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@dhbw.de +4962141051412

I. Eidesstattliche 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 Zeitrahmen 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

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Promptverzeichnis

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	Künstliche Intelligenz
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	Effektstärke nach Cohen
d	Referenztext
k	Gewicht
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 selbstaufmerksamen neuronalen Netzen ist ein neues Zeitalter der Entwicklung von Künstliche Intelligenz (KI) eingetreten. Diese Entwicklungen führten zu erheblichen Fortschritten in der Integration von KI-Technologien in vielfältige Anwendungsfelder, insbesondere in betriebswirtschaftliche Prozesse [10, S. 8-9]. In der Unternehmenswelt eröffnet der Einsatz von KI signifikante Potenziale zur Effizienzsteigerung und Automatisierung [11].

Die SAP SE nutzt KI in der SAP Business AI Produktreihe, welche KI-gestützte Lösungen bereitstellt, die sich nahtlos in bestehende Geschäftsprozesse einfügen [12]. Ein zentrales Anwendungsfeld stellt die Datenextraktion dar, die strukturierte Informationen automatisiert aus unstrukturierten Datenquellen gewinnt und damit ehemals manuelle Prozesse beschleunigt [13, S. 263-264]. Hierbei ist die SAP bemüht, ihre Prozesse hinsichtlich dieses Anwendungsfelds zu optimieren [14].

Die Aktivierung sowie der Absatz der SAP Business Artificial Intelligence (AI) Produkte in Systemlandschaften von Kunden ist eines der Hauptziele der SAP im Jahr 2025 [15]. Zur Messung der Erreichung dieses Ziels sind Reportingdaten notwendig, welche die Anzahl erfolgreicher Aktivierungen dokumentieren. Durch die steigenden Kaufzahlen von SAP Business AI Produkten ist die AI Regional Implementation Group (RIG), für die Aktivierung von Business AI zuständige Beratungsgruppe der SAP, unter erhöhtem Druck, wodurch eine Dokumentation in das Reporting vernachlässigt wird. Um die RIG zu entlasten, soll ein Ansatz entwickelt werden, welcher auf Basis von E-Mails strukturierte Reportingdaten

durch Large Language Model (LLM)-gestützte Datenextraktion extrahiert und diese automatisiert in das Reporting eingepflegt.

1.2. Ziel und Gang

Ziel dieser Arbeit ist die Optimierung des Reportingprozesses innerhalb der SAP, exemplarisch untersucht am Beispiel der AI RIG. Hierzu wird ein Ansatz zur automatisierten, KI-gestützten Datenextraktion 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 des hohen zeitlichen Aufwands soll ein automatisierter Ansatz erprobt werden, welcher strukturierte Daten extrahiert und den Status von Kundenaktivierungen 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 sollen in dieser Arbeit beantwortet werden:

- **Forschungsfrage 1**: Welche Verfahren der Datenextraktionsoptimierung aus der Literatur zeigen eine Verbesserung der Datenextraktionsgenauigkeit?
- **Forschungsfrage 2**: Wie kann eine erfolgreiche Datenextraktion den zugehörigen Eintrag in der Reporting-Anwendung identifizieren?

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 die Extraktion und Strukturierung interner E-Mail-Kommunikationsdaten zu optimieren, 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, endet der Prozess nach Abschluss der Evaluation.

Zur Datenextraktion werden Pre-trained Language Models (PLMs) verwendet. PLMs, Teilgruppe der Large Language Models (LLMs), übertreffen herkömmliche regelbasierte und statistische Natural Language Processing (NLP)-Ansätze, da sie komplexe Texte sowie kontextabhängige Bedeutungen erfassen können und im Gegensatz zu regelbasierten Verfahren mittels Optimierungsverfahren ihre Extraktionsgenauigkeit weiter gesteigert werden kann [19, S. 11-14], [20, S. 1-4]. In der Medizin findet LLM-basierte Datenextraktion aus Freitexten bereits Anwendung [10, S. 6-9], [21, S. 5-7]. E-Mails weisen, vergleichbar mit medizinischen Dokumenten, sowohl strukturierte als auch unstrukturierte Daten auf.

Zur Optimierung der Datenextraktionsqualität wird ein Experiment auf einem reduzierten Datensatz durchgeführt, um verschiedene Modellierungsansätze zu testen und deren Auswirkung auf die Datenqualität zu messen und zu validieren [22, 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 Datenstruk-

turierung systematisch untersucht werden und die für den Anwendungsfall besten Konfigurationen ausgewählt werden [23, S. 1014-1016].

Zur Evaluation der Ergebnisse werden die Metriken Precision, Recall, F1-Score und Recall-Oriented Understudy for Gisting Evaluation (ROUGE) berechnet [24, S. 6-12]. Diese Variation an Evaluationsmetriken ergibt sich durch die Varietät der zu extrahierenden Felder, da diese sowohl Klassifikationen als auch Freitexte enthalten [25, S. 8], [26, S. 1-2]. Als Referenzquelle werden manuell extrahierte Daten auf einem reduzierten Datensatz genutzt. Alle genannten Metriken sind weit verbreitet und werden bereits bei LLM-gestützten Datenextraktionen und NLP-Aufgaben in Medizin und Wirtschaft verwendet [27, S. 5], [28, 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 20 ersichtlich unterstützt die RIG in der „Discover“ bis hin zur „Realize“ Phase Kollegen aus dem Sales und dem Support.

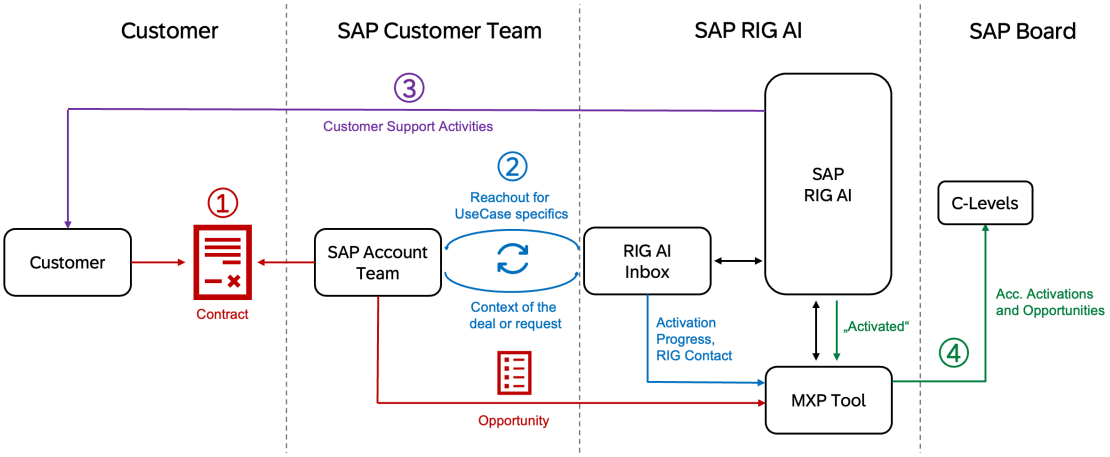

Abbildung 1: Abstrahierte Prozesse in der RIG AI in Anlehnung an [1]

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 das Produkt gekauft 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 erfolgreichem 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 Vorstand zur Verfügung gestellt [1, S. 9-14] [Abbildung 1, Prozess 4].

3.2. Cross Industry Standard Process for Data Mining

CRISP-DM ist ein auf KDD basiertes [29, S. 2] De-facto-Standard Vorgehensmodell, um Data-Science- und Data-Mining-Projekte strukturiert zu planen und durchzuführen [16, S. 529-532]. Der CRISP-DM-Prozess ist industrie- und technologieunabhängig konzipiert und in sechs iterative Phasen unterteilt, die von der geschäftlichen Zieldefinition bis zur finalen Implementierung reichen [30, S. 9-10]. Das Vorgehensmodell gliedert sich in die in Abbildung 2 dargestellten Phasen und wird im Folgenden erläutert:

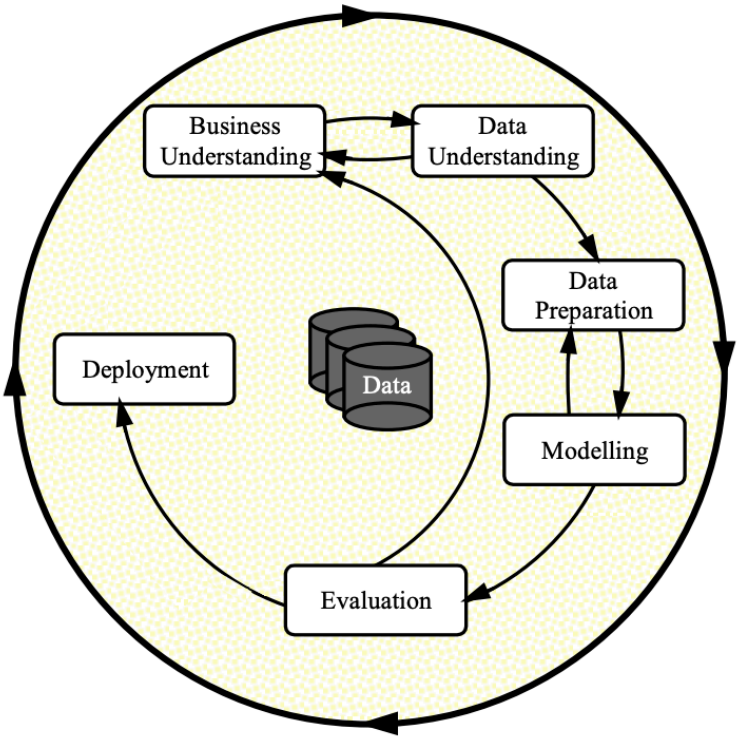

Abbildung 2: Phasen des CRISP-DM Prozess [2]

- **Business Understanding:** In der ersten Phase werden Ziele aus Unternehmensperspektive definiert und die Anforderungen an das Data-Science-Projekt präzisiert [30, S. 16-19]. Ebenfalls wird ein oder mehrere Erfolgskriterien zur Bewertung der Zielerfüllung ausgewählt. [16, S. 527].

