall, F1-Score, und Accuracy verwendet, welche auf Basis einer Re-

ferenzlösung gebildet werden. Das Extraktionsfeld „Analysis“ wird im Gegensatz zu den anderen Extraktionsfeldern mithilfe von ROUGE-n sowie einer N-Gramm basierten Accuracy evaluiert, da es sich hierbei um ein Freitextfeld handelt, welches im Gegensatz zu den anderen Feldern nicht exakt der Referenzlösung entsprechen muss.

4.2. Data Understanding

Zur Entwicklung des PoCs wird ein Auszug von zufällig gewählten E-Mails verwendet, welche mittels der Microsoft GraphAPI aus der Outlook-Inbox der RIG extrahiert wurden. Die Microsoft GraphAPI ist eine von Microsoft zur Verfügung gestellte Schnittstelle zu einem E-Mail-Postfach, welches die Abfrage von E-Mail-Attributen wie Betreff, Inhalt oder Absender ermöglicht [92]. Im Folgenden werden die Daten auf ihre Konsistenz und Relevanz für die Datenextraktion geprüft.

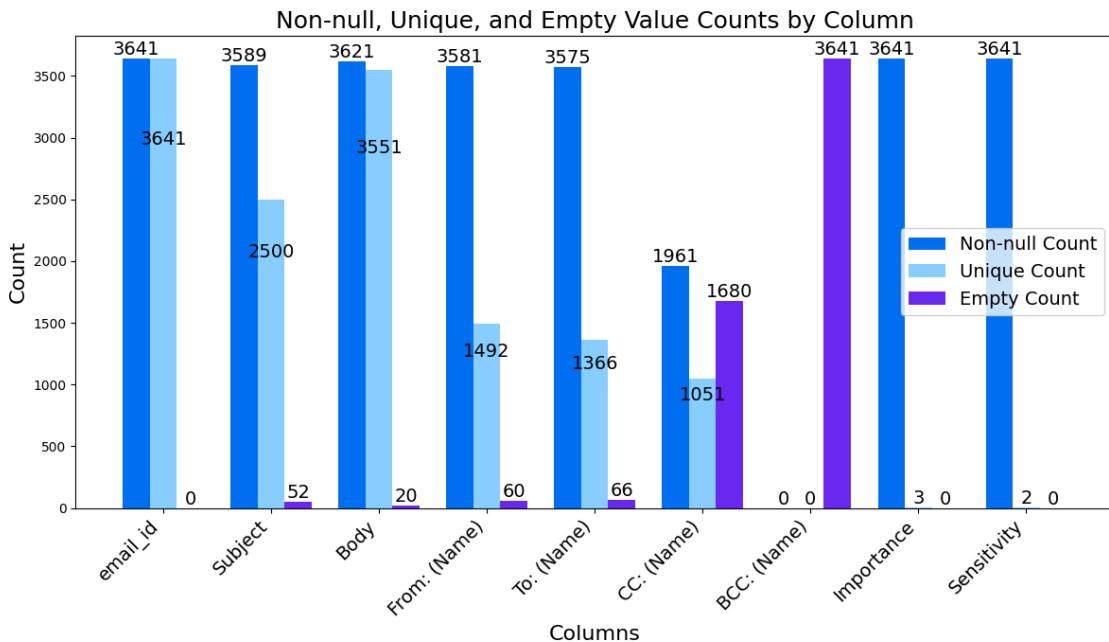


Abbildung 5: Attribute und Häufigkeit der Daten. Eigene Darstellung.

Der Datensatz wird durch die RIG zur Verfügung gestellt und enthält 3641 E-Mails abgeschlossener Kundenaktivierungen aus dem 4. Quartal des Jahres 2024. Eine

E-Mail wird mithilfe der in Abbildung 5 festgelegten Attribute dargestellt. Weitere Attribute der Microsoft GraphAPI werden aufgrund von fehlender Relevanz für den Anwendungsfall nicht in Betracht gezogen. Jede E-Mail bildet einen E-Mail-Verlauf ab, bestehend aus der zuletzt gesendeten E-Mail des Verlaufs sowie angehängter vorheriger E-Mails, getrennt durch Metadaten. Hinsichtlich einer Ermittlung des Status und Analyse.

Die Antworten des Kundenteams der SAP auf die Mailkampagnen der RIG sind wie aus Abbildung 6 ersichtlich meist kundenspezifisch, wobei in einigen E-Mails mehrere Kunden je E-Mail angesprochen werden. Um eine korrekte Identifikation der Opportunity zu garantieren, müssen diese in der weiteren Datenextraktion als alleinstehende mögliche Opportunities behandelt werden.

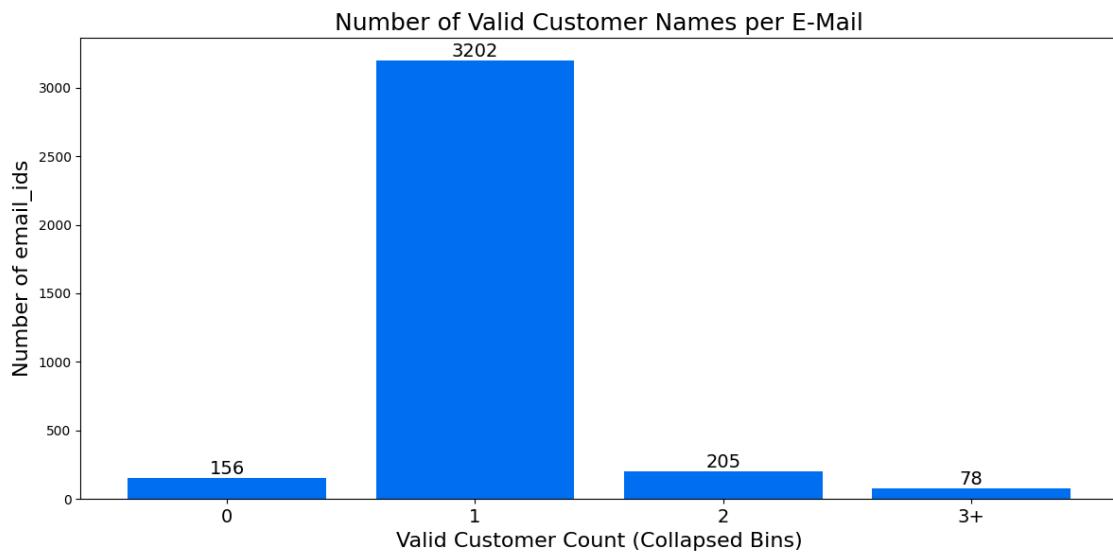


Abbildung 6: Anzahl der Kunden je E-Mail. Eigene Darstellung.

Anteilig sind in der RIG-Inbox 19,5% der E-Mails von der RIG gesendet bzw. angrenzenden Abteilungen und 1,56% von „Others“ [Abbildung 7]. Unter „Others“ fallen E-Mails von Kontakten, welche keine Relevanz für das Reporting der RIG haben, beispielsweise automatisch generierte Antworten von „Microsoft Outlook“. Somit sind 78,94% der gesendeten Daten von „Non-RIG“ Absendern.

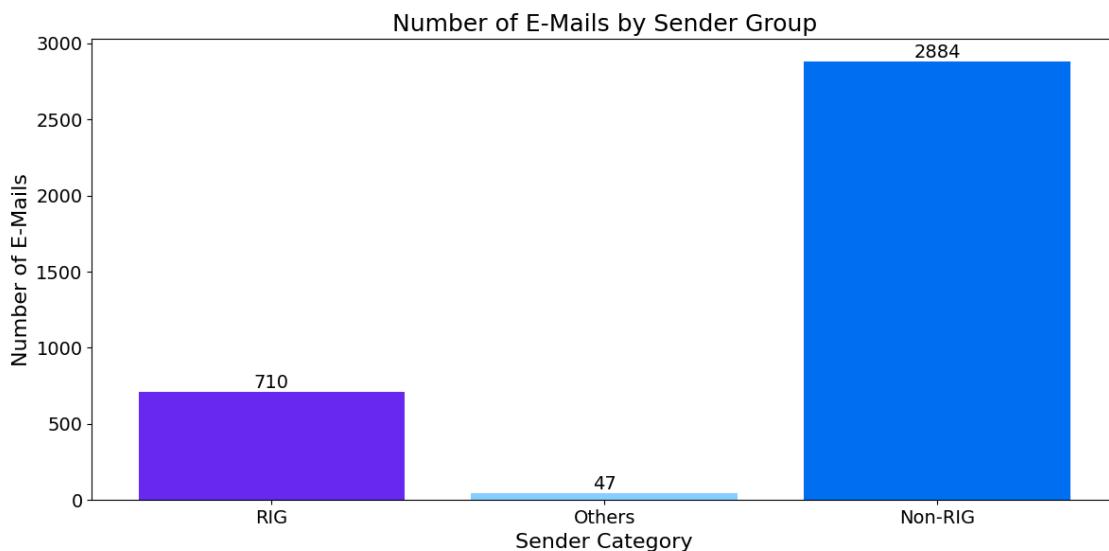


Abbildung 7: Anzahl an E-Mails je Sendergruppe. Eigene Darstellung.

Für die Identifikation der Opportunity sind die Felder ‚Subject‘, ‚Body‘ sowie ‚From: (Name)‘ notwendig, da diese die gesuchten Daten aus Abschnitt 4.1 enthalten:

- **Subject:** Der Betreff einer E-Mail. In den Mailkampagnen der RIG werden die AEs aufgefordert, den Kundennamen sowie den Produktnamen in den Betreff der E-Mail zu schreiben, wobei dies in der Realität nur bedingt umgesetzt wird und diese Daten oft in ‚Body‘ zu finden sind [Abbildung 8].
- **Body:** Der Inhalt eines E-Mail-Verlaufs, bestehend aus dem Inhalt der zuletzt versendeten E-Mail sowie E-Mail-Inhalte der vorherigen Konversation, i.d.R. abgetrennt durch „From: ... To: ... CC: ...“. Er enthält Informationen zu CRM-Account-ID, Opportunity-ID, Status, Analyse und On-Hold-Datum.
- **From: (Name):** Den Sender einer E-Mail im Format „Nachname, Vorname“. Er gibt Aufschluss, inwiefern die E-Mail relevant für weitere Datenverarbeitung ist.

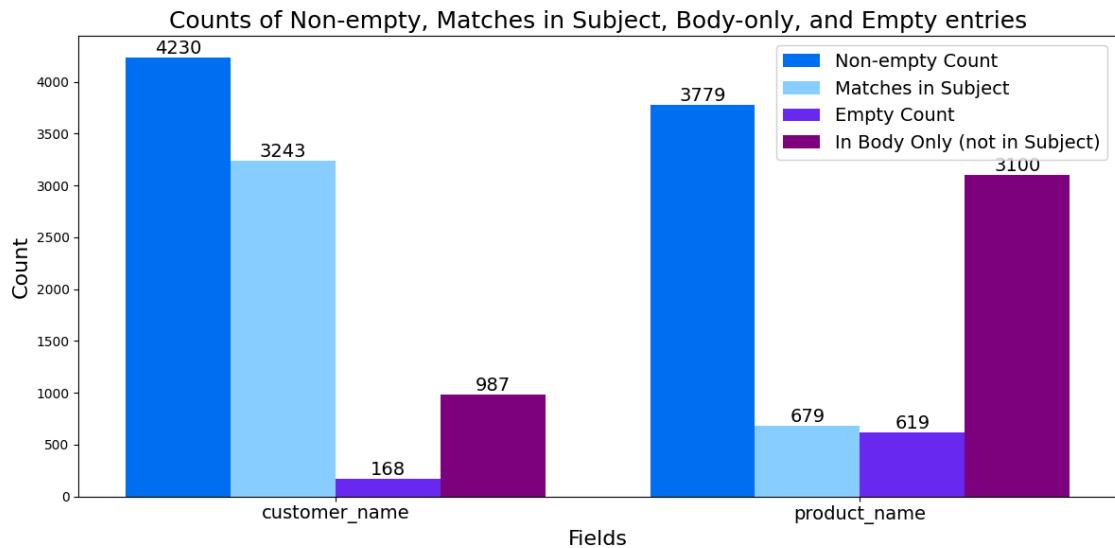


Abbildung 8: Analyse des E-Mail-Betreffs. Eigene Darstellung.

Zur Analyse der Datenqualität wurde eine erste LLM-gestützte Datenextraktion durchgeführt, bei der OpenAIs GPT-4o Modell mittels eines Zero-Shot-Basis-prompts zur Extraktion relevanter Informationen aufgefordert wurde. Diese Konfigurationen wurden aufgrund von guten Erfahrungen der RIG in anderen internen Prozessen gewählt und die Ergebnisse können als Vergleichswert zur Steigerung der Extraktionsqualität verwendet werden [1, S. 16]. Die extrahierten Daten wurden anschließend als Parameter in einen API-GET-Request eingebunden und falls dieser genau einen Eintrag zuordnen kann, wird dieser gespeichert.

Es konnten darunter 4398 mögliche Opportunities identifiziert werden. In diesen sind nur in 15,32% die Opportunity-ID und in 5,22% die CRM-Account-ID enthalten, während Daten wie der Kunden- (96,18%) und Produktnname (85,92%) in einer hohen Konzentration identifiziert werden können. Diese Daten werden an die MXP gesendet und können bei 295 Opportunities (6,71%) den zugehörigen Eintrag identifizieren.

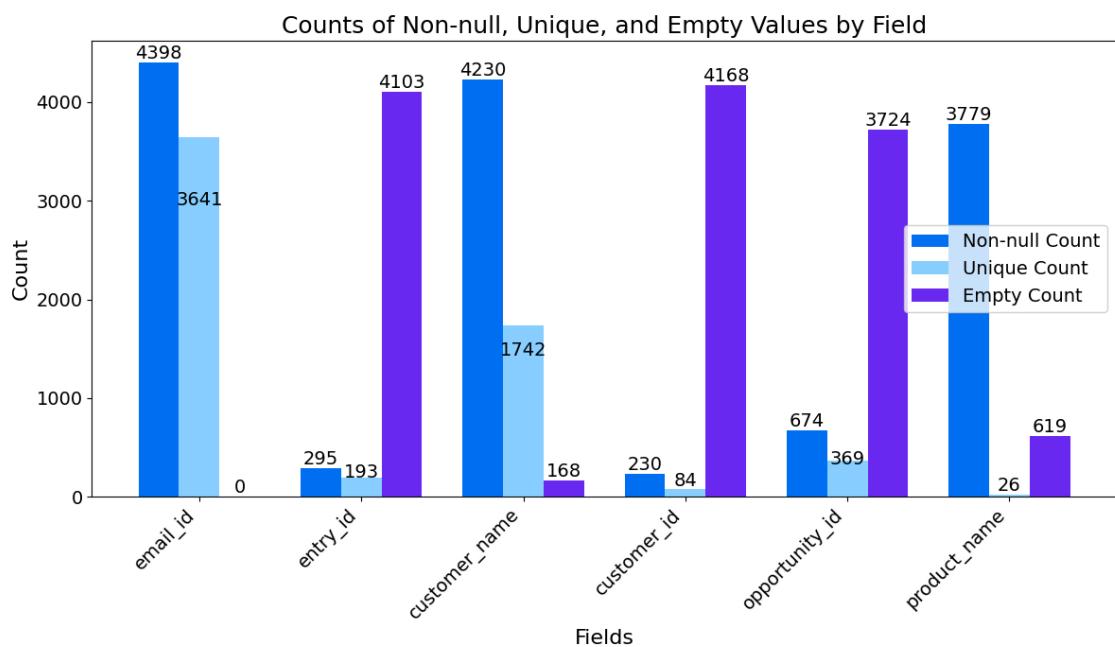


Abbildung 9: Ergebnisse einer initialen Extraktion. Eigene Darstellung.