- **Data Understanding**: In der zweiten Phase werden die zur Verfügung stehenden Daten gesammelt und analysiert. Datenquellen werden identifiziert, Daten werden gesammelt und erste Analysen durchgeführt, um einen Überblick über Datenqualität, Datenumfang sowie Hinweise zu Datenanalysemethoden zu erlangen. [30, 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 Attrbibuten und Einträgen, Transformation und Konvertierung in ein geeignetes Format. [30, 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.[30, S. 27-29]

- **Evaluation**: In dieser Phase wird das Modelling und die Ergebnisse anhand der Evaluationsmetriken bewertet und evaluiert, inwiefern die Geschäftsziele erreicht sind.[30, S. 30-31].

- **Deployment**: Zum Abschluss erfolgt die Planung und Durchführung der Produktivnahme (Deployment), Projektabschluss und Dokumentation des Modells. Hierbei werden die Verantwortungen nach Ende des Projektes hinsichtlich Wartung und Support festgehalten. [30, S. 32-34]

Die Verwendung von CRISP-DM bringt einige Vorteile mit sich. CRISP-DM erlaubt Iterativität in vorherige Phasen [Abbildung 2], beispielweise durch von den Geschäftszielen abweichende Datengrundlage [31, S. 9], Notwendigkeit weitere Datenvorverarbeitung oder Nichterfüllung der Geschäftsziele in der Evaluation [32, Abs. 2.1]. Erkenntnisse aus den vorherigen Iterationen fließen in den Prozess mit ein und verbessern die Datenqualität und Modellleistung nachhaltig [30, S. 9-10].

Zudem bietet CRISP-DM eine Struktur für Arbeit in einem Team. 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 inhaltlich sinnvoll verankert 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], [33, S. 32]. Gründe dafür sind fehlende Richtlinien, wann eine Iteration in eine vorherige Phase notwendig ist [33, 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], [33, 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 genommen. [16, S. 532-533]

3.3. Large Language Models

Large Language Models (LLMs) sind generative Sprachmodelle, die aus einer Modelleingabe (Prompt) anhand von Wahrscheinlichkeiten eine natürliche Ausgabe produzieren [34, S. 2]. Sie entspringen der Domäne des NLP und basieren in der Regel auf neuronalen Netzen, die auf großen Datensätzen trainiert werden [35, S. 3]. Ein neuronales Netz verbindet Eingabe- und Ausgabedaten über künstliche Neuronen, welche die Aufgabe haben, aus mehreren Eingaben, anhand von Voreingenommenheiten (Bias) aus vorherigen Iterationen, eine Ausgabe zu berechnen. Die Ausgaben werden anschließend von einer Softmax-Funktion, welcher einer Eingabe eine Ausgabewahrscheinlichkeit zuweist [36, S. 1], bewertet und die Antwort (Prediction) mit der höchsten Wahrscheinlichkeit ausgibt. [37, S. 442-443].

Eine Vielzahl der LLMs basieren auf Transformer-Architektur, die mithilfe von mehrschichtigen neuronalen Netzen eine Ausgabe auf Basis des Encoder-Decoder-Prinzip produzieren [9, S. 1-6], [38, 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. [39, S. 4-5], [40, S. 31-32]. Zur Generierung des Ergebnisses werden Relationen zwischen Tokens mittels Selbstaufmerksamkeit (Self-attention) extrahiert [40, 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]. [39, S. 6] Dies ermöglicht den Modell, kontextuelle Beziehungen zwischen allen Tokens der Eingabesequenz sowie gespeicherter Parametertokens fundierte Vorhersagen zu treffen, wodurch Transformer regelbasierte Verfahren in Genauigkeit und Effizienz übertreffen [39, S. 6], [40, S. 34].

LLMs grenzen sich gegenüber anderen Sprachmodellen besonders durch ihre Modellgröße und umfangreiches Training ab, wodurch sie eine hohe Leistung in diversen Anwendungen aufweisen [34, S. 2-3]. Es wird differenziert zwischen LLMs und PLMs. PLMs bilden eine Untermenge 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. [41, Abs. 1]. In der Literatur werden diese Begriffe häufig synonym verwendet, weswegen in dieser Arbeit die allgemeine Bezeichnung „LLM“ verwendet wird [41, Abs. 1]. [42, 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 [43, S. 1-2]. Aufgrund ihrer Transformer-Architektur ist es LLMs möglich, die Relationen zwischen Sätzen aufzufassen, wodurch sie in der Lage sind, komplexe Satzstrukturen aufzufassen und eine korrekte Ausgabe zu produzieren [44, 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 [45, S. 1]. Im Gegensatz zu regelbasierten NLP-Verfahren können LLMs Daten zum kontinuierlichen Lernen verwenden, wodurch ihre Antwortqualität kontinuierlich steigt [46, 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. [38, S. 5-8], [47, S. 8]. Im medizinischen Umfeld werden Modelle bereits seit langem eingesetzt und übertreffen die Arbeit von Nicht-Experten in diversen Aufgabenstellungen [48, S. 4-8].

Ebenso erlauben LLMs die Automatisierung diverser Prozesse, beispielweise in Unternehmen und der Forschung in Aufgaben wie Textgenerierung oder Zusammenfassung von Informationen [49, 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 [50, S. 4], [51, S. 1-3]. Diese können durch diverse Faktoren auftreten. Darunter gehören Fehlberechnungen statistischer Annahmen, unvollständiges Modelltraining oder die Hinzugabe eines zu geringen Kontext in das Modell [50, S. 5]. Mit diversen Methoden können minimiert, aber nicht vermieden werden [51, S. 11-12]

Einschränkungen treten bei LLMs hinsichtlich Kosten, Zeit und Daten auf [52, 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 tokenbasierte Abreechnung an, bei der die Kosten auf Basis der Nutzung kalkuliert werden [52, S. 1-3], [53]. Durch die Verwendung von Application Programming

Interface (API)-Schnittstellen zwischen Modellanbieter und Anwender kommt es zu Latenzen zwischen Eingabe und Modellausgabe [54, S. 15-16]. [55, 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 [56, S. 1:13], wodurch bei fehlender Zustimmung durch Herausgabe an Externe dem LLM-Nutzer Schaden drohen kann [57].

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 [38, S. 5-8], [58, S. 4-5].

In der Medizin werden bereits LLMs zur Extraktion von Daten aus unstrukturierten Daten genutzt [25, S. 7-12], [59, S. 4-7], [60, S. 5167-5174] und übertreffen bei der Extraktion klinischer Informationen aus Freitext klassische NLP-Verfahren [38, S. 5-8]. In der Wirtschaft steigern LLMs in einer Studie zur Extraktion von Finanzdokumenten die Leistung gegenüber nicht-expertengestützten Verfahren um bis zu 29% [58, 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 sensitiver Geschäftsdaten als zentrales Problem angesehen. [58, S. 7-11], [61, S. 3]. Darüber hinaus ist die Verwendung von LLMs gegenüber regelbasierten NLP-Exraktionsverfahren mit

erhöhtem Zeit- und Ressourcenaufwand verbunden, weswegen sie im geschäftlichem Kontext oft nicht in Betracht gezogen werden [52, S. 1], [54, 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 [62]. Der SAP AI Core Service fungiert als zentrale AI Laufzeitumgebung innerhalb der SAP Business Technology Platform, die zugrunde liegende Cloud-Plattform zur Entwicklung, Integration und Erweiterung von SAP Anwendungen. [63, 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. [62], [63, 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 mittels einer einheitlichen API-Schnittstelle 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 eine Kodierung von vordefinierten Daten bei der Eingabe in ein Modell [63, 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) zur Verfügung, wobei diese sich bisher nur auf wenige Modelle begrenzt [64], [65]. Diese Fähigkeit ist essenziell für Datenextraktionsaufgaben, damit Daten einheitlich strukturiert vorliegen für weitere Verarbeitung.

Alle im GenAI Hub verfügbaren Modelle sind konform mit den Geschäftsbedingungen der SAP SE hinsichtlich Datenschutz und Sicherheit [66]. Im Laufe des Jahres 2025 sollen weitere Modelle zum SAP GenAI Hub hinzugefügt werden und die Modellauswahl von Structured Outputs erweitert werden [67]. Im Hinblick auf die Ziele der SAP im Jahr 2025 wird der AI Core in folgenden Jahren weiter ausgebaut werden, womit er als zu verwendendes Tool zunehmend attraktiver wird [15].

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 die Integration mehrerer Datenquellen, deren Anreicherung mit eigenen Daten und anschließende Visualisierung der Daten [4, Abs. Introduction], [68]. MXP Applikationen eignen sich für dateninten-sive Anwendungen, beispielweise zur Visualisierung von Kennzahlen oder zum Reporting [4, Abs. Experience Building & Designing].

Die MXP nutzt Augmented Content Management Systems (aCMSs), um Aufwand bei der Integration mehrerer Datenquellen zur Applikationsentwicklung zu mindern. aCMS ist ein Datenverwaltungssystem, welches mittels einer Augmentierungsschicht 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 manipulieren. [4, Abs. Content Management]

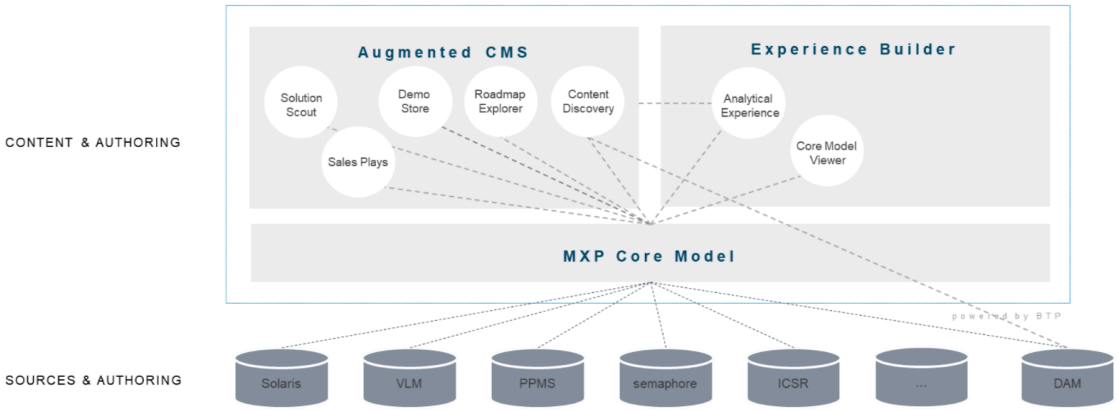

Abbildung 4: Architektur einer MXP Applikation [4]

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 Maximierung der Extraktionsgenauigkeit und damit verbundene Verbesserung der Datenqualität in Bezug auf ein Optimierungsziel [13, S. 279-281]. Verfahren zur Optimierung der Datenextraktionsgenauigkeit umfassen Vorverarbeitung und Normalisierung von Daten sowie eine Optimierung der LLM-Eingabe und -Modellwahl [7, S. 1-2], [60, S. 5165-5165].

3.5.1. Modellwahl

Die Wahl des LLMs ist maßgeblich für die Qualität der Datenextraktion [60, S. 5165-5165]. LLMs unterscheiden sich in Kriterien wie Parametergröße und Robustheit gegenüber Noise zueinander, wodurch ihre Performance stark an den Anwendungsfall gekoppelt ist [69, S. 39:6-39:12]. 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 [60, S. 5167-5171]. Diese Kriterien umfassen:

- **Modellgröße**: Modelle mit geringer Parametergröße sind meist schneller und günstiger als große Modelle, wobei große Modelle komplexe Aufgaben ohne Vortraining besser absolvieren können [70, S. 12-14].

- **Robustheit**: Robuste Modelle sind performanter bei erhöhter Komplexität der Anfrage, wobei sie meist durch die Anwendung von Reasoning ressourcenintensiv arbeiten [71, S. 4-6].
- **Verfügbarkeit**: Durch externe Faktoren wie Unternehmensrichtlinien, Lizenzen oder notwendige Tools (bspw. OpenAIs Structured -Outputs [65]) ist die Wahl des Modells oft eingeschränkt, wodurch nur eine limitierte Auswahl an Modellen zur Verfügung steht [72, S. 1].

3.5.2. Prompt Engineering

Prompt Engineering bezeichnet die gezielte Formulierung von Prompts, um die Qualität von LLM-Ausgaben zu optimieren [73, S. 1]. Dies kann durch die Anwendung von Prompting-Templates und -Techniken umgesetzt werden [73, S. 2-10]. Prompt Engineering zeichnet sich durch seine simplifizierte Umsetzbarkeit und Effektivität bei der Optimierung von PLMs-Outputs aus [73, S. 1], [74, S. 7-11].

Ein Prompt enthält Textsegmenten wie Aufgabenstellungen, Kontext oder Beispiele, die das LLM zur Generierung einer Antwort benötigt [75, S. 3-4]. Gute Prompts zeichnen sich sich durch eine präzise Zieldefinition, eine klare Struktur und kontextbezogene Informationen aus [76, S. 1-2]. Um die Gewichtung dieser durch das Modell zu kontrollieren, kann einem Textsegment eine Rolle zugewiesen werden [77]. 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 [78].
- **User Message**: Enthält die zu bewältigende Aufgabe [77].

Prompt Engineering nutzt diese Eigenschaften aus, um dem Modell gezielte Vorgaben zu geben, wie eine Aufgabe zu bewältigen ist [79]. Im Folgenden werden Prompting Techniken vorgestellt:

- **Zero-Shot Prompting**: Das LLM erhält ausschließlich eine Aufgabe, ohne die Hinzugabe weiterer Daten [73, S. 3]. Dies erlaubt es, LLMs unmittelbar einzusetzen, bedingt jedoch oftmals eine geringere Leistung bei komplexen Fragestellungen [74, S. 7-11].

- **One/Few-Shot Prompting**: Das LLM erhält eine Aufgabe mit Hinzugabe von einem oder mehrerer 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 [73, S. 2]. Mit steigender Beispielanzahl wächst die Genauigkeit der Extraktion [45, S. 20-21], wobei der Zuwachs der Genauigkeit ab vier bis acht Beispielen abflacht und nur geringfügig eine Verbesserung erzielt [80]

- **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 [75, S. 7].[73, 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 betrachten [75, S. 8-9]. Dadurch validiert sich das Modell selbst, wodurch Halluzinationen minimiert und die Quali-

tät der Modellausgabe verbessert gegenüber derer von CoT-Prompting in Datenextraktionsaufgaben ist [81, S. 5-8].

Die praktische Umsetzung vergangener Anwendungsfälle zeigt, dass durch die Kombination mehrerer Prompting-Techniken, die Qualität der Modellausgabe erheblich gesteigert werden kann [73, S. 2], [75, S. 14-23]. Ebenso übertreffen LLMs generierte Prompts manuell erstellte Prompts in Leistung, weswegen die in dieser Arbeit behandelten Prompts mithilfe von AI verbessert werden [82, 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 Teilen sich in Klassifikationsmetriken und Token-Similarity-Metriken. [24, S. 6-7]

Bei der Klassifizierung werden Daten in eine oder mehrere Klassen zugeordnet und anahnd einer Referenzklassifizierung evaluiert. [24, 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.

Bei der Verarbeitung im NLP Umfeld ist eine Klassifizierung von Texten schwierig, weswegen Token-Similarity-Metriken existieren. Sie evaluieren Wortsequenzen. Dafür werden Metriken wie ROUGE eingesetzt, die eine Wortsequenz gegenüber einer Referenzquelle evaluiert und anhand der Überschneidung einen Wert bestimmt [26, 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 [83, Abs. 1].

3.6.1. Precision, Recall und F1-Score

Precision, Recall und F1-Score sind drei weit verbreitete Metriken zur Evaluation von Klassifikationsaufgaben [27, S. 5]. Sie werden oft in Betracht gezogen zur Bewertung von LLM Klassifizierungsausgaben [10, S. 7-9], [21, S. 6], [25, S. 8].

Sei L die Menge an Klassen, $P \subset L$ die Menge an vorhergesagten Klassen und $H \subset L$ die Menge an tatsächlichen Klassen. Precision, Recall und F1-Score basieren auf der Mächigkeit der Mengen True Positive (TP), True Negative (TN), False Positive (FP) und False Negative (FN), welche die Anzahl der korrekten bzw. inkorrekten Klassen durch ein Klassifikationsmodell widerspiegeln [Tabelle 1] [24, S. 7-9].

	Actual Positive	**Actual Negative**
Predicted Positive	$\mathrm{TP} = H \cap P$	$\mathrm{TN} = L \setminus (H \cup P)$
Predicted Negative	$\mathrm{FP} = P \setminus H$	$\mathrm{FN} = H \setminus P$

Tabelle 1: Confusion Matrix für eindimensionale Klassifizierung [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: [24, S. 8]

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}|} \tag{1}$$

Recall, auch bekannt als „True Positive Rate“ oder „Sensitivity“, beschreibt, wie vollständig relevante Ergebnisse erkannt wurden. Sie bestimmt, inwiefern ein LLM positive Ergebnisse identifiziert. Die Formel für Recall lautet: [24, S. 7-8]

$$\text{Recall} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \tag{2}$$

Der F1-Score bildet das harmonische Mittel zwischen Precision und Recall. Die Formel für den F1-Score lautet: [24, S. 8]

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

Die dargestellten Formeln beziehen sich auf den mengenbasierten F1-Score. In der Literatur gibt es diverse Abwandlungen, die ihn einsetzbar in binärer Klasisfikation sowie der Evaluation von Texten machen. Eine textbasierte Evaluation anhand des F1-Scores wird in Abschnitt 3.6.3 vorgestellt. [24, S. 7-14]

3.6.2. Accuracy

Accuracy ist ein Maß zur Bewertung von Klassifikationsaufgaben. Sie misst den Anteil der korrekt klassifizierten Instanzen von allen Instanzen eines Datensatzes. [24, S. 7]

$$\text{Accuracy} = \frac{|\text{TP}| + |\text{TN}|}{|\text{TP}| + |\text{TN}| + |\text{FP}| + |\text{FN}|} = \frac{|\text{TP}| + |\text{TN}|}{|L|} \tag{4}$$

Die Accuracy kann tokenbasiert auf multiple Klassifikationen und Textevaluation berechnet werden. Hier wird die Genauigkeit je „Token“ berechnet und eine Überschneidung von generiertem und Referenztext gemessen. [24, S. 7-9]

3.6.3. Recall-Oriented Understudy for Gisting Evaluation

ROUGE ist eine Evaluationsmetrik des NLP, welche zur Evaluation von generierten Texten gegenüber einer referenzquelle verwendet wird [24, S. 11-12]. ROUGE-n, eine von vielen Umsetzungen von ROUGE, misst den Überlappungsgrad zwischen einem generiertem Text zu einer Referenz mittels N-Grammen (Wortfolgen bestimmter Länge) [26, S. 1-2]. Damit ist die Berechnung von ROUGE-n wie diese eines F1-Scores, wobei anstatt korrekter Klassifikation bei ROUGE die Übereinstimmung zweier Texte auf Basis von n-Grammen betrachtet wird. [24, S. 14]

Sei $b = (\omega_1, \omega_2, ..., \omega_N)$ ein vorhergesagter Text aus N Worten (ω) und $d = (\omega_1, \omega_2, ..., \omega_M)$ ein Referenztext aus M Worten. 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. Damit ergeben sich die folgenden Formeln [24, S. 14]:

$$\text{Precision}_n = \frac{|D \cap B|}{|B|} \tag{5}$$

$$\text{Recall}_n = \frac{|D \cap B|}{|D|} \tag{6}$$

$$\text{ROUGE-n} = 2 \cdot \frac{\text{Precision}_n \cdot \text{Recall}_n}{\text{Precision}_n + \text{Recall}_n} \tag{7}$$

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 das Board 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]. Status Quo werden diese Daten manuell über eine Experience der RIG-MXP je Opportunity eingepflegt [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 aggregierte Daten aller Opportunities der Periode, um fundierte Entscheidungen hinsichtlich der SAP AI Strategie zu fällen [84].