Die Ergebnisse aus Abbildung 9 sollen durch die in Abschnitt 3.5 vorgestellten Ansätze der Steigerung der Datenqualität hinsichtlich Genauigkeit und MXP-Identifikationsrate verbessert werden. Auf Basis dieser Erkenntnisse wird im Folgenden ein Datensatz erstellt, welcher zur Modellierung und Evaluierung des Experiments geeignet sind.

4.3. Data Preparation

Nach der Analyse der Daten werden die E-Mails des Datensatzes nach ihrer Relevanz für die Datenextraktion klassifiziert. Ausgehend von Abschnitt 4.2 muss eine E-Mail zur Identifikation der Opportunity im MXP folgende Merkmale aufweisen:

- **Betreff und Inhalt:** Eine E-Mail muss die Attribute ‚Subject‘ und ‚Body‘ aufweisen, damit ein LLM notwendige Attribute auslesen kann.
- **Non-RIG:** Viele der E-Mails aus dem Datensatz kommen aus der RIG selbst. Diese haben aus Geschäftsperspektive keine Relevanz, da diese keine zu extrahierenden Daten enthalten.
- **Datenqualität:** Nach erfolgreicher Extraktion der Daten muss der Produkt- und Kundennamen für eine erfolgreiche Identifikation des Eintrags auf der MXP gesetzt sein.

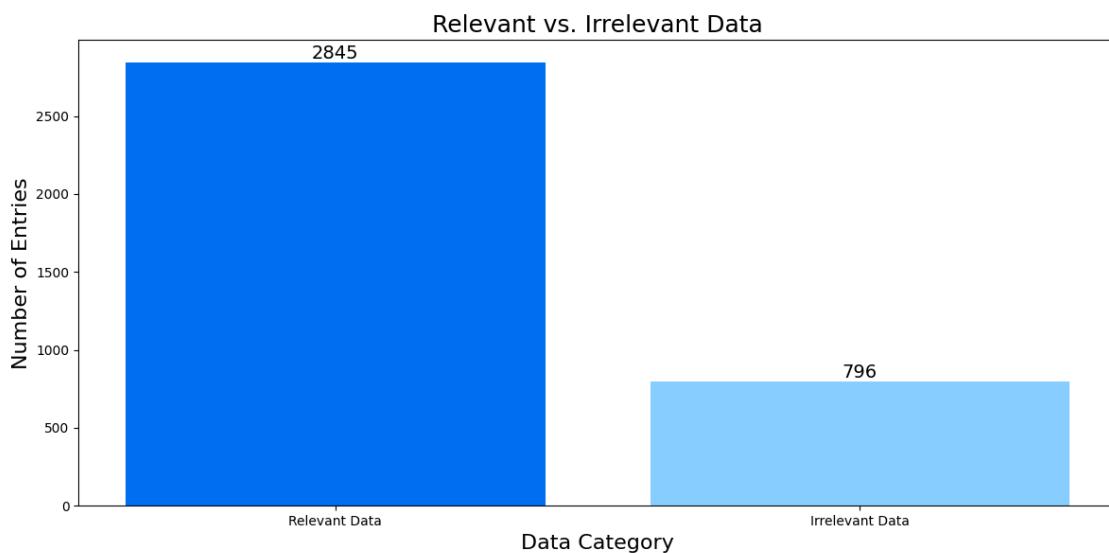


Abbildung 10: Anteil relevanter Daten. Eigene Darstellung.

Von allen 3641 E-Mails des Datensatzes sind 2845 E-Mails relevant für weitere Datenextraktion, da diese Werte in den Attributen ‚Subject‘ und ‚Body‘ als auch von nicht von der RIG oder ‚Others‘ versendet wurden [Abbildung 10]. Die verbleibenden 796 E-Mails werden aus dem Datensatz entfernt.

Zur Optimierung der Datenextraktion hinsichtlich der Opportunity-Zuordnung werden die in Abschnitt 4.1 definierten Hypothesen auf einem reduzierten Validierungsdatensatz durchgeführt, um Zeit und Kosten zur Modellierung und manueller Erstellung einer Referenzquelle (Labeling) zu schonen. Dieser trennt sich in das Segment „Train“, welches zur Angabe von Beispielen an das LLM dient, „Dev“, welches einen Datensatz zur Validierung der Konfigurationsergebnisse beschreibt, und „Test“, auf welchem die Ergebnisse auf dem Validierungsdatensatz hinsichtlich Plausibilität evaluiert werden [90, S. 1138-1139]. Zu jedem der Datensätze wird ein manuell durch den Verfasser extrahierter Referenzdatensatz erstellt.

Da zur Extraktion PLMs eingesetzt werden, welche in ähnlichen Domänen ohne Vortraining der Modelle hohe Extraktionsergebnisse aufweisen, umfasst der Trainingsdatensatz 5 E-Mails, welche bei One- und Few-Shot-Prompts dem Prompt angehängt werden [19, S. 5167-5171]. Die Anzahl wurde auf Basis von Literatur zur Beispielanzahl von Few-Shot-Prompting gewählt, da diese ab einer Anzahl von 5 Beispielen keine signifikante Steigerung der Ausgabequalität mit zunehmender Beispielanzahl in anderen Domänen nachweisen konnten [47, S. 20]. Diese wurden zufällig aus dem Datensatz entnommen und werden als Beispiele für One- und Few-Shot-Prompting eingesetzt.

| Datenqualität | Anzahl | Anteil mit extr. Entry-ID | Fehlende Attr. |
|---------------|--------|---------------------------|--------------------------------|
| Hoch | 14 | 10/14 = 71,43% | - |
| Mittel | 23 | 6/23 = 26,09% | Opportunity-ID |
| Mittel | 30 | 19/30 = 63,33% | Customer-ID |
| Niedrig | 33 | 0/33 = 0% | Customer-ID, Opportunity-ID |

Tabelle 2: Übersicht des Validierungsdatensatzes auf Basis von Abbildung 9

Der Validierungsdatensatz umfasst 100 E-Mails unterschiedlicher Datenqualität, die in den in Tabelle 2 dargestellten Teilmengen enthalten sind. Die ausgewählte Anzahl an E-Mails erlaubt die Überrepräsentation seltener Klassen bei weniger zeitintensiver Modellausführung und Datensatzlabeling, da derselbe Datensatz für jede Modell- und Promptkonfiguration durchlaufen werden muss. Hinzu kann durch die kontrollierte Variation an Datenqualität Overfitting, eine zu starke Anpassung des Modells auf den Trainingsdatensatz, reduziert werden [93, S. 1-2]. Die Datenqualität wurde anhand von Abbildung 9 zugeordnet und alle Daten liegen in einer tabellarischen Form mit den Attributen ‚Subject‘, ‚Body‘ sowie einer zugewiesenen E-Mail-ID vor.

Der Testdatensatz, bestehend aus 20 E-Mails mit zufälliger Datenqualität, wird verwendet, um die Ergebnisse des Validierungsdatensatzes an einem dem Modell unbekannten Datensatz zu überprüfen. Die zufällige Verteilung der Datenqualität ermöglicht eine realitätsnahe Validierung hinsichtlich auftretender Overfitting-Effekte, während die geringe Größe des Datensatzes ein kostengünstiges manuelles Labeling der Daten erlaubt [94, S. 1].

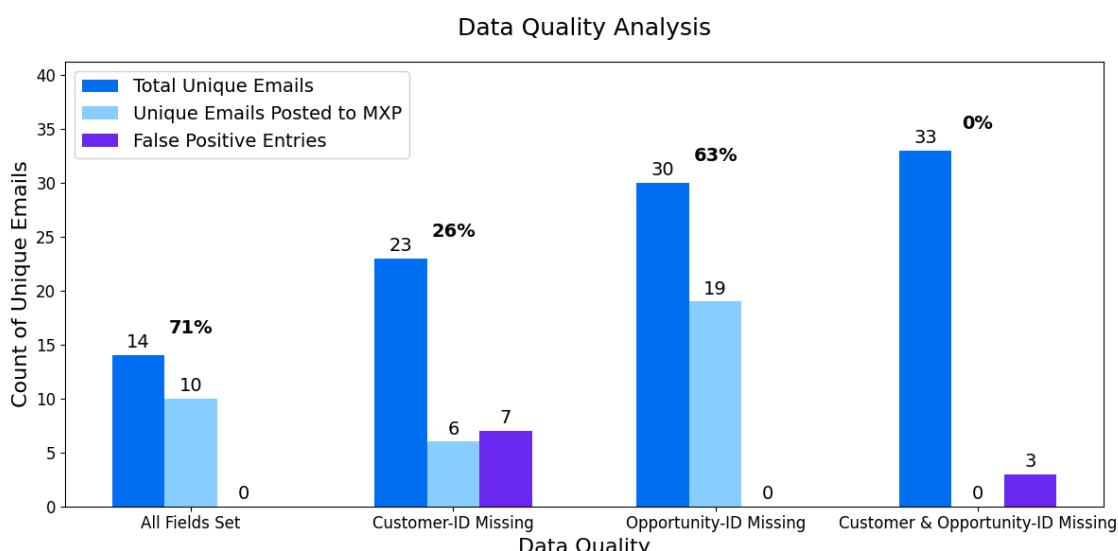


Abbildung 11: Anteil identifizierten Opportunities je Qualität. Eigene Darstellung.

Wie aus Abbildung 11 ersichtlich ist eine korrekte Extraktion mit bisherigem Ansatz bei 35% der E-Mails des Validierungsdatensatzes möglich. Dennoch wurde bei 10% der E-Mails eine falsche MXP-Opportunity zugewiesen, womit der Anteil der falsch positiv identifizierten MXP-Einträge bei 22,22% der Opportunities liegt. Dieser Wert ist nach den von der RIG definierten Zielen aus Abschnitt 4.1 zu hoch für eine Produktivnahme.

Dieser Anteil soll durch Abschnitt 4.4 verringert und der Anteil der korrekt identifizierten Einträge erhöht werden, indem Daten korrekt identifiziert werden. Im Folgenden wird das Vorgehensmodell beschrieben, dieses Ziel zu erreichen.

4.4. Modelling

Zur Optimierung der Extraktion strukturierter Daten werden, wie in Abschnitt 3.3.3 erläutert, Large Language Models (LLMs), aufgrund ihrer hohen Performance gegenüber alternativen Verfahren, eingesetzt [39, S. 5-8], [61, S. 4-5]. Vom Training eines eigenen Modells wird aufgrund von geringer Datengrundlage sowie Kosten- und Rechenleistung abgesehen und verschiedene PLMs eingesetzt [59, S. 1:13].

Da es sich bei der Verarbeitung der Daten, um sensitive Geschäftsdaten der SAP SE handelt, wird zur Verwaltung von Modellanfragen der Orchestration Service des SAP GenAI Hub verwendet, welcher eine Anonymisierung von Kunden- und Mitarbeiterdaten vor Modelleingabe mittels „Data Masking“ sowie intern eine kostengünstige Verarbeitung erlaubt [3]. Darüber hinaus werden die Funktionen „Prompt Templating“ zur dynamischen Generierung von LLM-Prompts sowie „Structured Outputs“ verwendet [66]. Neben der Verwendung von „Structured Outputs“ existiert über den SAP GenAI Hub keine Möglichkeit, ein Modell zu einem einheitlich strukturierten Ausgabeformat zu zwingen, weswegen

die Modellwahl dahingehend angepasst wird. Ein einheitliches Format der LLM-Ausgabe ist notwendig, um eine automatisierte Weiterverarbeitung der Daten zu ermöglichen. Hinsichtlich der Unterstützung von „Structured Outputs“ stehen bislang nur die Modelle von OpenAI mit einer einheitlichen Schnittstelle über den Orchestration Service zur Verfügung [66], weswegen in der Modellierung nur diese betrachtet werden. Die dadurch aufkommenden Limitationen werden in Abschnitt 5.3 gewürdigt.

Zur Validierung der Hypothesen aus Abschnitt 4.1 werden zur Extraktion von E-Mail Daten Modelle verwendet, welche sich hinsichtlich Parameterumfang und Robustheit zueinander unterscheiden und eine Verwendung von „Structured Outputs“ ermöglichen [66], [95], [96]:

- **OpenAI GPT-4o:** GPT-4o, veröffentlicht am 13. Mai 2024 [97], ist im Vergleich zu den anderen Modellen das älteste Modell, erreicht dennoch hohe Genauigkeit bei Datenextraktionsaufgaben [98, S. 3]. Im Vergleich zu o1 und o3-mini zeichnet sich GPT-4o durch seine geringe Time-To-First-Token (TTFT) von 0,421 Sekunden sowie hohe Parametergröße aus [95], [97].
- **OpenAI o1:** Das Modell o1, eingeführt am 5. Dezember 2024, ist ein Modell, welches Reasoning durch internes Chain-of-Thought umsetzt und somit Extraktionsergebnisse von GPT-4o übertrifft [99, S. 9-12]. Gegenüber GPT-4o und o3-mini basiert o1 auf einem größeren Datensatz, ist mit einer TTFT von 12,763 Sekunden im Vergleich das langsamste Modell [95]. [96, S. 2-6]
- **OpenAI o3-mini:** o3-mini, veröffentlicht am 31. Januar 2025, ist das neueste und kompakteste Reasoning-Modell des SAP GenAI Hub. Es zeichnet sich besonders durch sein schnelles und kostengünstiges Reasoning aus, welches in logikbasierten Aufgaben jenes von o1 übertrifft. [100]

Zur Durchführung der Modellierung wird ein Python-Skript verwendet. Python ist eine im Bereich Data Science weit verbreitete Programmiersprache, die, im Gegensatz zu anderen Sprachen, diverse Methoden zur Datenanalyse, -manipulation und -visualisierung als frei-verfügbaren Codepakete (Libraries) bereitstellt [101, S. 27-29]. Die RIG nutzt bereits Python zur Automatisierung anderer interner Prozesse, weswegen Python für die Fortführung der Entwicklung sowie die Integration in die RIG-Inbox geeignet ist. Die im Rahmen dieser Arbeit verwendeten Libraries sind in Tabelle 3 im Anhang dargestellt.

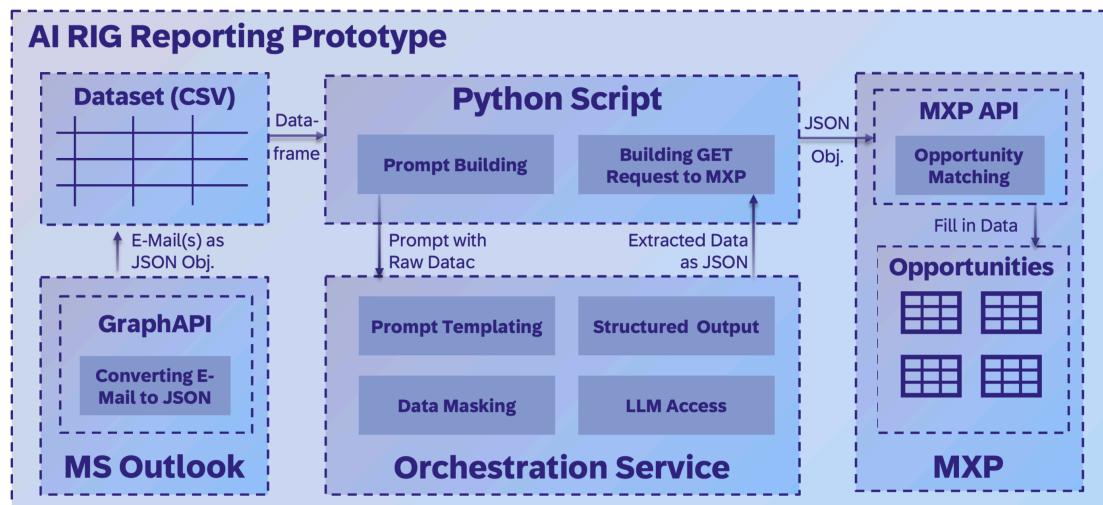


Abbildung 12: Extraktionspipeline des PoCs. Eigene Darstellung.

Abbildung 12 zeigt den Aufbau der Extraktionspipeline. Jede E-Mail des reduzierten Datensatzes aus Abschnitt 4.3 wird einzeln in das Python Skript gelesen und ein LLM-Prompt mittels Prompt Templating erstellt. Dieser wird an das ausgewählte LLM des Orchestration Service gesendet und das zurückgegebene JSON-Object der Struktur Tabelle 4 vom Python Skript verarbeitet. Die extrahierten Daten werden anschließend als Parameter eines GET-Request an die MXP gesendet für jede zum extrahierten Produktnamen zugehörige mögliche Material-ID [Tabelle 6]. Enthält die Antwort exakt eine Opportunity, wird ihr Primärschlüssel

zur Durchführung eines PATCH-Requests genutzt und die extrahierten Reportingdaten auf der MXP gesetzt.

Im Rahmen des Experiments zur Erreichung der maximalen Datenextraktionsgenauigkeit werden an die zuvor ausgewählten Modelle Prompts mit Text aus dem Validierungsdatensatz gesendet. Verwendet wird ein Basisprompt in Anlehnung an M. Moundas, J. White, und D. C. Schmidt [7], ein Chain-of-Thought-Prompt sowie ein Self-consistency-Prompt. Diese werden jeweils als Zero-, One- und Few-Shot-Prompt mit Beispielen aus einem getrennten Testdatensatz durchgeführt, um zu prüfen, inwiefern eine Kombination diverser Prompting-Techniken höhere Extraktionsergebnisse erzielt, wie es in vergleichbaren Anwendungsfällen auftritt [102, S. 22]. Alle Prompts wurden hinsichtlich präziser Zieldefinition und klarer Struktur [80, S. 1-2] mittels mehrerer LLMs optimiert [86, S. 8-10] und für alle E-Mail des Validierungsdatensatzes je Modell durchgeführt.

| Rolle | Inhalt |
|---------------|--|
| System | You are a helpful email data extraction assistant. Your task is to extract key elements from an email. |
| User | EXTRACT DATA FROM THE FOLLOWING EMAIL: Subject: {{?subject}} Body: {{?main_body}} E-Mail-Context: {{?context_body}} |

Prompt 1: Struktur des verwendete Basisprompts in Anlehnung an [7, S. 2-3]

Prompt 1 bietet eine Vorlage für den Aufbau eines Prompts, exemplarisch dargestellt am Basisprompt. Die als {{?Parameter}} gekennzeichneten Variablen werden durch „Prompt Templating“ des Orchestration Service dynamisch pro E-Mail mit Werten ersetzt. Innerhalb des Prompts wird zwischen dem „Body“, dem aktuellen Inhalt der E-Mail, und dem „E-Mail-Context“, dem Inhalt vorheriger E-Mails derselben Konversation, unterschieden, um im Extraktionsschema in

Tabelle 4 den Fundort einzelner Elemente präzise anzugeben. [7, S. 3]. Chain-of-Thought- [Prompt 2], Self-consistency- [Prompt 3] und Addition einer Beispielstruktur für One- und Few-Shot-Prompt [Prompt 4] folgen demselben Aufbau und stehen wegen der Leserlichkeit der Arbeit im Anhang.

Im Modelling häufig aufkommende Risiken umfassen Data-Leakage, der Vermischung von Trainings-, Validierungs- und Testdatensatz [103, S. 4485-4486] sowie falsch-validierte Testdaten. Diese werden mittels einer strikten Trennung von Testdaten (Generierung von Beispielen) und Validierungsdaten (Evaluation der LLM- und Prompt-Konfiguration) sichergestellt [103, S. 4486]. Die manuelle Extraktion durch den Verfasser wurde vor Evaluierung der Daten durch eine Person des Backoffice-Teams validiert, welche mit der Extraktion von E-Mail-Daten vertraut ist.

Für jedes der drei in Abschnitt 4.4 ausgewählten Modelle wurden je neun Prompt-Kombinationen auf 100 E-Mails, somit insgesamt 2700 Extraktionen, durchgeführt. Im Folgenden werden die Extraktionsergebnisse evaluiert.

4.5. Evaluation

4.5.1. Datenextraktionsanalyse

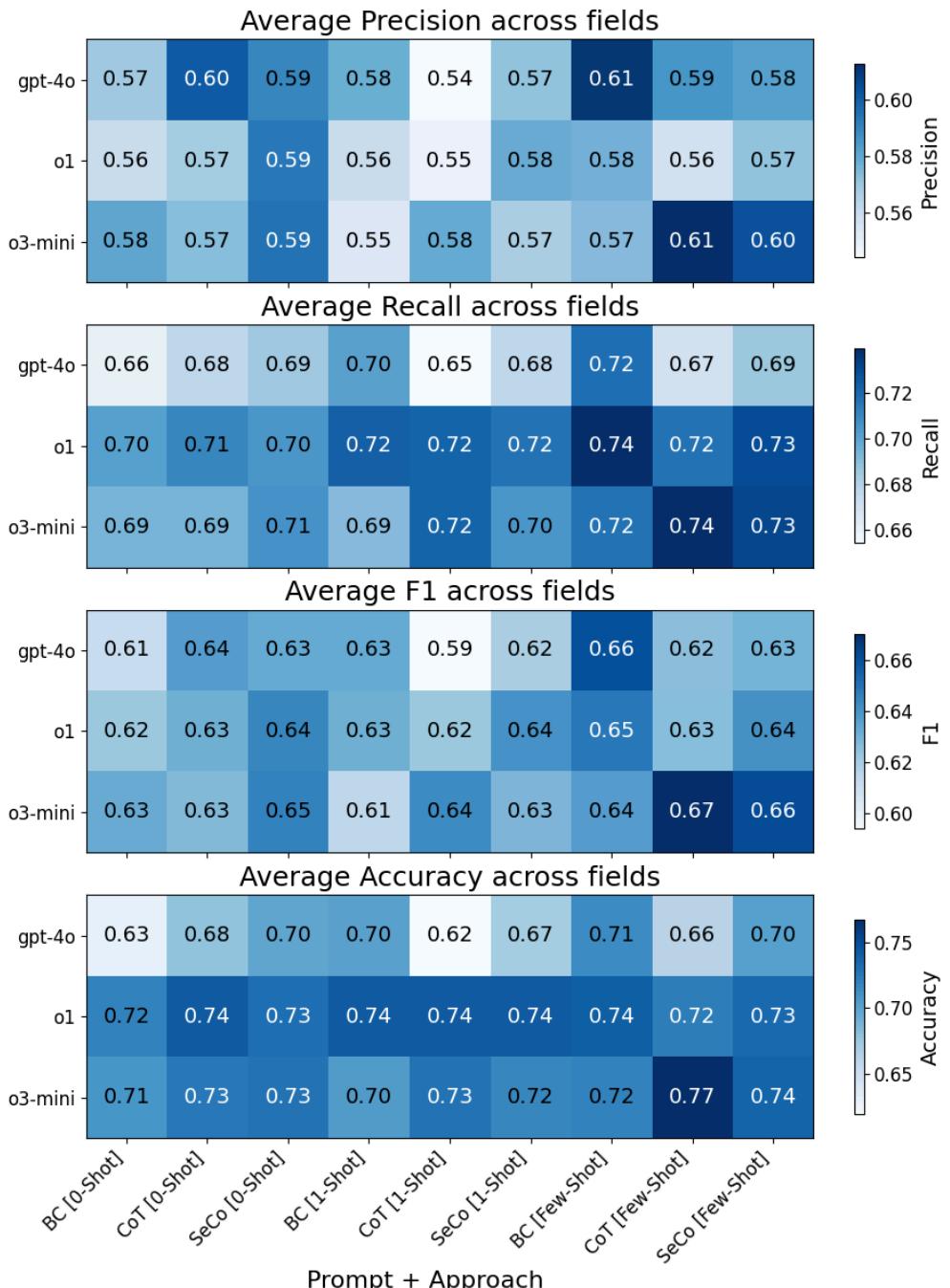


Abbildung 13: Durchschnittliche Metriken je Feld (aggregiert). Eigene Darstellung.

Abbildung 14 und Abbildung 15 stellen die durchschnittliche F1- und Accuracy-Scores in Form einer Heatmaps je Extraktionsfeld dar, Abbildung 13 die über alle Felder aggregierten Metriken.

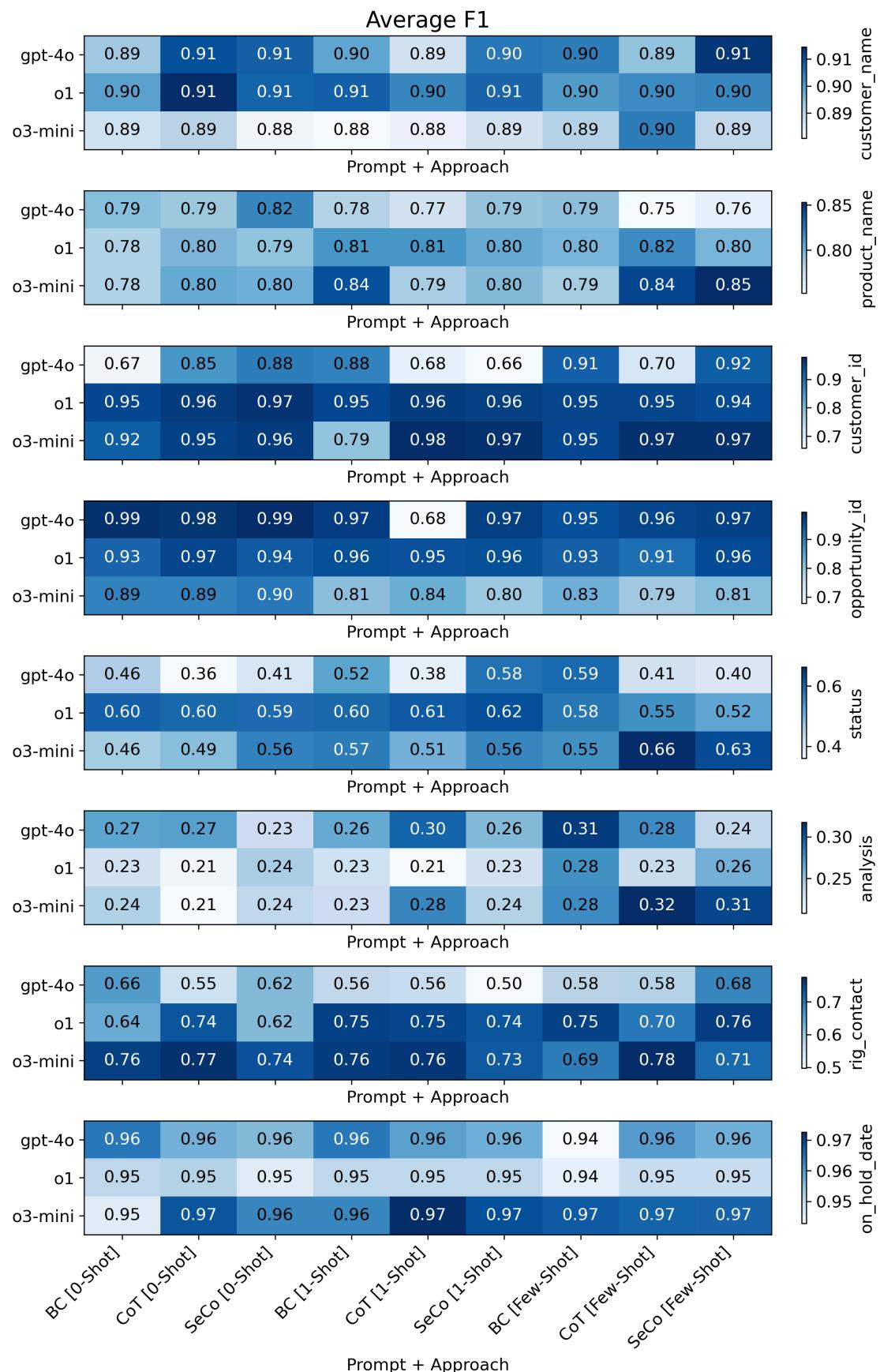


Abbildung 14: Durchschnittlicher F1-Score je Feld. Eigene Darstellung.