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 [84]:

- **CRM-Account-ID**: Die ID eines Kunden bzw. Customer Relationship Management (CRM)-Accounts im internen SAP System
- **Opportunity-ID**: Die ID einer Opportunity, welche den Vertragsbestandteil abbildet.
- **Material-ID**: Die im System hinterlegte Produktnummer des gekauften SAP Produktes. Zu einem Produkt können aufgrund rechtlicher Bestimmungen mehrere Material-IDs existieren.

Sind diese Daten in der E-Mail nicht gesetzt, wird versucht, über einen Sekundärschlüssel die zugehörige Opportunity zu identifizieren. Zu diesen Feldern gehören:

- **Kundenname**: Der Name des vom AE betreuten Kunden.
- **Produktname**: 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. 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-Contact**: 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 zur zugehörigen Opportunity auf der MXP. 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 eine halluzinationsfreie Extraktion von Reporting- und Zuordnungsdaten. Von der RIG zu erreichende Zahlen sind hierbei eine Extraktionsgenauigkeit über alle Felder von 70% sowie eine Identifikationsratenabweichung von der manuellen Zuordnung von maximal 10%. Darüber hinaus darf die falsch-positiv Rate an identifizierten MXP-Opportunities nicht 5% übersteigen.

Um dieses Ziel zu erreichen, wird ein Experiment durchgeführt, welches die Datenextraktionsrate und -qualität mittels der Optimierungsverfahren aus Abschnitt 3.5 verbessern soll. Folgende Hypothesen werden an das Experiment gestellt:

- **Hypothese 1**: Der Einsatz von One-Shot Prompting verbessert die Datenexraktionsgenauigkeit gegenüber Zero-Shot Prompting.
- **Hypothese 2**: Few-Shot Prompting führt zu höherer Datenextraktionsgenauigkeit gegenüber Zero- und One-Shot Prompting.
- **Hypothese 3**: Die Anwendung Chain-of-Thought Prompting extrahiert qualitativ bessere Daten als ein nicht optimierter, natürlichsprachiger Prompt.

- **Hypothese 4**: Self-consistency Prompting extrahiert Daten mit einer höheren Datenextraktionsgenauigkeit als ein Basisprompt.
- **Hypothese 5**: Reasoning-Modelle wie OpenAI „o1“ und „o3-mini“ weisen höhere Extraktionsgenauigkeit und geringere Halluzinationen auf gegenüber Non-Reasoning Modellen.
- **Hypothese 6**: Modelle mit großem Parameterumfang weisen bessere Datenextraktionsergebnisse als kleine Modelle auf.

Bei der Evaluation des PoCs werden die in Abschnitt 3.6 vorgestellten Metriken Precision, Recall, F1-Score, Accuracy und ROUGE verwendet, welche auf Basis einer Referenzlösung gebildet werden. Zur Optimierung des Modells sowie zur Validierung des Experiments wird ein reduzierter Datensatz genutzt, um Durchlaufzeiten zu minimieren.

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 [85]. Im Folgenden werden die Daten auf ihre Konsistenz und Relevanz für die Datenextraktion geprüft.

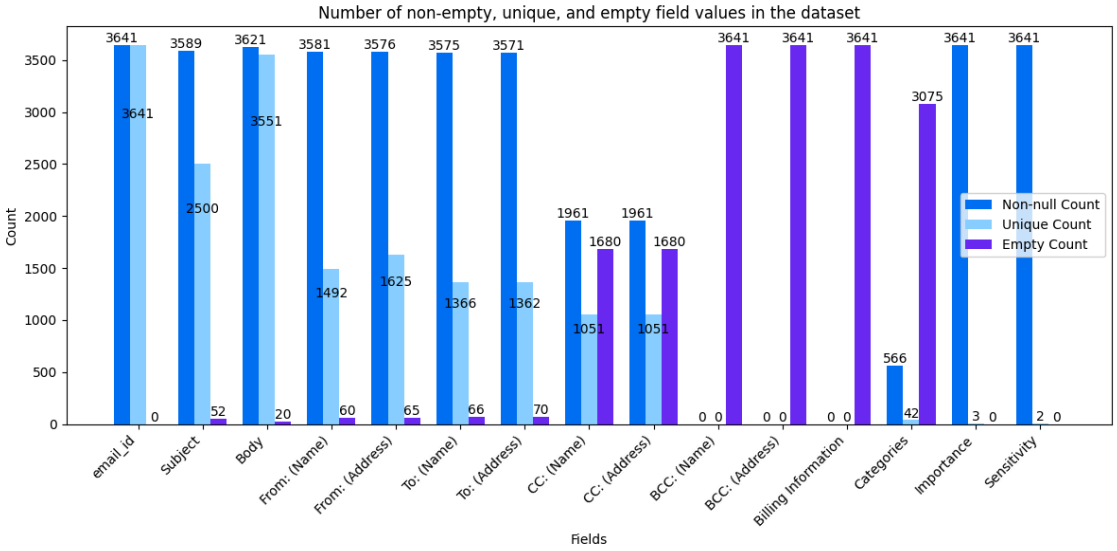

Abbildung 5: Attribute und Häufigkeit der Daten. Eigene Darstellung.

Der Datensatz enthält 3641 E-Mails aus Q4, 2024. Eine E-Mail wird von Microsoft in die in Abbildung 5 dargestellten Attribute aufgespalten. 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.

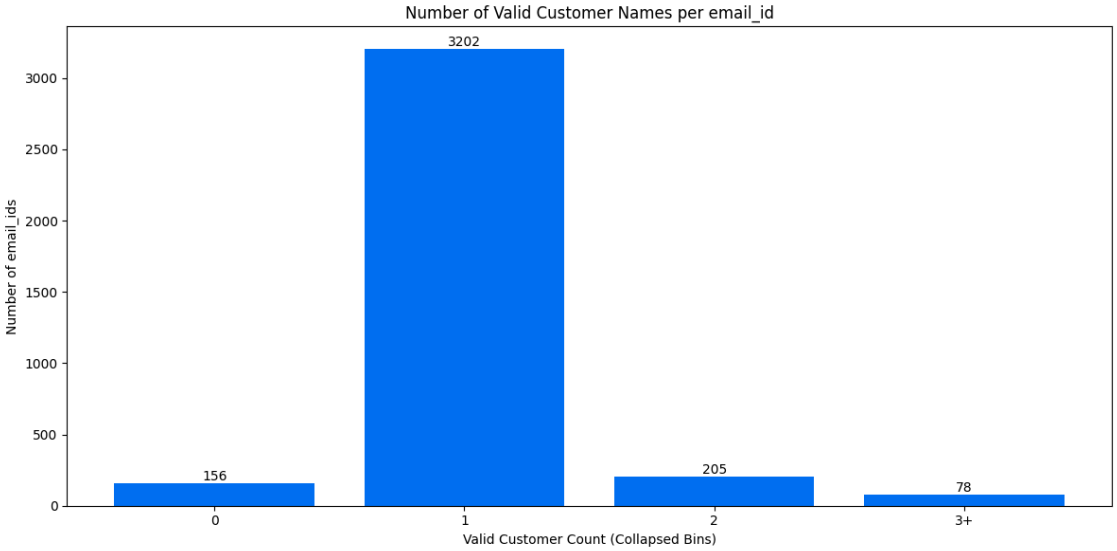

Abbildung 6: Anzahl der Kunden je E-Mail. Eigene Darstellung.

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, beispielweise automatisch generierte Antworten von „Microsoft Outlook“. Somit sind 78,94% der gesendeten Daten von „Non-RIG“ Absendern.

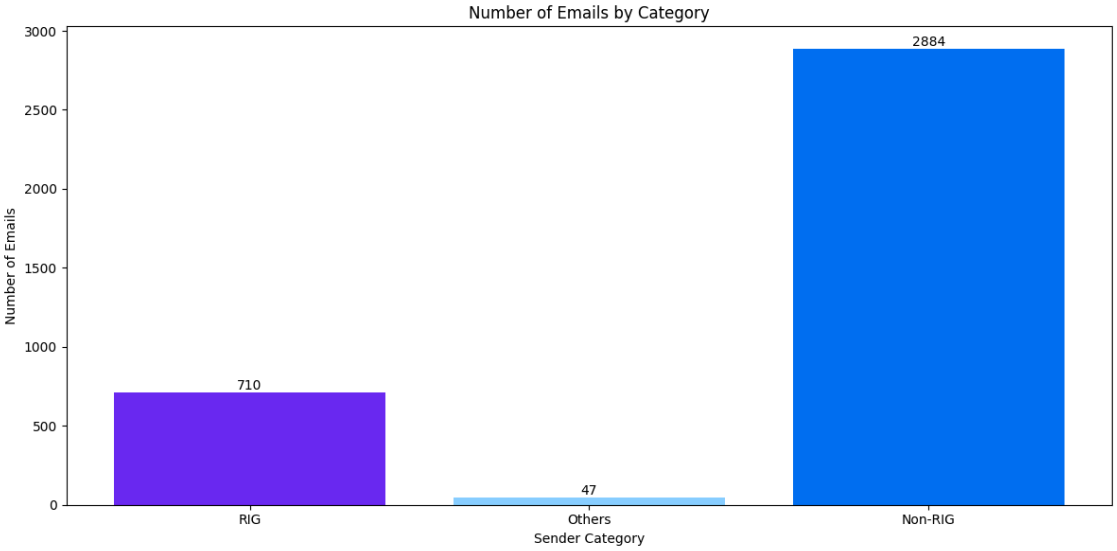

Abbildung 7: Anzahl an E-Mails je Sendergruppe. Eigene Darstellung.

Für die Identifikation der Opportunity sind die Felder ‚Subject', ‚Body', ‚From: (Name)' sowie ‚From: (Address)' 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.

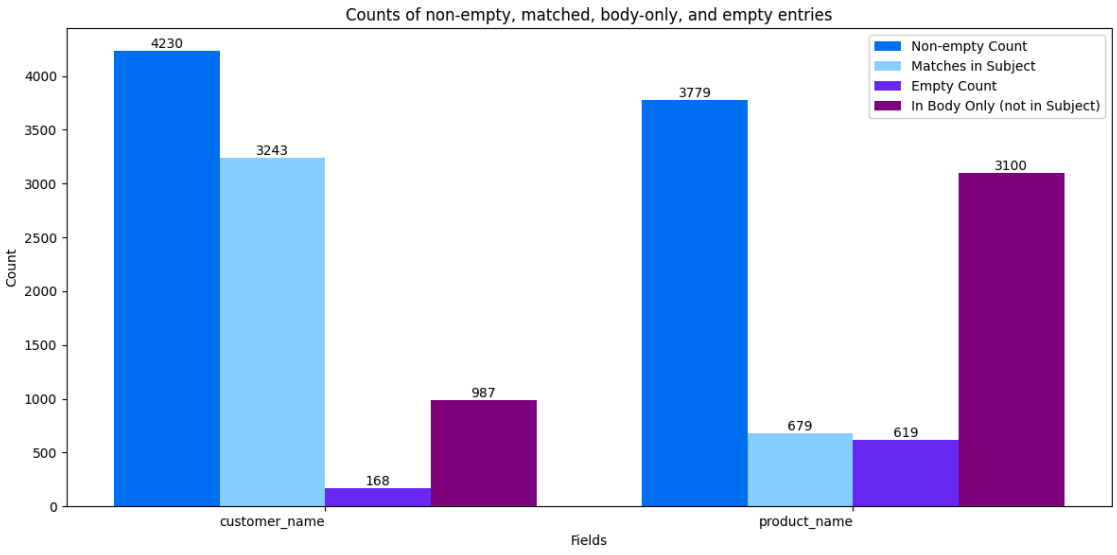

Abbildung 8: Analyse des E-Mail Betreffs. Eigene Darstellung.

Zur Analyse der Datenqualität wurde eine erste LLM-gestützte Datenextraktion durchgeführt, bei der OpenAIs GPT-4 mittels eines Zero-Shot-Prompts zur Extraktion relevanter Informationen aufgefordert wurde. Die extrahierten Daten

wurden anschließend als Parameter in einen API-GET-Request eingebunden und falls dieser genau einen Eintrag zuordnen kann, wird dieser gespeichert.

Diese konnte darunter 4398 mögliche Opportunities identifizieren. 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 Produktname (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.

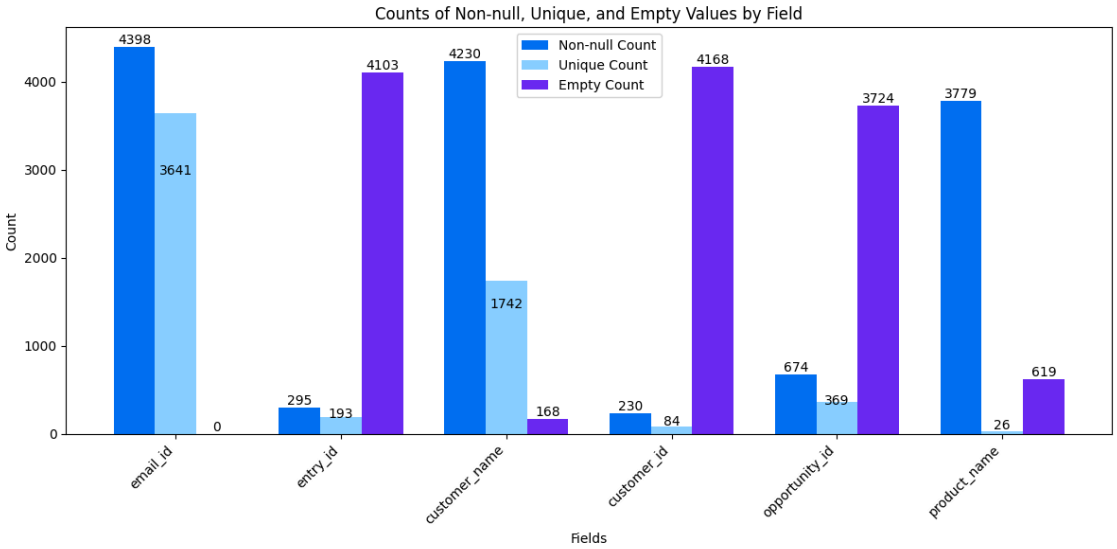

Abbildung 9: Ergebnisse einer initialen Extraktion. Eigene Darstellung.

Die in Abbildung 9 dargestellten Daten werden als Referenz zur weiteren Optimierung verwendet. Angestrebt ist eine höhere Extraktionsrate der Felder Opportunity- und Customer-ID um die Rate an richtig identifizierten Opportunities zu erhöhen. Auf Basis dieser Erkenntnisse werden die Daten aus dem Datensatz vorverarbeitet, um die Extraktion strukturierter Daten sowie Zuweisung zur Opportunity auf der MXP zu optimieren.

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 von der Microsoft GraphAPI extrahierte 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.

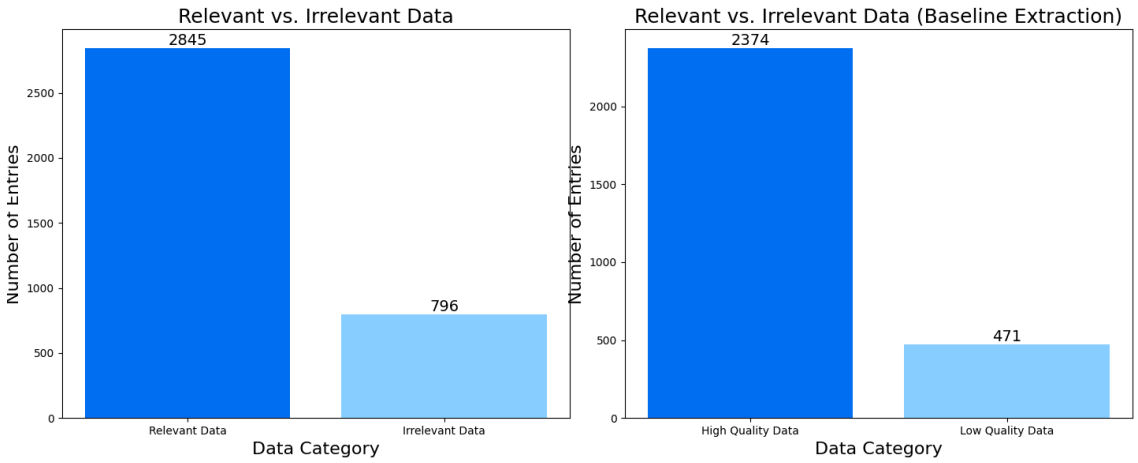

Abbildung 10: Anteil an relevanter Daten. Eigene Darstellung.

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]. Hinsichtlich der Datenqualität kann vor der Datenextraktion durch das Modell keine Aussage

getroffen werden, weswegen diese im Datensatz bestehen bleiben. 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.

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 reduzierten Datensatzes auf Basis von Abbildung 9

Der Datensatz umfasst 100 E-Mails unterschiedlicher Datenqualität, die in den in Tabelle 2 dargestellten Teilmengen enthalten sind. Die Datenqualität wurde manuell pro E-Mail auf Basis fehlender Attribute bewertet und bildet sämtliche potenziellen Extraktionsszenarien ab. Bei der Auswahl wurde bewusst eine Überrepräsentation seltener Klassen mit hoher Datenqualität im Validierungsdatensatz angestrebt, um Overfitting, eine zu starke Anpassung des Modells auf dem Trainingsdatensatz, zu minimieren. [86, S. 1-2]. Die Daten liegen in einer tabellarischen Form mit den Attributen ‚Subject', ‚Body' sowie einer E-Mail-ID vor.

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 die Produktivnahme zu hoch.