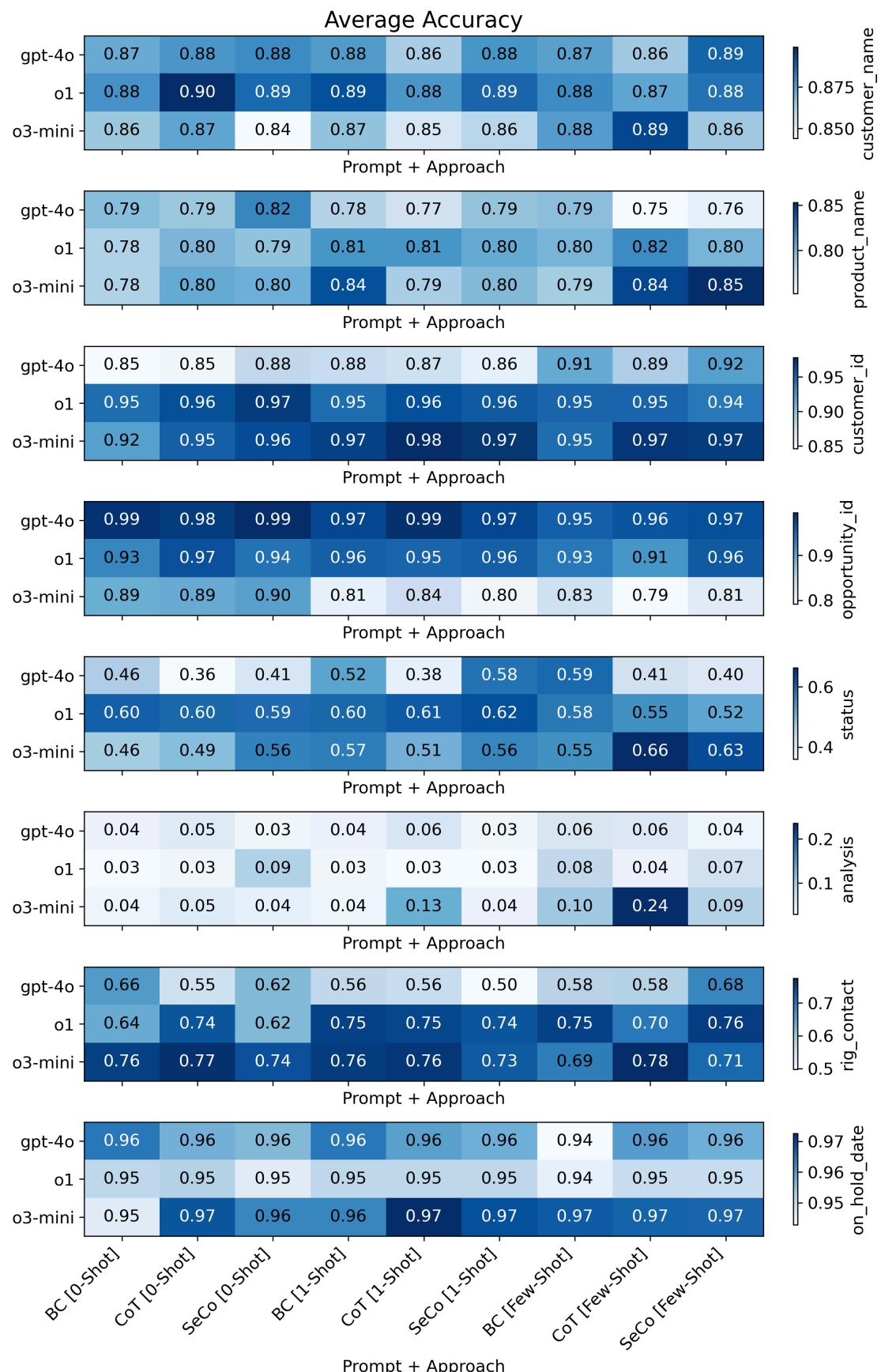


Abbildung 15: Durchschnittliche Accuracy je Feld. Eigene Darstellung.

Die Analyse von Abbildung 14 und Abbildung 15 zeigt modellspezifische Unterschiede: GPT-4o weist gegenüber den anderen Modellen bei der Extraktion der Felder „opportunity_id“ und „analysis“ hohe Ergebnisse auf, erzielt aber in der Extraktion anderen Feldern geringe Ergebnisse. Das Modell o1 liefert konsistent hohe Ergebnisse über alle Felder, während o3-mini zwar die höchsten F1- und Accuracy-Scores pro Feld erreicht, jedoch bei „opportunity_id“ niedrige Ergebnisse erzielt. Für „customer_name“, „analysis“ und „on_hold_date“ bestehen keine signifikanten Modellunterschiede, ebenso zeigen sich keine Auffälligkeiten in Bezug auf Prompting-Technik.

Die Ergebnisse in Abbildung 13 zeigen einen geringen Spread bei Precision und F1-Score, der sich durch die jeweilige Feldvarianz aus Abbildung 14 erklären lässt. Das Modell o3-mini erreicht mit einem CoT Few-Shot-Prompt die höchsten Ergebnisse (F1: 0,67; Accuracy: 0,77), gefolgt von o1 mit einem Basecase Few-Shot-Prompt (F1: 0,65; Accuracy: 0,74). Die Tendenz geringerer Ergebnisse des Modells GPT-4o bestätigt sich, hinsichtlich der Prompting-Techniken können aus Abbildung 13 keine Erkenntnisse abgeleitet werden.

4.5.2. Analyse der Opportunity-Zuordnung

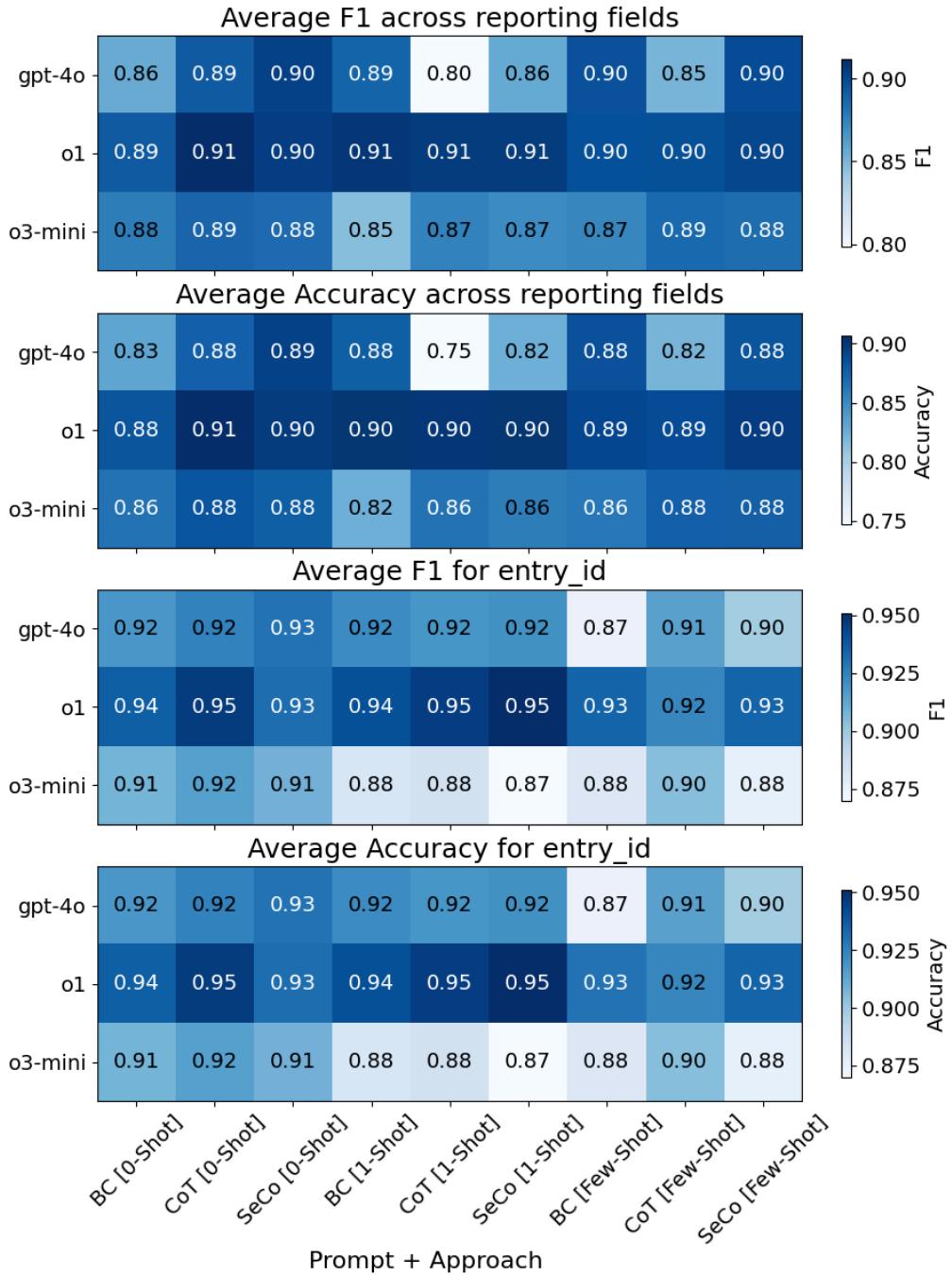


Abbildung 16: Durchschnittliche Metriken (Reportingfelder). Eigene Darstellung.

Abbildung 16 betrachtet die Evaluationsergebnisse jener Felder, welche auf der MXP je Opportunity gesetzt werden, sowie die Metriken für die Entry-ID. Die Ergebnisse zeigen einen Zusammenhang zwischen hohen F1-/Accuracy-Scores der Extraktionsfelder und hoher Identifikation der Entry-ID. Auch hier gibt es modellspezifische, aber keine promptspezifischen Auffälligkeiten: o1 weist bei

der Identifikation der MXP-Opportunity mit einem F1- und Accuracy-Score von 0,95 über drei Konfigurationen hinweg die höchste Rate auf, wobei o3-mini und GPT-4o schlechtere Ergebnisse mit maximaler Genauigkeit von 0,92 bis 0,93 erreichten. Den höchsten F1-Score der Entry-ID erreicht o1 mit drei Prompts, wobei CoT (Zero-Shot) die höchste Accuracy bei den Extraktionsdatenfeldern aufwies.

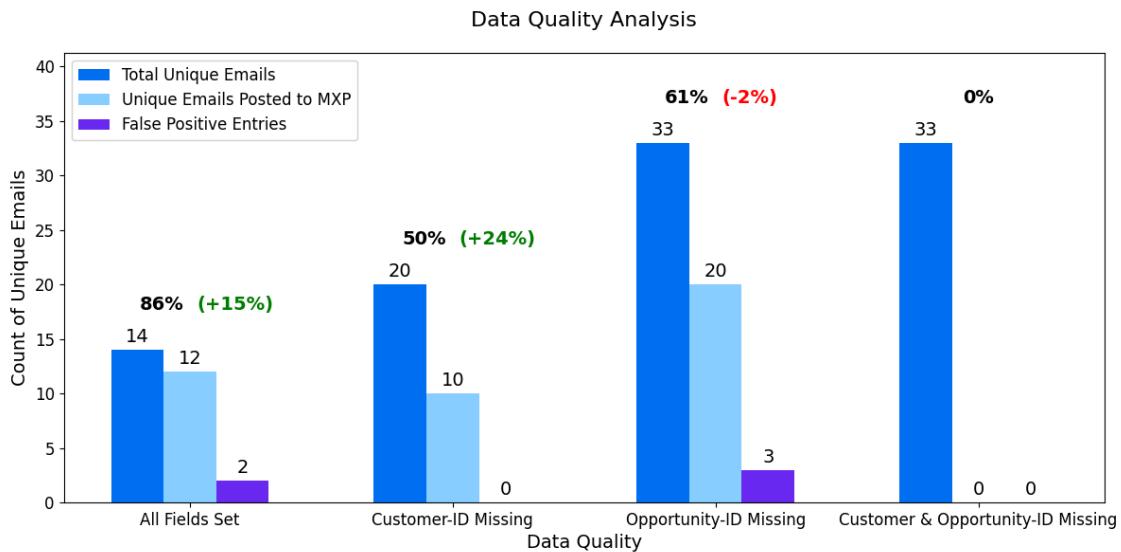


Abbildung 17: Identifikationsrate (beste Konfiguration). Eigene Darstellung.

Im Vergleich zu Abbildung 11 wurde die korrekte Identifikation der Opportunities erheblich verbessert werden, insbesondere, wenn alle Felder Werte enthalten oder die Kunden-ID nicht gesetzt ist [Abbildung 17]. Die niedrigere Rate an falsch-positiven Zuordnungen von MXP-Opportunities legt nahe, dass viele Einträge, die zuvor als qualitativ hochwertig galten, auf fehlerhaften oder unvollständigen Daten basierten. Dies trifft insbesondere auf die Kunden-ID zu, da diese keiner klar erkennbaren Struktur folgt. Diese Annahme wird durch die Beobachtungen aus Abbildung 14 und Abbildung 15 gestützt.

Auf Basis dieser Erkenntnisse wird im Folgenden eine Handlungsempfehlung ausgesprochen. Sie beschreibt, welches Modell mit welcher Prompt-Konfiguration im Rahmen der RIG zum Einsatz kommen sollte.

4.5.3. Gesamtbetrachtung & Konfigurationswahl

Folgende Kombinationen aus LLM und Prompting-Technik erzielten die besten Ergebnisse in beiden Prozessen:

- **o3-mini CoT (Few-Shot)**: Erzielte die höchsten Ergebnisse in der Datenextraktion mit F1: 0,67 und Accuracy: 0,77.
- **o1 CoT (Zero-Shot)**: Erreichte die höchste Accuracy von 0,91 der Extraktionsfelder und eine hohe MXP-Zuordnung.
- **o1 Self-consistency (One-Shot)**: Erzielte den höchsten Anteil an korrekt identifizierten Opportunities.

Anhand der definierten Ziele der RIG aus Abschnitt 4.1 muss eine Konfiguration sowohl eine hohe Extraktionsgenauigkeit als auch eine hohe Identifikationsrate des zugehörigen MXP-Eintrags aufweisen, da für ein erfolgreiches Reporting einer E-Mail beide Prozesse erfüllt sein müssen. Daher werden im Folgenden beide Prozesse gleich stark gewichtet.