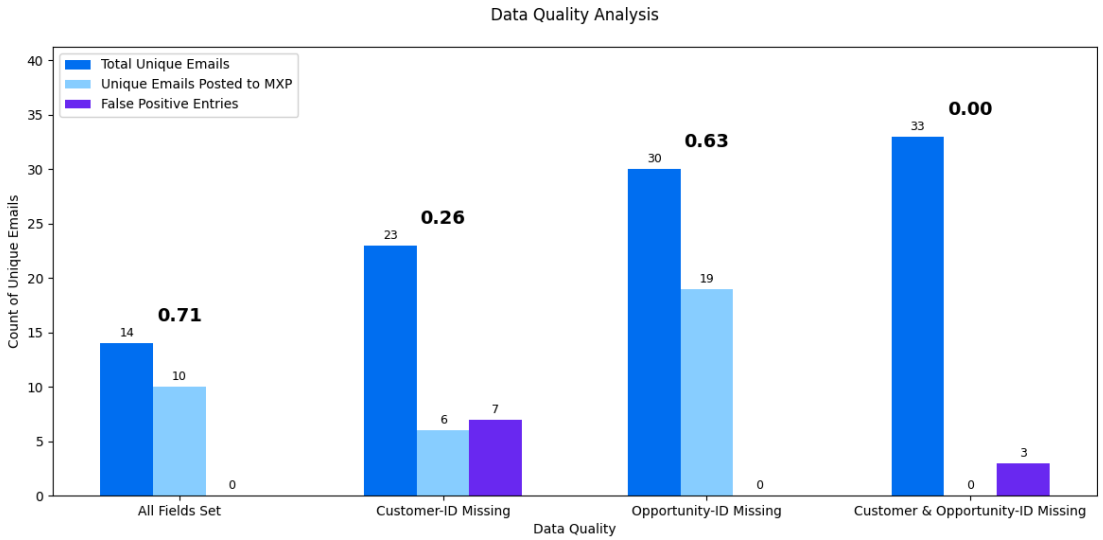

Abbildung 11: Anteil an identifizierten Opp. je Datenqualität. Eigene Darstellung.

Dieser Anteil soll durch Abschnitt 4.4 minimiert und der Anteil der korrekt identifizieren Einträge erhöht werden, indem Daten halluzinationsfrei 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 alternativer Verfahren, eingesetzt [38, S. 5-8], [58, S. 4-5]. Vom Training eines eigenen Modells wird aufgrund von geringer Datengrundlage sowie Kosten- und Rechenleistung abgesehen und zur Extraktion verschiedene PLMs verwendet [56, 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“ zur Verarbeitung der extrahierten Daten in Code eingesetzt, wobei dieser in Kombination mit dem Orchestration Service bislang nur die Modelle von OpenAI unterstützt [64].

Folgende Modelle werden zur Extraktion von E-Mail Daten über den GenAI Hub verwendet und auf ihre Performance untersucht:

- **OpenAI GPT-4o**: GPT-4o, veröffentlicht am 13. Mai 2024 [87], ist im Vergleich zu den anderen Modellen das älteste Modell, erreicht dennoch hohe Genauigkeit bei Datenextraktionsaufgaben [88, 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 [87], [89].
- **OpenAI o1**: Das Modell o1, eingeführt am 5. Dezember 2024, ist ein Modell, welches Reasoning durch internes Chain-of-Thought umsetzt und somit Extrak-

tionsergebnisse von GPT-4o übertrifft [90, 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 [89]. [91, 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 das von o1 übertrifft. [92]

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 im Form von frei-verfügbaren Codepaketen (Libraries) bereitstellt [93, S. 27-29]. Die im Rahmen dieser Arbeit verwendeten Libraries sind in Tabelle 3 dargestellt.

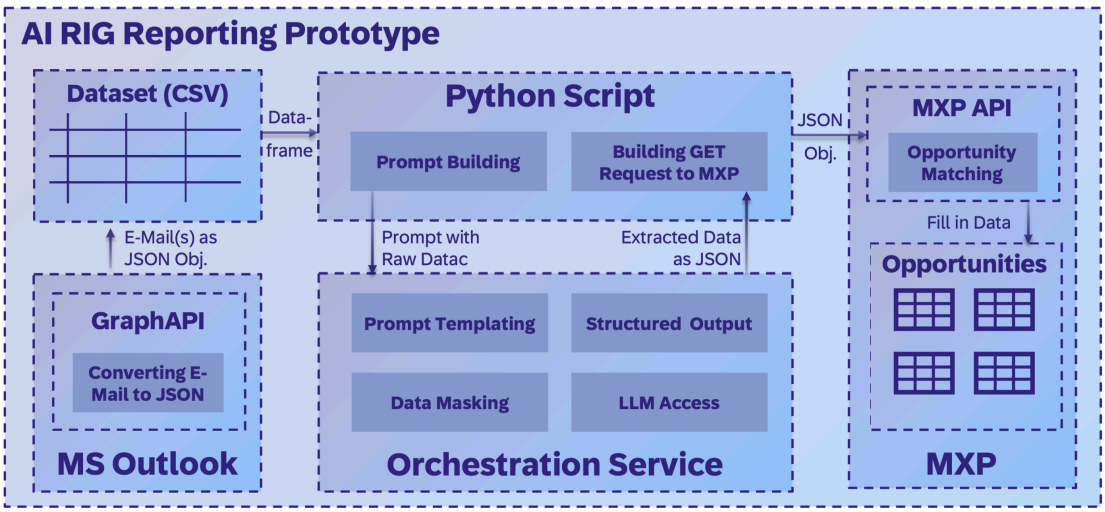

Abbildung 12: Extraktionspipeline des PoCs. Eigene Darstellung.

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 vom Python Skript verarbeitet. Die extrahierten Daten werden anschließend als Parameter eines GET-Request an die MXP gesendet für jede zum extrahierten Produktname 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 des Orchestration Service verschiedene Prompts gesendet mit Daten aus dem reduzierten Datensatz. 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-Shot, One-Shot 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 [94, S. 22]. Alle Prompts wurden hinsichtlich präziser Zieldefinition und klarer Struktur [76, S. 1-2]. mittels mehrerer LLMs optimiert [82, 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 stellt den Aufbau des Basisprompt dar. In Die als `{{?Parameter}}` gekennzeichneten Variablen werden durch „Prompt Templating“ des Orchestration Service als Werte dynamisch je E-Mail in den Prompt eingefügt. Im Prompt wird zwischen „Body“, dem letzten Inhalt der E-Mail, sowie dem „E-Mail-Context“, dem Inhalt vergangener E-Mails der E-Mail-Konversation, differenziert, um im Extraktionsschema den Fundort der Elemente zu spezifizieren [7, S. 3]. Der Aufbau des Chain-of-Thought Prompt [Prompt 2], Self-consistency Prompt [Prompt 3] sowie Zusatz eines Beispiels für One- und Few-Shot Prompting[Prompt 4] befinden sich im Anhang.

Im Modelling häufig aufkommende Risiken umfassen Data-Leakage, der Vermischung von Trainings- und Validierungsdatensatz [95, S. 4485-4486] , und falsch-validierte Testdaten. Diese werden mittels einer strikter Trennung von Testdaten (Generierung von Beispielen) und Validierungsdaten (Evaluation der LLM- und Prompt-Konfiguration) sichergestellt [95, S. 4486]. Der Testdatensatz mit fünf Beispielen wurde aus E-Mails und der MXP gewonnen, die die RIG bereits analysiert hat, um eine Deckunsggleichheit mit vergangener Extraktionen zu garantieren [96, S. 4].

4.5. Evaluation

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 Exrraktionsergebnisse evaluiert.

4.5.1. Datenextraktionsanalyse

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

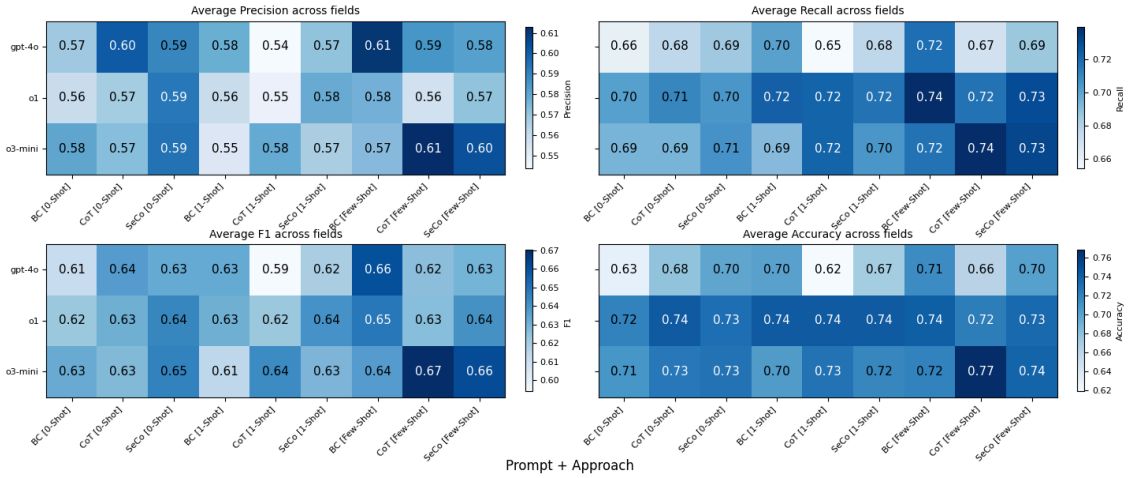

Abbildung 13: Durchschnittl. Metriken [alle Felder]. Eigene Darstellung.

Die Analyse von Abbildung 14 zeigt modellspezifische Unterschiede: GPT-4o weist bei der Extraktion von „opportunity_id“ und „analysis“ hohe Ergebnisse auf, erzielt aber in der Extraktion anderen Feldern geringe Ergebnisse auf. 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.

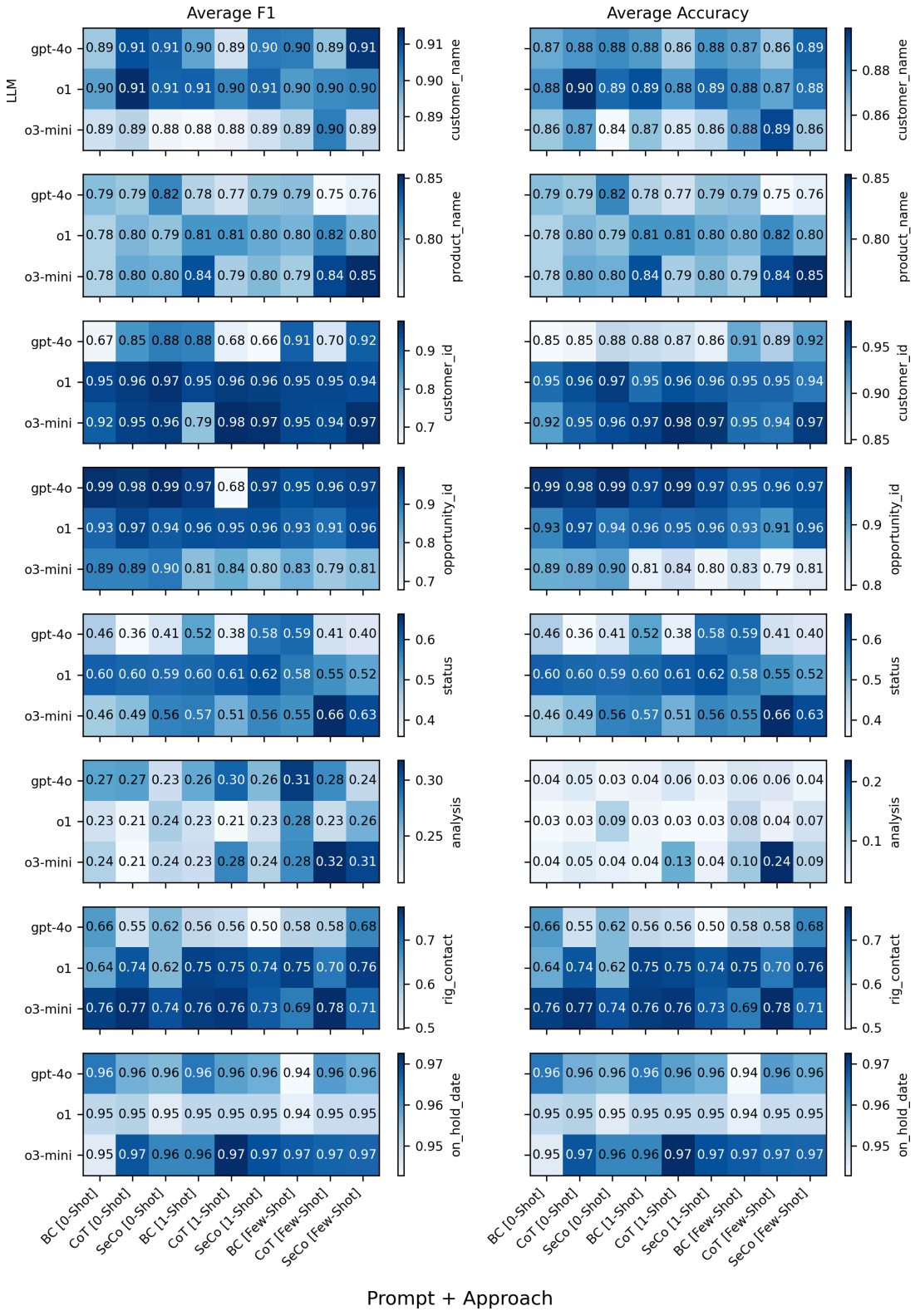

Abbildung 14: Durchschnittl. Metriken [je Feld]. Eigene Darstellung.

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

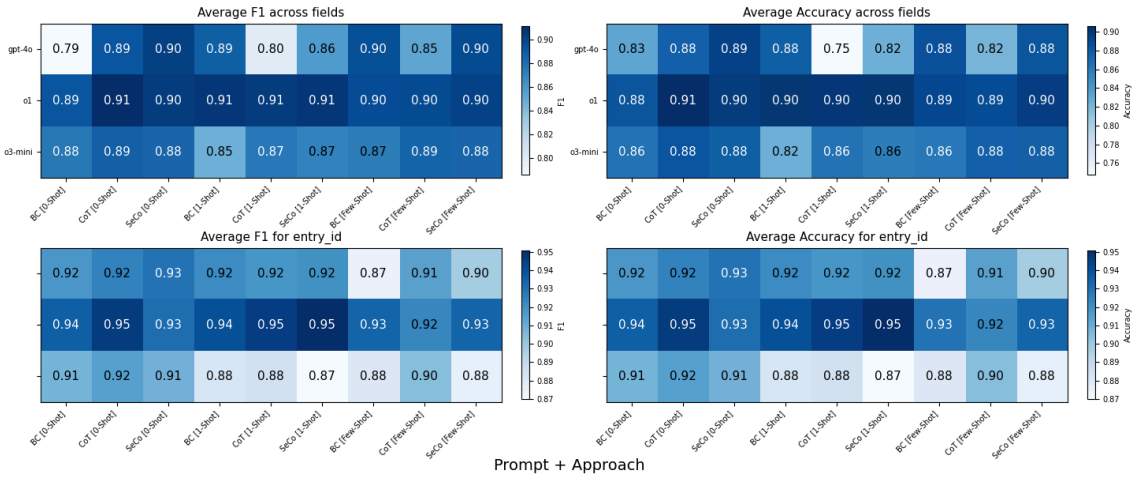

Abbildung 15: Durchschnittl. Metriken [Entry-ID]. Eigene Darstellung.

Die Ergebnisse von Abbildung 15 zeigen bei Betrachtung der Ergebnisse 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 die höchste Rate auf, wobei o3-mini schlechtere Ergebnisse als GPT-4o erreicht. Den höchsten F1-Score der Entry-ID erreicht o1 mit drei Prompts, wobei CoT (Zero-Shot) die höchste Accuracy bei den Extraktionsdatenfeldern aufwies.

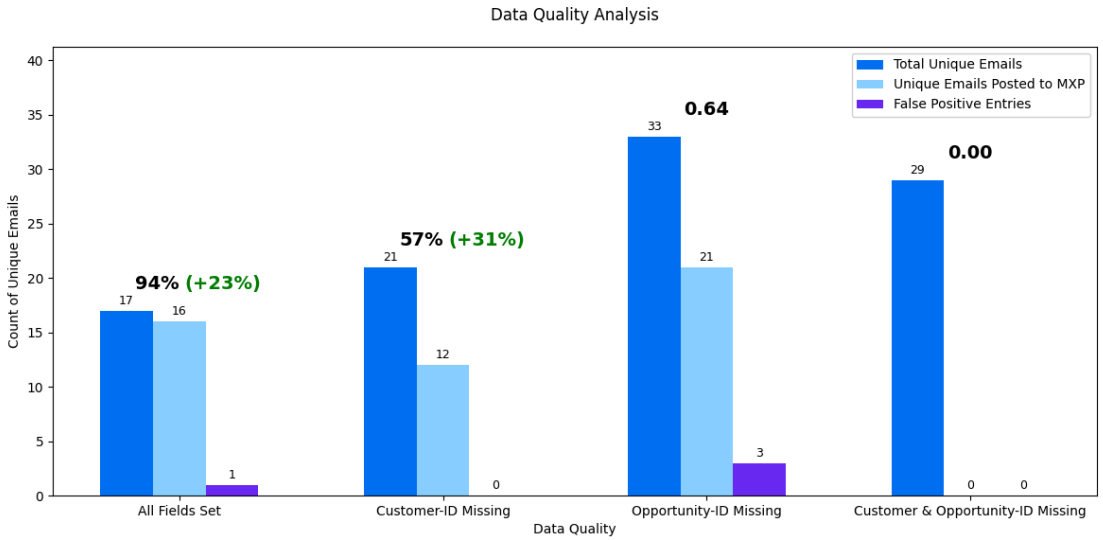

Abbildung 16: Identifikationsraten [alle Konfig.]. Eigene Darstellung.

Im Vergleich zur Abbildung 11 konnte die korrekte Identifikation der Opportunities, aggregiert über alle Konfigurationen, erheblich verbessert werden. Die geringere Rate halluzinierter MXP-Opportunities weist darauf hin, dass viele vormals als qualitativ hochwertig bewertete Einträge auf fehlerhaften Daten basierten – insbesondere bei der Customer-ID, deren Struktur keinem ersichtlichen Schema folgt. Diese Annahme wird durch die Beobachtungen aus Abbildung 14 gestützt.

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.

Zur finalen Evaluation werden die Ergebnisse der F1- und Accuracy mittels eines gewichteten arithmetischen Mittels bestimmt. Hierbei wird der Prozess der Datenextraktion aller acht Felder gleich stark gegenüber dem Prozess der MXP-Opportunity-Zuordnung bewertet. Bei vollständiger Attributbetrachtung gilt $k = 8$, bei ausschließlicher Analyse der Reportingdaten $k = 4$. Die entsprechenden Resultate sind in Abbildung 17 und Abbildung 18 dargestellt [97, S. 6–7].

$$\mathrm{F1}_{\mathrm{all}} = \frac{\sum_{i=1}^{k} \mathrm{F1}_i + k \cdot \mathrm{F1}_{\mathrm{MXP}}}{2 \cdot k}, \quad \mathrm{Acc}_{\mathrm{all}} = \frac{\sum_{i=1}^{k} \mathrm{Acc}_i + k \cdot \mathrm{Acc}_{\mathrm{MXP}}}{2 \cdot k} \tag{8}$$

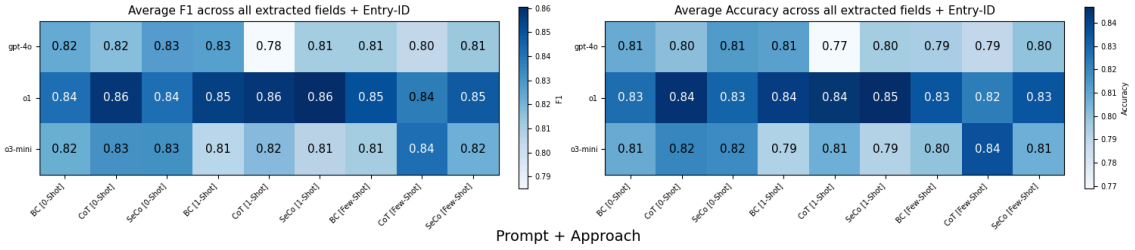

Abbildung 17: Durchschnittl. Metriken [Extraktion + Entry-ID]. Eigene Darstellung.