Um dies zu gewährleisten, werden die Ergebnisse der F1- und Accuracy mittels eines gewichteten arithmetischen Mittels bestimmt, bei welcher jedem F1/Accuracy-Score eine Metrik-Gewichtung (k) zugewiesen wird. Hierbei wird der Prozess der Datenextraktion aller acht Felder gleich stark gegenüber dem Prozess der MXP-Opportunity-Zuordnung bewertet, für den stellvertretend die Metriken der Entry-ID betrachtet wird, da diese nur bei erfolgreicher Identifikation des MXP-Eintrags gesetzt wird. Bei vollständiger Attributbetrachtung gilt $k = 8$. Die folgende Formel wird zu Berechnung der konfigurationsspezifischen Metriken verwendet. Dabei ist $F1_{conf}$ der konfigurationsspezifische F1-Score und Acc_{conf} die konfigurationsspezifische Accuracy in Abhängigkeit der Gewichtung k des MXP-Opportunity-Zuordnungsprozess.

$$F1_{\text{conf}}(k) = \frac{\sum_{i=1}^k F1_i + k \cdot F1_{\text{MXP}}}{2 \cdot k}, \quad k \in \mathbb{N} \quad (8)$$

$$\text{Acc}_{\text{conf}}(k) = \frac{\sum_{i=1}^k \text{Acc}_i + k \cdot \text{Acc}_{\text{MXP}}}{2 \cdot k}, \quad k \in \mathbb{N} \quad (9)$$

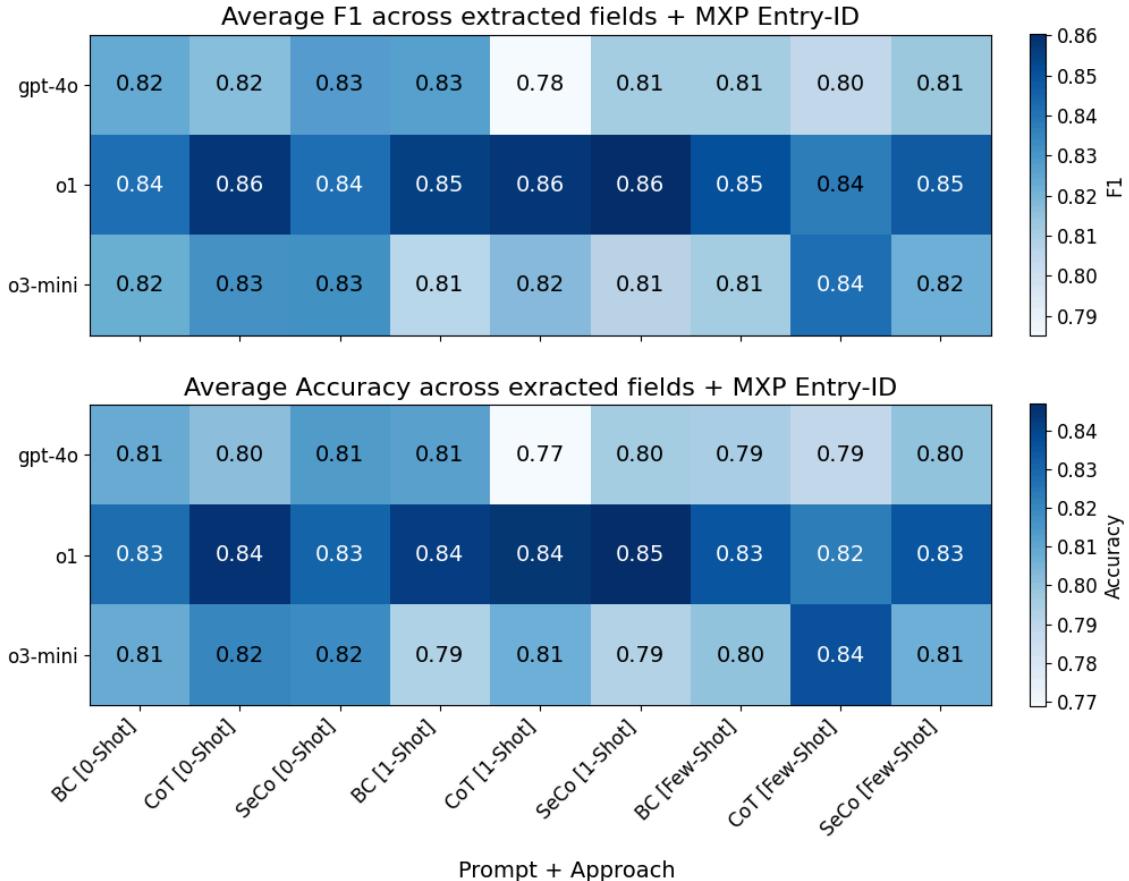


Abbildung 18: Durchschnittliche Metriken je Prozess. Eigene Darstellung.

Wie aus Abbildung 18 hervorgeht erzielt o1 in F1- und Accuracy-Score bessere Ergebnisse als GPT-4o und o3-mini. Die beste Konfiguration erzielt o1 mit Self-consistency (One-Shot)-Prompting ($F1_{\text{config}}(8) = 0,86$; $\text{Acc}_{\text{config}}(8) = 0,85$). Abseits des Modells o1 erreicht o3-mini mit einem CoT Few-Shot-Prompt die besten Metriken und kommt somit als Modellalternative in Frage ($F1_{\text{config}}(8) = 0,84$; $\text{Acc}_{\text{config}}(8) = 0,84$).

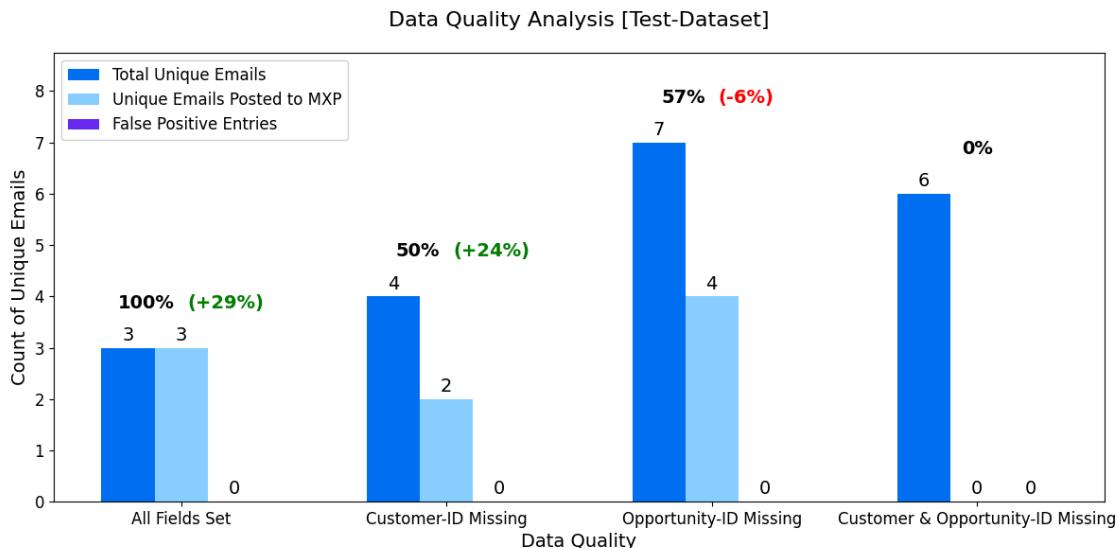


Abbildung 19: Durchschnittliche Metriken je Prozess. Eigene Darstellung.

Aus Abbildung 19 ergibt sich gegenüber der individuellen Modellleistung aus Abbildung 17 je Datenqualität vergleichbare Ergebnisse, wodurch die Ergebnisse explorativ als plausibel einzuordnen sind, wobei für eine empirische Verbesserung ein größerer Datensatz notwendig ist [90, S. 1138-1139].

Damit erzielt o1 bei Analyse beider Abbildungen mit einem Self-consistency One-Shot-Prompt die höchsten Extraktionsergebnisse. o3-mini erreicht mit einem CoT Few-Shot-Prompt hingegen die beste MXP-Opportunity-Zuordnung und zeigt insgesamt eine vergleichbare Extraktionsqualität, trotz deutlich geringerer Kosten und reduzierter TTFT. Für weitere Optimierungsansätze sollten beide Modelle berücksichtigt werden.

4.5.4. Hypothesenvalidierung

Zur Bewertung der in Abschnitt 4.1 definierten Hypothesen werden die F1- und Accuracy Ergebnisse aus Abschnitt 4.5 betrachtet. Eine Hypothese gilt als:

- **angenommen:** Beide Kennzahlen steigen gegenüber der Referenz (Baseline) erkennbar an, zeigen über alle Modelle hinweg eine positive Entwicklung und der Unterschied ist statistisch signifikant [104, S. 2-3].
- **teilweise angenommen:** Mindestens eine Kennzahl verbessert sich gegenüber der Baseline, während die andere Metrik stabil bleibt [104, S. 2-3]. Der Unterschied muss nicht statistisch signifikant sein, solange keine gegenläufigen Entwicklungen zwischen den Modellen auftreten [105, S. 363].
- **abgelehnt:** Die Veränderung stagniert oder es treten sich widersprechende Ergebnisse auf. Ebenso wird die Hypothese abgelehnt, wenn der Unterschied statistisch nicht aussagekräftig ist [104, S. 2].

Ein Effekt wird in dieser Arbeit als statistisch signifikant klassifiziert, wenn das Signifikanzniveau $\alpha < 0,05$ beträgt [106, S. 163-185], [107, S. 6]. Zur Bestimmung dieser Wahrscheinlichkeit wurde ein gepaarter, nicht-parametrischer Bootstrap-Ansatz nach B. Efron und R. J. Tibshirani [108] verwendet, bei dem F1- und Accuracy-Metriken wiederholt berechnet und deren Abweichungen vom arithmetischen Mittel analysiert wurden. Signifikanz liegt vor, wenn die Gesamtabweichung kleiner als 0,05 ist. [108, S. 5-8]

Abbildung 20 zeigt die Auswirkungen der Parameter Modell, Prompt und Beispielumfang. Einträge mit der Notation * gelten als statistisch signifikant. Die Ergebnisse werden als Hauptquelle für die Annahme bzw. Ablehnung der Hypothese verwendet.

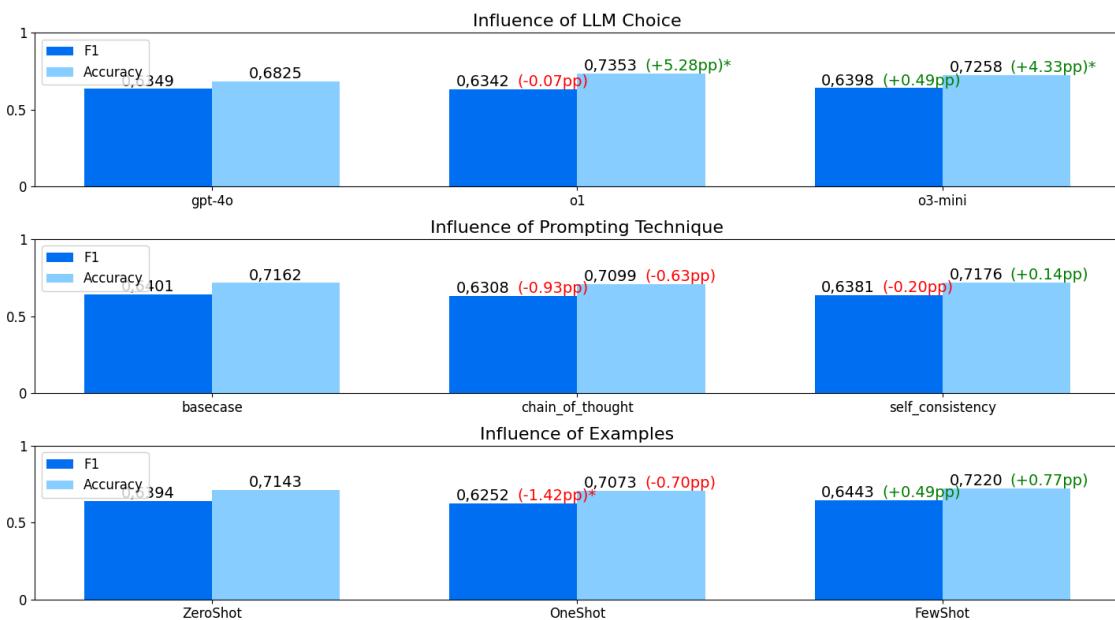


Abbildung 20: Auswirkung von Prompt/LLM. Eigene Darstellung.

Im Folgenden erfolgt die Annahme bzw. Ablehnung der Hypothesen 1-6 auf Basis der Datenextraktion aller Felder:

- **Hypothese 1 - One-Shot > Zero-Shot:** Die Ergebnisse aus Abbildung 20 zeigen eine statistisch signifikante Reduktion des F1 um 1,42 Prozentpunkt (pp) und der Accuracy um 0,63pp. Da keine Verbesserung vorliegt, wird diese Hypothese **abgelehnt**.
- **Hypothese 2 - Few-Shot > Zero-Shot und One-Shot:** Gegenüber Hypothese 1 weist die Anwendung von Few-Shot-Prompting marginale Verbesserungen gegenüber Zero-Shot von unter 1pp, wobei eine signifikante Verbesserung gegenüber One-Shot zu erkennen ist. Aufgrund fehlender Signifikanz aller Resultate wird die Hypothese **teilweise angenommen**.
- **Hypothese 3 - Chain-of-Thought > Basisprompt:** Gegenüber dem Basisprompt zeigt CoT eine geringe Verschlechterung der Metriken weswegen die Hypothese **abgelehnt** wird.

- **Hypothese 4 - Self-consistency > Basisprompt:** Gegenüber Hypothese 3 gibt es bei Anwendung von Self-consistency-Prompting keine signifikante Verbesserung, weswegen die Hypothese **abgelehnt** wird.
- **Hypothese 5 - Reasoning Modell > Non-Reasoning-Modell:** Wie aus Abbildung 20 ersichtlich zeigt die Verwendung von o1 und o3-mini gegenüber GPT-4o einen signifikanten Zuwachs der Accuracy von 4,33-5,28pp, der F1-Score zeigt marginale Veränderung. Aufgrund dessen wird die Hypothese **teilweise angenommen**.
- **Hypothese 6 - Größere Modelle > Kleinere Modelle:** Wie aus Abbildung 20 und Hypothese 5 herausgehen gibt es eine Extraktionsverbesserung zwischen o1 und o3-mini, wobei GPT-4o eine signifikante Verschlechterung der Accuracy gegenüber o3-mini aufweist. Daher wird die Hypothese **abgelehnt**.

Insgesamt wurden aus allen sechs Hypothesen vier Hypothesen abgelehnt und zwei Hypothesen teilweise angenommen. Eine Analyse zeigt, dass mithilfe der Modellwahl die Datenextraktion optimiert werden kann, die Anwendung von Prompt Engineering aber keine signifikante Steigerung der Metriken aufweist.

4.6. Deployment

Die Ergebnisse aus Abschnitt 4.4 zeigen eine erfolgreiche Steigerung der Evaluationsmetriken gegenüber Abbildung 9. Beide in Abschnitt 4.5.3 ausgewählten Konfigurationen erreichen die in Abschnitt 4.1 für eine Produktivnahme definierten Vorgaben der RIG von mindestens 70% Genauigkeit, einer Übereinstimmung der manuellen MXP-Identifikation von 95% Genauigkeit [Abbildung 16] und einen Anteil falsch-positiv identifizierter Opportunities von 5% [Abbildung 17]. Damit erfüllt der PoC die Vorgaben der RIG und ein Deployment mit Anbindung an die RIG-Inbox kann durchgeführt werden. Es empfiehlt sich ein schrittweises Vorgehen, bei dem die Änderungen auf der MXP zunächst als Vorlage gespeichert werden und eine Begutachtung durch die RIG vor dem Reporting erfolgen muss.

Nächste Schritte umfassen die Validierung der Ergebnisse auf einem zeitlich aktuellen Datensatz, um Abweichungen durch die Datengrundlage auszuschließen [93, S. 2]. Durch folgende Maßnahmen kann die RIG die Extraktionsqualität vor der produktiven Implementierung steigern:

- **Anpassung der E-Mail-Kampagnen:** Die RIG kann durch Hinzubrügen von MXP-Identifikatoren wie die CRM-Account- und Opportunity-ID die Datenqualität zukünftiger E-Mails durch die Anpassung ihrer E-Mail-Kampagnen steigern [Abbildung 17].
- **Konkretisierung der Reportingdaten:** Durch die Konkretisierung der Reportingwerte, wie z.B. dem Status aus Tabelle 5 im Anhang oder einem RIG-Kontakt innerhalb der E-Mail, wird die korrekte Identifizierung der Daten erleichtert.

Nach erfolgreicher Validierung der Ergebnisse übernimmt die produktive Entwicklung sowie Wartung und Support die RIG. Diese Arbeit dient als Dokumentation der bisherigen Entwicklung.

5. Schlussbetrachtung

5.1. Zusammenfassung der Ergebnisse

Eine Analyse der Ergebnisse der Modellierung zeigt, dass die Datenextraktion durch die in Abschnitt 3.5 vorgestellten Ansätze und damit verbunden das Setzen von strukturierten E-Mail-Kommunikationsdaten verbessert wird. Reasoning-Modellen wie o1 und o3-mini steigern die Extraktionsgenauigkeit, eine Verbesserung der Ergebnisse durch Prompt Engineering konnte in dieser Arbeit nicht nachgewiesen werden [Abbildung 20].

Die von der RIG in Abschnitt 4.1 definierten KPIs konnten im Rahmen des PoC erreicht werden. Die besten Ergebnisse erzielte das Modell o1 mit einem Self-consistency One-Shot-Prompt mit einem F1-Score von 0,64, einer Genauigkeit von 0,74 und einer MXP-Opportunity-Identifikationsgenauigkeit von 95% [Abbildung 13, Abbildung 16].

Aufgrund der hohen Variation der Ergebnisse je Modellkonfiguration und einer fehlenden statistischen Aussagefähigkeit durch einen geringen Datensatz spricht sich die Arbeit für eine Validierung der Ergebnisse auf einem größeren, zeitlich aktuelleren Datensatz aus. Darüber hinaus soll die RIG durch Anpassung ihrer E-Mail-Kampagnen die Datenqualität je E-Mail erhöhen, um die MXP-Identifikationsrate weiter zu steigern.

Nach Erfüllung der definierten Kriterien übernimmt die RIG die produktive Implementierung des PoC sowie Aufgaben im Bereich Wartung und Support. Aufgrund der erfolgreichen Steigerung der Datenqualität empfiehlt die Arbeit, dass die RIG Maßnahmen zur Verbesserung der Datenqualität fortsetzt.

5.2. Einordnung der Ergebnisse

Die erzielten Ergebnisse aus Abschnitt 4.5 reihen sich in die aktuelle Literatur zu LLM-basierter Datenextraktion ein. Die erreichte Genauigkeit von 74% ist vergleichbar mit der Genauigkeit in anderen Bereichen [19, S. 5167 - 5174] und zeigt ein ähnliches Verhalten beim Einsatz von Reasoning-Modellen [99, S. 9-12]. Diese Arbeit erzielt, entgegen Datenextraktionen aus anderen Bereichen [10, S. 8-9], [79, S. 7-9] keine signifikante Genauigkeitsveränderung durch die Anwendung von Prompt Engineering. Mögliche Ursachen dieser Abweichung, wie anwendungsspezifische Faktoren oder Fehler bei der Modellierung, können durch eine wiederholte Durchführung des Experiments sowie die Anpassung der Prompts identifiziert werden.

5.3. Herausforderungen und Limitationen

Die Ergebnisse aus Abschnitt 4.5 unterliegen Limitation hinsichtlich ihrer Aussagefähigkeit durch Herausforderungen bei Erhebung und Interpretation der Daten.

Der zur Modellierung verwendete Datensatz, bestehend aus 125 E-Mails, aufgeteilt in Trainings-, Validierungs- und Testdatensatz, weist Herausforderungen auf: im Vergleich zu anderen Datenextraktionsanalysen [10, S. 3-4], [99, S. 8-9] ist der Validierungsdatensatz klein und besitzt eine geringe Datenvarianz, wodurch die Ergebnisse dieser Arbeit nicht als Optimierung und empirisch angesehen werden können, sondern als explorative Untersuchung beschränkt auf die Domäne der E-Mail-Kommunikation der RIG [93, S. 1-2]. Darüber hinaus können bei der manuellen Extraktion Fehler durch den Verfasser, trotz Validierung eines RIG-Mitarbeiters, Fehler auftreten, die die Evaluationsergebnisse verzerren und Hypothesen zu Unrecht abgelehnt werden [10, S. 9-11].

Weitere Herausforderungen ergeben sich durch die Verwendung von LLMs gegenüber regelbasierten Verfahren: durch das persistierende Risiko falsch extrahierter Daten sowie Halluzinationen ist ein angestrebter Anteil von 0% falsch-positiver MXP-Opportunity-Zuordnung erschwert möglich [54, S. 11-12]. Des Weiteren wurde im Rahmen dieser Arbeit je Prompting-Technik nur ein Prompt [Prompt 2, Prompt 3] generiert, wodurch Fehler bzw. modellspezifische Bias auftreten können [86, S. 8-10]. Im Rahmen einer anderen Arbeit kann dies untersucht werden und Methoden wie LLM-Stacking zur Validierung der Ergebnisse eingesetzt werden [109].

Die Interpretation der Ergebnisse ist limitiert durch die in dieser Arbeit verwendeten Tools: die Verwendung des Orchestration Service des SAP GenAI Hub limitiert die einsetzbaren Modelle auf den Anbieter OpenAI [66], [76, S. 1], wodurch ein hohes Risiko von Vendor-Lock-In besteht, was die Ergebnisse verzerren kann [110, S. 3-7]. Zur Validierung der Ergebnisse ist eine erneute Durchführung mit Modellen weiterer Anbieter empfehlenswert, um dieses Risiko zu minimieren.

Weitere Limitationen treten durch die zu extrahierenden Daten auf: in dieser Arbeit wurden die in Tabelle 4 dargestellten Felder extrahiert. Wie aus Abbildung 14 und Abbildung 15 ersichtlich variieren die Metriken einer Modell- und Prompting-Konfiguration stark je Feld, wodurch die in dieser Arbeit erhobenen Evaluationsergebnisse für andere Anwendungsfälle erneut evaluiert werden müssen.

Abschließend ist die Extraktion von Reportingdaten bislang limitiert auf die Extraktion von Betreff und Inhalt der E-Mail. Eine Analyse von angehängten Dateien sowie strukturelle Informationen wie Formatierung oder eingebettete Hyperlinks wird im Rahmen des PoCs nicht unterstützt, bietet aber eine Initiative für weitere Forschung.

5.4. Ausblick

Die in Abschnitt 1.2 definierten Forschungsfragen konnten im Rahmen der Entwicklung eines PoCs beantwortet werden. Dabei wurde an mehreren Stelle Potential für weitere Forschung identifiziert. Diese können im Rahmen einer produktiven Entwicklung durch die RIG miteinbezogen werden, um die Extraktionsqualität weiter nachhaltig zu verbessern. Diese umfassen:

- **Modellauswahl:** Durch die Veröffentlichung weiterer Modelle diverser Anbieter, die die Leistung der alten Modelle übertreffen, bieten diese Potential für bessere Extraktionsergebnisqualität in der Zukunft [111, S. 17-19].
- **Hyperparameter-Tuning:** Eine Verbesserung der Extraktionsqualität anhand von Hyperparametern erzielt in anderen Datenextraktionsaufgaben eine verbesserte Extraktionsgenauigkeit [37]. Der SAP GenAI Hub soll in Zukunft das Setzen von Temperatur und Tokenlimit der Modelle o1 und o3-mini unterstützen.
- **Algorithmus zur MXP-Identifikation:** Bislang werden die extrahierten Daten in Form von Parametern an einen MXP-GET-Request angehängt. Anhand eines Algorithmus, welcher falsch extrahierte Daten eliminiert und Duplikate auf der MXP identifiziert, kann die Rate der an das MXP gesendeten Einträge verbessert werden.
- **Einbindung mehrerer Datenquellen:** Wie in Abschnitt 5.3 erläutert können zur Erhöhung der Extraktionsrate weitere Datenquellen wie bspw. der Anhang einer E-Mail sowie über Microsoft Outlook hinausgehende Datenquellen genutzt werden, um eine höhere Datenqualität aufzuweisen.

Hierbei sollten die in Abschnitt 5.3 ausgeführten Herausforderungen und Limitationen beachtet werden, sowie ein größerer Datensatz verwendet werden, um die Aussagefähigkeit der Ergebnisse zu verbessern.

i. Literaturverzeichnis

- [1] SAP RIG AI, „Drive & Track Booked AI Opportunities“. Zugegriffen: 28. Februar 2025. [Online]. Verfügbar unter: https://sap.sharepoint.com/:p/r/sites/209179/_layouts/15/Doc.aspx?sourcedoc=%7B4A2B54EE-0AB6-4351-9FE7-490DE2BA822C%7D&file=SAP%20AI%20RIG%20-%20Drive%20%26%20Track%20booked%20AI%20opportunities.pptx&action=edit&mobileredirect=true&previoussessionid=10858943-7f66-a109-83e0-14a3653705f1
- [2] R. Wirth und J. Hipp, „CRISP-DM: Towards a standard process model for data mining“, in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000, S. 29–39. Zugegriffen: 28. Februar 2025. [Online]. Verfügbar unter: <https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf>
- [3] SAP SE, „Solving Your Business Problems Using Prompts and LLMs in SAP's Generative AI Hub“. Zugegriffen: 6. März 2025. [Online]. Verfügbar unter: <https://learning.sap.com/learning-journeys/solving-your-business-problems-using-prompts-and-llms-in-sap-s-generative-ai-hub/using-generative-ai-hub-sdk-to-leverage-power-of-llms>
- [4] SAP, „Multi Experience Platform Documentation“. Zugegriffen: 3. März 2025. [Online]. Verfügbar unter: <https://documentation.value-experience-hub.for.sap/docs/intro>
- [5] P. Hoffmann, „SAP AI Services - Strategic Positioning“. Zugegriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://workzone.one.int/>

sap/site#workzone-home&/groups/i5bfKiR7lvKpC0QI282gfZ/documents/
XtjgLMPFW8LAIGnoROj6R/slide_viewer

- [6] Z. C. Lipton, C. Elkan, und B. Narayanaswamy, „Thresholding Classifiers to Maximize F1 Score“, *arXiv preprint*, 2014, Zugriffen: 7. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/1402.1892>
- [7] M. Moundas, J. White, und D. C. Schmidt, „Prompt Patterns for Structured Data Extraction from Unstructured Text“. Zugriffen: 4. Februar 2025. [Online]. Verfügbar unter: https://www.dre.vanderbilt.edu/~schmidt/PDF/Prompt_Patterns_for_Structured_Data_Extraction_from_Unstructured_Text.pdf
- [8] C. E. Shannon, „A Mathematical Theory of Communication“, *The Bell System Technical Journal*, Bd. 27, Nr. 3, 4, S. 379–423623–656, 1948.
- [9] A. Vaswani u. a., „Attention Is All You Need“, in *Advances in Neural Information Processing Systems (NeurIPS)*, Curran Associates, Inc., 2017. Zugriffen: 27. März 2025. [Online]. Verfügbar unter: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf
- [10] R. Srivastava, S. Prasad, L. Bhat, S. Deshpande, B. Das, und K. Jadhav, „MedPromptExtract (Medical Data Extraction Tool): Anonymization and Hi-fidelity Automated data extraction using NLP and prompt engineering“. Zugriffen: 7. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2405.02664v3>
- [11] McKinsey u. a., „The Economic Potential of Generative AI: The Next Productivity Frontier“, Juni 2023. Zugriffen: 27. Februar 2025. [Online].

Verfügbar unter: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>

- [12] SAP, „SAP Business AI – Künstliche Intelligenz für Unternehmen“. Zugegriffen: 22. April 2025. [Online]. Verfügbar unter: <https://www.sap.com/germany/products/artificial-intelligence.html>
- [13] S. Sarawagi, „Information Extraction“, *Foundations and Trends in Databases*, S. 261–377, 2007, Zugegriffen: 12. März 2025. [Online]. Verfügbar unter: <https://www.cse.iitb.ac.in/~soumen/readings/papers/Sarawagi2008ie.pdf>
- [14] SAP, „SAP Concur enhanced by Generative AI“. Zugegriffen: 22. April 2025. [Online]. Verfügbar unter: <https://www.sap.com/assetdetail/2023/10/3c45ed47-937e-0010-bca6-c68f7e60039b.html>
- [15] SAP News Center, „AI in 2025: Five Defining Themes“, *SAP News Center*, Jan. 2025, Zugegriffen: 4. März 2025. [Online]. Verfügbar unter: <https://news.sap.com/2025/01/ai-in-2025-defining-themes/>
- [16] C. Schröer, F. Kruse, und O. Görke, „A Systematic Literature Review on Applying CRISP-DM Process Model“, 2021, Zugegriffen: 25. Februar 2025. [Online]. Verfügbar unter: <https://www.sciencedirect.com/science/article/pii/S1877050921002416>
- [17] J. Oliveira und A. Oliveira, „Text Mining: Crossing the Chasm Between the Academy and Industry“, in *Data Mining III*, WIT Press, 2002, S. 351–360. Zugegriffen: 4. März 2025. [Online]. Verfügbar unter: <https://www.witpress.com/Secure/elibrary/papers/DATA02/DATA02035FU.pdf>

- [18] O. S. Choudhry und others, „Data Collection and Analysis of French Dialects“, *arXiv preprint arXiv:2208.00752*, 2022, Zugriffen: 4. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2208.00752>
- [19] M. Wang *u. a.*, „Precision Structuring of Free-Text Surgical Record for Enhanced Stroke Management: A Comparative Evaluation of Large Language Models“, *Journal of Multidisciplinary Healthcare*, Bd. 17, S. 5163–5175, 2024, Zugriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://www.dovepress.com/precision-structuring-of-free-text-surgical-record-for-enhanced-stroke-peer-reviewed-fulltext-article-JMDH>
- [20] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, und Z. Wang, „A Survey of Information Extraction Based on Deep Learning“, *Applied Sciences*, Bd. 12, Nr. 19, S. 9691, 2022, doi: 10.3390/app12199691.
- [21] B. Adamson *u. a.*, „Approach to Machine Learning for Extraction of Real-World Data Variables from Electronic Health Records“, *Frontiers in Pharmacology*, Bd. 14, S. 1180962, 2023, Zugriffen: 7. März 2025. [Online]. Verfügbar unter: <https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1180962/full>
- [22] M. Labonne und S. Moran, „Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection“, *arXiv*, 2023, doi: 10.48550/arXiv.2304.01238.
- [23] X. Pu, M. Gao, und X. Wan, „Summarization is (Almost) Dead“, *arXiv*, 2023, Zugriffen: 24. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2309.09558>

- [24] J. Saltz, I. Shamshurin, und K. Crowston, „Comparing Data Science Project Management Methodologies via a Controlled Experiment“, in *Proceedings of the Annual Hawaii International Conference on System Sciences*, Hawaii International Conference on System Sciences, 2017, S. 1031–1022. doi: 10.24251/hicss.2017.120.
- [25] T. Hu und X.-H. Zhou, „Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions“, *arXiv preprint*, Apr. 2024, Zugegriffen: 12. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/pdf/2404.09135>
- [26] D. Hein *u. a.*, „Prompts to Table: Specification and Iterative Refinement for Clinical Information Extraction with Large Language Models“, *medRxiv*, 2025, Zugegriffen: 7. März 2025. [Online]. Verfügbar unter: <https://doi.org/10.1101/2025.02.11.25322107>
- [27] C.-Y. Lin, „ROUGE: A Package for Automatic Evaluation of Summaries“, in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Juli 2004, S. 74–81. doi: 10.3115/1075163.1075200.
- [28] N. Ghamrawi und A. McCallum, „Collective Multi-Label Classification“, in *Proceedings of the University of Massachusetts Amherst*, Amherst, Massachusetts, USA: University of Massachusetts Amherst, 2005. Zugegriffen: 11. März 2025. [Online]. Verfügbar unter: <https://scholarworks.umass.edu/server/api/core/bitstreams/ee4f8c19-e9e4-4a0f-bb2a-669cdfe09706/content>
- [29] D. Li *u. a.*, „From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge“, *arXiv preprint*, 2024, Zugegriffen: 12. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2411.16594>

- [30] F. Martínez-Plumed *u. a.*, „CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories“, *IEEE Transactions on Knowledge and Data Engineering*, 2020, Zugegriffen: 4. März 2025. [Online]. Verfügbar unter: https://research-information.bris.ac.uk/ws/portalfiles/portal/220614618/TKDE_Data_Science_Trajectories_PF.pdf
- [31] P. Chapman *u. a.*, „CRISP-DM 1.0: Step-by-step Data Mining Guide“. 2000. Zugegriffen: 3. März 2025. [Online]. Verfügbar unter: <https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf>
- [32] IBM, „IBM SPSS Modeler CRISP-DM Guide“. 2023. Zugegriffen: 4. März 2025. [Online]. Verfügbar unter: https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPD.pdf
- [33] D. Dean, „Advanced Data Analytics for Organizations“. Zugegriffen: 4. März 2025. [Online]. Verfügbar unter: <https://app.myeducator.com/reader/web/1421a>
- [34] J. Baijens, R. Helms, und D. Iren, „Applying Scrum in Data Science Projects“, in *2020 IEEE 22nd Conference on Business Informatics (CBI)*, Antwerp, Belgium: IEEE, Juni 2020, S. 29–38. doi: 10.1109/CBI49978.2020.00011.
- [35] W. X. Zhao *u. a.*, „A Survey of Large Language Models“, *arXiv*, 2023, Zugegriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2303.18223>

- [36] R. Bommasani *u. a.*, „On the Opportunities and Risks of Foundation Models“, *arXiv*, 2021, Zugegriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2108.07258>
- [37] W. Du, Y. Yang, und S. Welleck, „Optimizing Temperature for Language Models with Multi-Sample Inference“, *arXiv preprint*, 2025, Zugegriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2502.05234>
- [38] A. Lapedes und R. Farber, „How Neural Nets Work“, in *Proceedings of the Neural Information Processing Systems (NeurIPS)*, Los Alamos, NM, USA: American Institute of Physics, 1988. Zugegriffen: 31. März 2025. [Online]. Verfügbar unter: https://proceedings.neurips.cc/paper_files/paper/1987/file/09c653c3ae9d116e5f288ff988283a06-Paper.pdf
- [39] Y. Hu *u. a.*, „Information Extraction from Clinical Notes: Are We Ready to Switch to Large Language Models?“, *arXiv preprint arXiv:2411.10020*, 2024, Zugegriffen: 12. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2411.10020>
- [40] C. Raffel *u. a.*, „Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer“, *arXiv preprint*, 2019, [Online]. Verfügbar unter: <https://doi.org/10.48550/arXiv.1910.10683>
- [41] U. Kamath, K. Keenan, G. Somers, und S. Sorenson, *Large Language Models: A Deep Dive - Bridging Theory and Practice*. Cham, Switzerland: Springer Nature Switzerland AG, 2024.
- [42] S. Wadhwa, S. Amir, und B. C. Wallace, „Revisiting Relation Extraction in the Era of Large Language Models“, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Volume 1)*:

- Long Papers*), Toronto, Canada: Association for Computational Linguistics, 2023, S. 15482–15495. [Online]. Verfügbar unter: <https://aclanthology.org/2023.acl-long.868>
- [43] X. Tang, J. Wang, und Q. Su, „Small Language Model Is a Good Guide for Large Language Model in Chinese Entity Relation Extraction“, *arXiv preprint*, 2024, Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/html/2402.14373v1>
- [44] Y. Ge, W. Hua, J. Ji, J. Tan, S. Xu, und Y. Zhang, „OpenAGI: When LLM Meets Domain Experts“, *arXiv*, 2023, Zugegriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2304.04370>
- [45] M. Cheung, „A Reality Check of the Benefits of LLM in Business“, *arXiv*, 2024, Zugegriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2406.10249>
- [46] J. Wu *u. a.*, „Can Large Language Models Understand Uncommon Meanings of Common Words?“, *Preprint*, 2025, Zugegriffen: 1. April 2025. [Online]. Verfügbar unter: <https://scispace.com/pdf/can-large-language-models-understand-uncommon-meanings-of-13x88kv0gx.pdf>
- [47] T. B. Brown *u. a.*, „Language Models are Few-Shot Learners“, *Advances in neural information processing systems*, Bd. 33, S. 1877–1901, 2020, Zugegriffen: 13. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2005.14165>
- [48] J. Devlin, M. Wei, C. Kenton, und L. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, in *Proceedings of NAACL-HLT 2019, North American Chapter of the Association for*

Computational Linguistics, 2019, S. 4171–4186. Zugegriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://aclanthology.org/N19-1423.pdf>

- [49] P. Sawicki, M. Grzes, D. Brown, und F. Góes, „Can Large Language Models Outperform Non-Experts in Poetry Evaluation? A Comparative Study Using the Consensual Assessment Technique“, *arXiv preprint*, Feb. 2025, Zugegriffen: 1. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/pdf/2502.19064.pdf>
- [50] N. Wan *u. a.*, „Humans and Large Language Models in Clinical Decision Support: A Study with Medical Calculators“, *arXiv preprint*, 2025, doi: 10.48550/arXiv.2411.05897.
- [51] D. V. Veen *u. a.*, „Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization“, *arXiv preprint*, 2023, [Online]. Verfügbar unter: <https://arxiv.org/pdf/2309.07430.pdf>
- [52] OpenAI, „GPT-4 Technical Report“, *arXiv preprint arXiv:2303.08774*, 2023, Zugegriffen: 27. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/pdf/2303.08774.pdf>
- [53] Z. Ji *u. a.*, „Survey of Hallucination in Natural Language Generation“, *ACM Computing Surveys*, Bd. 55, Nr. 12, S. 1–38, 2023, Zugegriffen: 27. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2202.03629>
- [54] S. M. T. I. Tonmoy *u. a.*, „A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models“, *arXiv preprint*, 2024, Zugegriffen: 26. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2401.01313>

- [55] S. Shekhar, T. Dubey, K. Mukherjee, A. Saxena, A. Tyagi, und N. Kotla, „Towards Optimizing the Costs of LLM Usage“, *arXiv*, 2024, Zugegriffen: 27. Februar 2025. [Online]. Verfügbar unter: <https://doi.org/10.48550/arXiv.2402.01742>
- [56] OpenAI, „OpenAI API Pricing“. [Online]. Verfügbar unter: <https://openai.com/api/pricing/>
- [57] J. Yang *u. a.*, „Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond“, *arXiv*, 2023, Zugegriffen: 27. Februar 2025. [Online]. Verfügbar unter: <https://doi.org/10.48550/arXiv.2304.13712>
- [58] B. Waltl, G. Bonczek, und F. Matthes, „Rule-Based Information Extraction: Advantages, Limitations, and Perspectives“, *Technical University of Munich, Department of Informatics*, 2017, Zugegriffen: 27. Februar 2025. [Online]. Verfügbar unter: <https://wwwmatthes.in.tum.de/pages/1w12fy78ghug5/Rule-based-Information-Extraction-Advantages-Limitations-and-Perspectives>
- [59] H. Wang *u. a.*, „A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy“, *arXiv preprint arXiv:2501.09431*, 2024.
- [60] „Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)“, *Official Journal of the European Union*, S. 1–88, 2016, Zugegriffen: 11. März 2025. [Online]. Verfügbar unter: <http://data.europa.eu/eli/reg/2016/679/oj>

- [61] T. Aguda u. a., „Large Language Models as Financial Data Annotators: A Study on Effectiveness and Efficiency“, *arXiv preprint arXiv:2403.18152*, 2024, Zugegriffen: 27. Februar 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2411.10020>
- [62] S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, und et al., „Opportunities and challenges for ChatGPT and large language models in biomedicine and health“, *Briefings in Bioinformatics*, Bd. 25, Nr. 1, S. bbad493, 2024, doi: 10.1093/bib/bbad493.
- [63] D. Xu, W. Chen, W. Peng, C. Zhang, Y. Zheng, und Y. Wang, „Large Language Models for Generative Information Extraction: A Survey“, *arXiv preprint arXiv:2312.17617*, 2023.
- [64] A. Roth, „How SAP’s Generative AI Hub facilitates embedded, trustworthy, and reliable AI“, Feb. 2024, Zugegriffen: 6. März 2025. [Online]. Verfügbar unter: <https://community.sap.com/t5/technology-blogs-by-sap/how-sap-s-generative-ai-hub-facilitates-embedded-trustworthy-and-reliable/ba-p/13596153>
- [65] SAP SE, „SAP AI Launchpad“. März 2025. Zugegriffen: 6. März 2025. [Online]. Verfügbar unter: <https://help.sap.com/doc/5945759df2d34b69b681c53bb2dd7b9f/CLOUD/en-US/038a6194f65c4ef68885f6f16360dbc4.pdf>
- [66] Y. Li, „From Unstructured Input to Structured Output: LLM meets SAP“. Zugegriffen: 17. April 2025. [Online]. Verfügbar unter: <https://community.sap.com/t5/artificial-intelligence-and-machine-learning-blogs/from-unstructured-input-to-structured-output-llm-meets-sap/ba-p/13772506>

- [67] OpenAI, „Structured Outputs“. [Online]. Verfügbar unter: <https://platform.openai.com/docs/guides/structured-outputs?api-mode=responses>
- [68] P. Herzig, „How SAP's Generative AI Architecture Redefines Business Applications“, Dez. 2023, Zugegriffen: 6. März 2025. [Online]. Verfügbar unter: <https://community.sap.com/t5/technology-blogs-by-sap/how-sap-s-generative-ai-architecture-redefines-business-applications/ba-p/13580679>
- [69] SAP SE, „SAP Roadmap 2025“. Zugegriffen: 6. März 2025. [Online]. Verfügbar unter: https://roadmaps.sap.com/board?range=FIRST-LAST&FT=GEN_AI#Q4%202024
- [70] SAP, „Multi Experience Platform (MXP) Overview“. [Online]. Verfügbar unter: [https://sap.sharepoint.com/sites/206579/SitePages/Multi%20Experience%20Platform%20\(MXP\)%20Overview.aspx](https://sap.sharepoint.com/sites/206579/SitePages/Multi%20Experience%20Platform%20(MXP)%20Overview.aspx)
- [71] SAP, „Solution Overview: Multi Experience Platform“. [Online]. Verfügbar unter: https://sap.sharepoint.com/:b/r/teams/MXPCommunity/Shared%20Documents/General/Slides%20%26%20General%20Information/20240605_MXP%20Solution%20Overview.pdf?csf=1&web=1&e=cNPb9h
- [72] Y. Chang u. a., „A Survey on Evaluation of Large Language Models“, *ACM Transactions on Intelligent Systems and Technology*, Bd. 15, Nr. 3, März 2024, Zugegriffen: 13. März 2025. [Online]. Verfügbar unter: <https://doi.org/10.1145/3641289>
- [73] B. Bala und S. Behal, „A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques“, in *Proceedings of the 2024 8th*

- International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, IEEE, 2024, S. 1755–1762. Zugegriffen: 12. März 2025. [Online]. Verfügbar unter: <https://ieeexplore.ieee.org/document/10714767>
- [74] J. Kaplan *u. a.*, „Scaling Laws for Neural Language Models“, *arXiv preprint arXiv:2001.08361*, 2020, Zugegriffen: 14. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2001.08361>
- [75] J. Wei *u. a.*, „Chain-of-Thought Prompting Elicits Reasoning in Large Language Models“, *arXiv preprint*, 2022, Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/pdf/2201.11903>
- [76] Q. Xu, F. Hong, B. Li, C. Hu, Z. Chen, und J. Zhang, „On the Tool Manipulation Capability of Open-source Large Language Models“, *arXiv preprint arXiv:2305.16504*, 2023, Zugegriffen: 14. April 2025. [Online]. Verfügbar unter: <https://doi.org/10.48550/arXiv.2305.16504>
- [77] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, und A. Chadha, „A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications“, *arXiv preprint*, Feb. 2024, Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/pdf/2402.07927>
- [78] Z. A. Nazi, M. R. Hossain, und F. A. Mamun, „Evaluation of Open and Closed-Source LLMs for Low-Resource Language with Zero-Shot, Few-Shot, and Chain-of-Thought Prompting“, *Natural Language Processing Journal*, Bd. 10, S. 100124, 2025, Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: https://www.sciencedirect.com/science/article/pii/S2949719124000724?ref=pdf_download&fr=RR-2&rr=926fbc7d8b9592b9

- [79] F. Polat, I. Tiddi, und P. Groth, „Testing Prompt Engineering Methods for Knowledge Extraction from Text“, *Semantic Web*, 2024, Zugegriffen: 22. April 2025. [Online]. Verfügbar unter: <https://journals.sagepub.com/doi/full/10.3233/SW-243719>
- [80] Q. Ye, M. Axmed, R. Pryzant, und F. Khani, „Prompt Engineering a Prompt Engineer“. Zugegriffen: 17. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2311.05661>
- [81] OpenAI, „Prompt Engineering“. Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: <https://platform.openai.com/docs/guides/prompt-engineering>
- [82] Microsoft, „System Message Design“. Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/advanced-prompt-engineering>
- [83] Microsoft, „Prompt Engineering techniques“. Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/prompt-engineering>
- [84] S. Min, X. Lyu, A. Holtzman, M. Artetxe, H. Hajishirzi, und L. Zettlemoyer, „Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?“, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, S. 727–740. Zugegriffen: 16. April 2025. [Online]. Verfügbar unter: <https://aclanthology.org/2022.acl-long.8.pdf>
- [85] X. Wang u. a., „Self-Consistency Improves Chain of Thought Reasoning in Language Models“, in *Proceedings of the International Conference on*

Learning Representations (ICLR), 2023. Zugegriffen: 27. März 2025.
[Online]. Verfügbar unter: <https://arxiv.org/pdf/2203.11171>

- [86] Y. Zhou *u. a.*, „Large Language Models Are Human-Level Prompt Engineers“, *arXiv preprint*, Nov. 2022, Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/pdf/2211.01910>
- [87] C. Li, D. Zhang, und J. Wang, „LLM-assisted Labeling Function Generation for Semantic Type Detection“, *arXiv preprint*, 2024, Zugegriffen: 31. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2408.16173>
- [88] R. M. Rifkin und A. Klautau, „In Defense of One-Vs-All Classification“, *Journal of Machine Learning Research*, Bd. 5, S. 101–141, 2004.
- [89] SAP, „Business AI - Adoption Overview“. Zugegriffen: 2. März 2025. [Online]. Verfügbar unter: https://launcher.value-experience-hub.for.sap/experiences/business-ai%E2%80%94adoption-overview/groups/live/pages/detailspage?nf-model-version=latest&selectedTab=Tab-1&account_id=0005482525&opportunity_id=0305420457&material_id=8018592
- [90] R. Kohavi, „A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection“, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montréal, Québec, Canada: Morgan Kaufmann, 1995, S. 1137–1143. [Online]. Verfügbar unter: <https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf>
- [91] R. Figueiroa, Q. Zeng-Treitler, S. Kandula, und L. Ngo, „Predicting sample size required for classification performance“, *BMC Medical Informatics and*

Decision Making, Bd. 12, 2012, [Online]. Verfügbar unter: <https://doi.org/10.1186/1472-6947-12-8>

- [92] Microsoft, „Microsoft Graph REST API v1.0 endpoint reference“. Zugegriffen: 5. April 2025. [Online]. Verfügbar unter: <https://learn.microsoft.com/en-us/graph/api/overview?view=graph-rest-1.0>
- [93] S. Salman und X. Liu, „Overfitting Mechanism and Avoidance in Deep Neural Networks“. Zugegriffen: 19. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/1901.06566>
- [94] Q. Huang und T. Zhao, „Data Collection and Labeling Techniques for Machine Learning“. Zugegriffen: 28. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2407.12793>
- [95] A. Kirkovska und A. Sharma, „Analysis: OpenAI o1 vs GPT-4o vs Claude 3.5 Sonnet“. Zugegriffen: 17. April 2025. [Online]. Verfügbar unter: <https://www.vellum.ai/blog/analysis-openai-o1-vs-gpt-4o>
- [96] OpenAI *u. a.*, „OpenAI o1 System Card“. Zugegriffen: 17. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2412.16720>
- [97] OpenAI, „GPT-4o“. [Online]. Verfügbar unter: <https://platform.openai.com/docs/models/gpt-4o>
- [98] J. B. Balasubramanian *u. a.*, „Leveraging large language models for structured information extraction from pathology reports“. Zugegriffen: 17. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2502.12183>
- [99] A. Brokman, X. Ai, Y. Jiang, S. Gupta, und R. Kavuluru, „A Benchmark for End-to-End Zero-Shot Biomedical Relation Extraction with LLMs: Ex-

- periments with OpenAI Models“. Zugegriffen: 17. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2504.04083>
- [100] OpenAI, „OpenAI o3-mini“. [Online]. Verfügbar unter: <https://openai.com/index/openai-o3-mini/>
- [101] M. K. Ranjan, K. Barot, V. Khairnar, V. Rawal, und others, „Python: Empowering Data Science Applications and Research“, *Journal of Operating Systems Development & Trends*, Bd. 10, Nr. 1, S. 27–33, Aug. 2023, doi: 10.37591/joosdt.v10i1.576.
- [102] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, und G. Neubig, „Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing“, *arXiv preprint*, 2021, Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: <https://arxiv.org/pdf/2107.13586>
- [103] R. van der Goot, „We Need to Talk About train-dev-test Splits“, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, Association for Computational Linguistics, 2021. Zugegriffen: 19. April 2025. [Online]. Verfügbar unter: <https://aclanthology.org/2021.emnlp-main.368.pdf>
- [104] K. Huang, Y. Jin, R. Li, M. Y. Li, E. Candès, und J. Leskovec, „Automated Hypothesis Validation with Agentic Sequential Falsifications“. [Online]. Verfügbar unter: <https://arxiv.org/abs/2502.09858>
- [105] B. E. Wampold, B. Davis, und R. H. Good, „Hypothesis validity of clinical research“, *Journal of Consulting and Clinical Psychology*, Bd. 58, Nr. 3, S. 360–367, 1990, [Online]. Verfügbar unter: <https://www.researchgate.net/publication/20961145>

- [106] S. C. Loftus, „Chapter 15 - Hypothesis tests for two parameters“, *Basic Statistics with R*. Elsevier, S. 163–185, 2022. [Online]. Verfügbar unter: <https://www.sciencedirect.com/science/article/abs/pii/B9780128207888000286>
- [107] J. Gorri *u. a.*, „A hypothesis-driven method based on machine learning for neuroimaging data analysis“, *Neurocomputing*, Bd. 510, S. 159–171, 2022, Zugegriffen: 29. April 2025. [Online]. Verfügbar unter: <http://dx.doi.org/10.1016/j.neucom.2022.09.001>
- [108] B. Efron und R. J. Tibshirani, *An Introduction to the Bootstrap*. in Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Boca Raton; London; New York; Washington, D.C.: Chapman & Hall/CRC, 1993. Zugegriffen: 21. April 2025. [Online]. Verfügbar unter: <https://www.hms.harvard.edu/bss/neuro/bornlab/nb204/statistics/bootstrap.pdf>
- [109] D. H. Wolpert, „Stacked Generalization“, *Neural Networks*, Bd. 5, Nr. 2, S. 241–259, 1992, Zugegriffen: 12. März 2025. [Online]. Verfügbar unter: https://www.researchgate.net/publication/222467943_Stacked_Generalization
- [110] W. C. Choi und C. I. Chang, „Advantages and Limitations of Open-Source versus Commercial Large Language Models (LLMs): A Comparative Study of DeepSeek and OpenAI's ChatGPT“. [Online]. Verfügbar unter: <https://www.researchgate.net/publication/390313410>
- [111] S. Huang, K. Yang, S. Qi, und R. Wang, „When Large Language Model Meets Optimization“. Zugegriffen: 22. April 2025. [Online]. Verfügbar unter: <https://arxiv.org/abs/2405.10098>

ii. Anhang

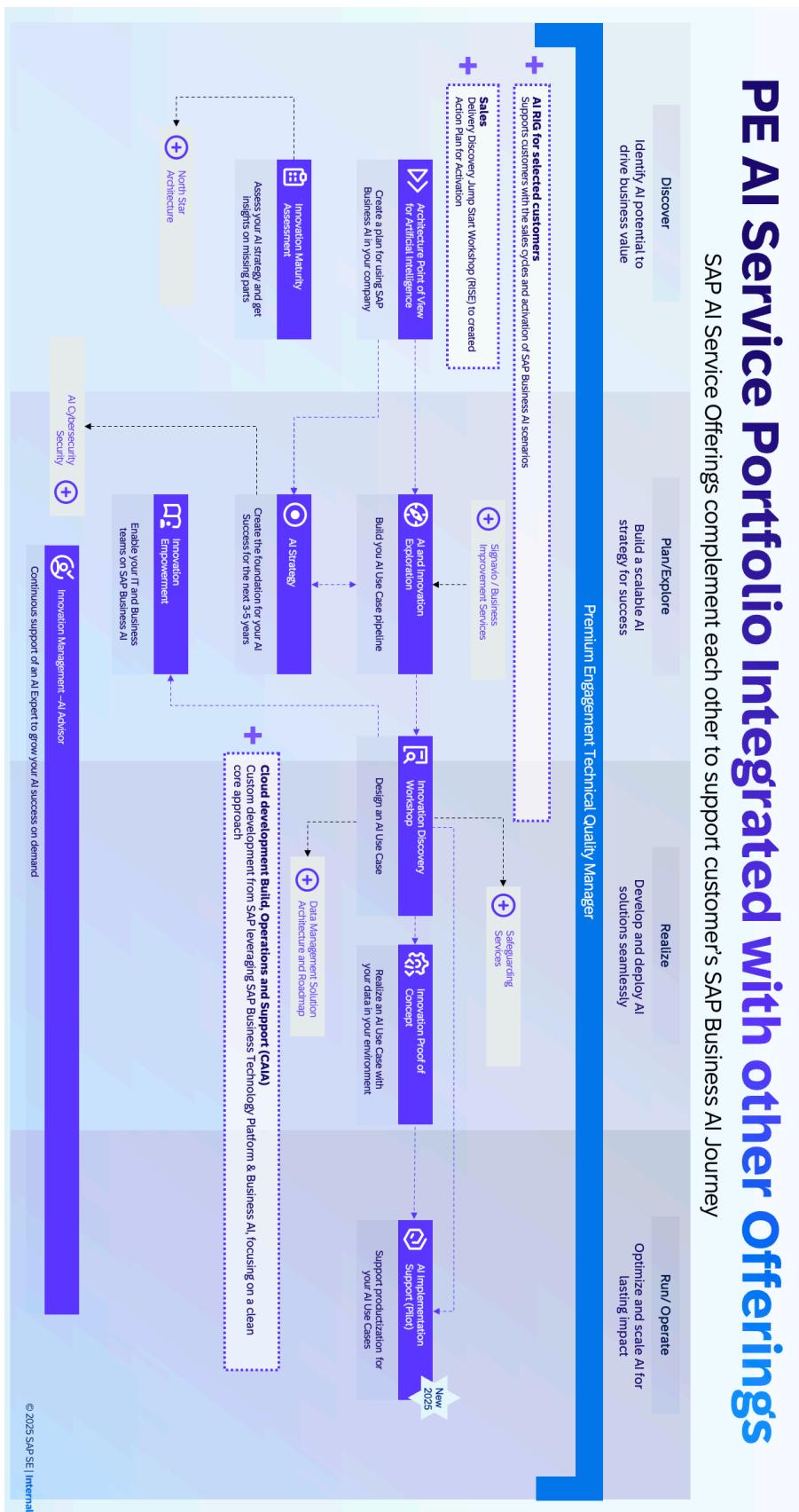


Abbildung 21: SAP AI Services - Aufgaben der RIG [5, S. 8]

| Library | Beschreibung |
|----------------|---|
| json | Verarbeitung von JSON-Daten |
| csv | Lesen und Schreiben von CSV-Dateien |
| numpy | Numerische Berechnungen und Arrays |
| pandas | Datenanalyse und -manipulation |
| matplotlib | Erstellung von Diagrammen und Plots |
| gen_ai_hub | Einbindung des SAP GenAI Hub |
| requests | HTTP-Anfragen senden und empfangen |
| oauthlib | Implementierung von OAuth-Authentifizierung |

Tabelle 3: Verwendete Python Libraries

| Feld | Enum | Beschreibung | Merkmal(e) |
|----------------|------------------------|---|---|
| Customer Name | - | Der Name des Kunden | Teil des Betreffs |
| Customer ID | - | Die ID eines Kunden | 10-stellige Nummer |
| Product Name | siehe Tabelle 6 | Der Name des SAP Business AI Produktes | - |
| Opportunity ID | - | Die ID des Kundenvertrags | 10-stellige Nummer, beginnend mit ,030‘ |
| Status | siehe Tabelle 5 | Der Status der Produktaktivierung | - |
| Analysis | - | Zusammenfassung des Status | - |
| RIG Kontakt | Mitarbeiter der AI RIG | Zuständiger RIG-Berater | Beantwortet eingehende E-Mails |
| On-Hold Datum | - | Datum der Weiterführung der Aktivierung | Nur gesetzt, wenn Status = ,On Hold‘ |

Tabelle 4: Zu extrahierende Datenfelder je Kunde

| Status | Beschreibung |
|------------------------------------|--|
| Preparing Outreach | New opportunity, no reach-out email has been sent yet to account teams (opportunity owners). |
| In Analysis | After the first reach-out email, before we receive the first response/context of the opportunity deal. |
| Waiting For Answer | Awaiting feedback for outreach emails after a reminder was sent. |
| On Hold | Currently not planning to do any AI use case preparation/activation work, due to other priorities/restrictions. |
| In Preparation | After we receive the first feedback from the right contact person; in the process of clarifying/finalizing the use cases. |
| Awareness session / JSD suggested | For RISE/potential RISE customers, if they need support on use case discovery, we connect account teams to the JSD (Jump Start Discovery) team. RIG contact to keep the "Workshop Status" field up to date (not planned, scheduled, in progress, or delivered) in MXP. |
| In Activation | After use cases are identified, start the technical activation (but before the first activation is completed). |
| Activated | First AI (GenAI, Joule, BTP) use case activated in any system. |
| Customer not interested in product | The customer has rejected implementing AI use cases, with no plan to activate any use case. We still suggest exploring possible use cases. |
| Discontinued | The deal was closed without an AI unit or does not apply to this particular account |

Tabelle 5: Anzunehmende Statuswerte einer Aktivierung

| Produktnam | zugehörige Material-IDs |
|--|---------------------------|
| SAP Ariba Category Management | 8015105 |
| SAP Adv VC and Pricing, Commerce, access | 8015476 |
| SAP Adv VC & Pricing, add-on for SAP CPQ | 8015503 |
| SAP Enterprise Service Management | 8015863 |
| RISE wSAP S/4HANA Cld, priv ed, prem pl | 8016421, 8018501 |
| SAP AI Unit | 8016532, 8016551, 8018592 |
| SAP Joule embedded entitlement | 8017178 |
| SAP AI Core extended | 8017491 |
| SAP CX AI Toolkit | 8017592 |
| SAP IPR | 8017891 |
| RISE with SAP S/4HANA Cld, priv ed, prem | 8018418 |
| RISE w SAP S4HANA Cld, priv ed, base | 8018511 |
| Joule limited promotion | 8018808 |

Tabelle 6: Material ID's der SAP Business AI Produkte

| Rolle | Inhalt |
|---------------|--|
| System | You are a helpful email data extraction assistant with strong chain-of-thought reasoning abilities. Your task is to first perform detailed, step-by-step reasoning to analyze and extract key elements from an email. Do not include your reasoning steps in the final output. |
| User | EXTRACT DATA FROM THE FOLLOWING EMAIL USING DETAILED CHAIN-OF-THOUGHT REASONING: Subject: {{?subject}} Body: {{?main_body}} E-Mail-Context: {{?context_body}} Break down the email content step-by-step and then provide only a valid JSON object as a result |

Prompt 2: Struktur des verwendeten Chain-of-Thought Prompts

| Rolle | Inhalt |
|---------------|---|
| System | You are a highly reliable email data extraction assistant, specialized in self-consistency reasoning. For this task, generate multiple independent chains-of-thought to extract the email details. Then, evaluate these reasoning paths internally and converge on the most consistent final answer. |
| User | EXTRACT DATA FROM THE FOLLOWING EMAIL USING SELF-CONSISTENCY TECHNIQUES: Subject: {{?subject}} Body: {{?main_body}} E-Mail-Context: {{?context_body}} Generate several chains-of-thought reasoning paths to process the content, evaluate them and provide only a valid JSON object as a result |

Prompt 3: Struktur des verwendeten Self-consistency Prompts

| Rolle | Inhalt |
|-------------|---|
| User | ... Exmaple: Subject: {{?example_subject}} Body: {{?example_main_body}} E-Mail-Context: {{?example_context_body}} Solution: {{?example_solution}} ... |

Prompt 4: Anhang je Beispiel bei Verwendung von One-/Few-Shot-Prompting