Da bei $k = 8$ auch Felder berücksichtigt werden, die ausschließlich zur Identifikation der MXP-Opportunity notwendig sind, führt die in Abbildung 15 gezeigte Korrelation dazu, dass dieser Aspekt in der Bewertung stärker gewichtet wird als die Qualität der Reportingdaten. Wie aus Abbildung 17 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: 0,86; Accuracy: 0,85).

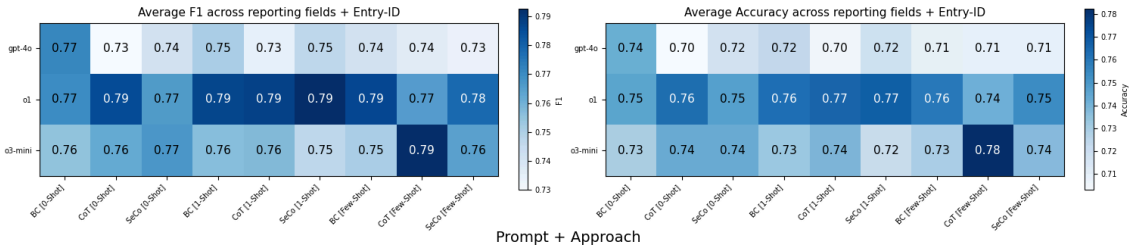

Abbildung 18: Durchschnittl. Metriken [Reporting + Entry-ID]. Eigene Darstellung.

Bei isolierter Betrachtung der Reportingqualität (k = 4) verringert sich die Varianz der Ergebnisse, wodurch modellspezifische Unterschiede weniger deutlich ausfallen. o3-mini erzielt mit einem CoT Few-Shot Prompt die höchsten Ergebnisse (F1: 0,79, Accuracy: 0,78), gefolgt von o1 Self-consistency One-Shot Prompt (F1: 0,79, Accuracy: 0,77).

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 geringeren Tokenkosten 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 Ausgangslage (Baseline) um ≥ 2 Prozentpunkte (pp), zeigen über alle Modelle hinweg eine positive Entwicklung und der Unterschied ist statistisch signifikant.
- **teilweise angenommen**: Mindestens eine Kennzahl verbessert sich um ≥ 1Prozentpunkt (pp), während alle anderen stabil bleiben oder sich ebenfalls leicht verbessern. Der Unterschied muss nicht statistisch signifikant sein, solange keine gegenläufigen Entwicklungen zwischen den Modellen auftreten.
- **abgelehnt**: Die Veränderung bleibt unter 1pp oder es treten sich widersprechende Ergebnisse auf. Ebenso wird die Hypothese abgelehnt, wenn der Unterschied statistisch nicht aussagekräftig ist.

Ein Effekt wird als statistisch signifikant klassifiziert, wenn die zweiseitige Irrtumswahrscheinlichkeit $\alpha < 0,05$ beträgt. Zur Bestimmung dieser Wahrscheinlichkeit wurde ein gepaarter, nicht-parametrischer Bootstrap-Ansatz nach B. Efron und R. J. Tibshirani [98] verwendet, bei dem F1- und Accuracy-Metriken wiederholt berechnet und deren Abweichungen vom arithmetischen Mittel analysiert wurden. Signifikanz liegt vor, wenn die Gesamtabweichung $< \alpha$ ist. [98, S. 5–8]

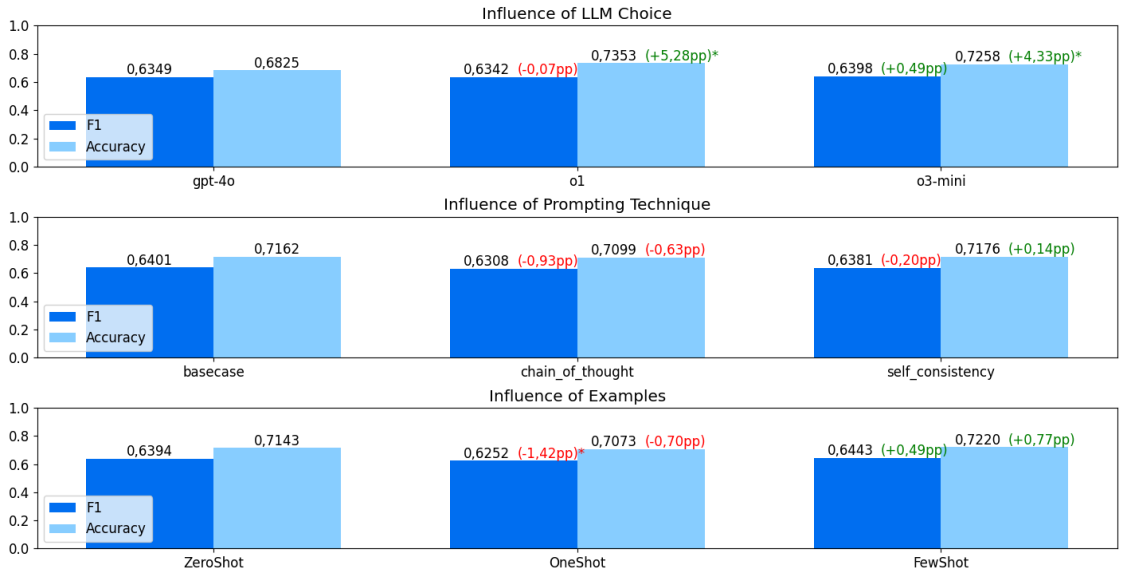

Abbildung 19: Auswirkung von Prompt/LLM. Eigene Darstellung.

Abbildung 19 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.

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 19 zeigen eine statistisch signifikante Reduktin des F1 um 1,42pp 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 19 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 19 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 Hypothese 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 Optimierung gegenüber einer ersten Datenextraktion. Beide in Abschnitt 4.5.3 ausgewählten Modellkonfigurationen erreichen die in Abschnitt 4.1 für eine Produktivnahme definierten Vorgaben der RIG von > 70% Genauigkeit. Die Abweichung der manuellen MXP-Identifikation liegt wie aus Abbildung 15 ersichtlich bei 95% Genauigkeit und somit unter 10% Abweichung, wobei der Anteil der falsch-positiv identifizieren Opportunities den Schwellenwert von 5% überteigt [Abbildung 21, Abbildung 22]. Damit gelten die Vorgaben der RIG als teilweise erfüllt und müssen vor einer finalen Produktivnahme weiter verbessert werden. Dennoch steht eine Produktivnahme durch Teilerfüllung der Vorgaben nach weiterer Optimierung in Aussicht.

Nächste Schritte umfassen eine Prüfung der Konfigurationen auf einem zeitlich aktuelleren Datensatz, um eine Abweichung von der Datengrundlage zu vermeiden und die Ergebnisse auf Overfitting zu prüfen [86, S. 2]. Darüber hinaus kann die RIG folgende Schritte durchführen, um die Extraktionsgenauigkeit anhand der E-Mails zu optimieren:

- **Anpassung der E-Mail-Kampagnen**: Die RIG nimmt i.d.R. den Erstkontakt mit dem Kundenteam in Form der E-Mail-Kampagnen auf. Mit der Hinzugabe von MXP-Identifikatoren wie die CRM-Account- und Opportunity-ID steigt die Datenqualität aller nachfolgender E-Mails des Verlaufs und die MXP-Identifikationsrate steigt [Abbildung 16].
- **Kundenanzahl je E-Mail-Verlauf**: Die Analyse ergibt, dass ein LLM bei der Verarbeitung mehrerer Kunden je E-Mail-Verlauf insbesondere bei der Zuordnung der CRM-Account-ID Schwierigkeiten hat, weswegen AEs aufgefordert werden sollen, je E-Mail Verlauf nur einen Kunden zu beschreiben.

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.

Die von der RIG definierten Vorgaben einer Extraktionsgenauigkeit von > 70% sowie einer mit manueller Extraktion abweichende Identifikation der MXP-Opportunity von < 10% konnte erreicht werden, eine falsch-positive Identifikation der zugehörigen Opportunity konnte aufgrund von halluzinierten Daten des Modells nicht erreicht werden. Trotz nur teilweise erreichter Vorgaben erweisen sich die Optimierungsergebnisse als wertvoll und stellen positive Aussichten auf eine Produktivnahme nach Durchführung weiterer Optimierungsmaßnahmen.

Aufgrund der Identifikation einer zu hohen Zahl falsch positiv identifizierter Einträge wird sich gegen eine sofortige Produktivnahme ausgesprochen und für eine weitere Optimierung der Datenextraktion. Nächste Schritte umfassen die Prüfung und Verbesserung der Datengrundlage und -qualität, eine Verbesserung des MXP-Algorithmus sowie die Durchführung weiterer Experimente hinsichtlich stärkerer LLMs.

5.2. Einordnung der Ergebnisse

Die erzielten Ergebnisse aus Abschnitt 4.5 reihen sich in aktuelle Literatur zu LLM-basierter Datenextraktion ein. Die erreichte Genauigkeit von 74% ist vergleichbar mit der Genauigkeit in anderen Bereichen [60, S. 5167 - 5174] und zeigt ein ähnliches Verhalten beim Einsatz von Reasoning-Modellen [90, S. 9-12]. Diese Arbeit erzielt, entgegen Datenextraktionen aus anderen Bereichen [10, S. 8-9], [75, 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 Anpassung der Prompts identifiziert werden.x

5.3. Herausforderungen und Limitationen

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

Der zur Modellierung verwendete Datensatz, bestehend aus 100 E-Mails, weist Herausforderungen auf: im Vergleich zu anderen Datenextraktionsanalysen [10, S. 3-4], [90, S. 8-9] ist der Validierungsdatensatz klein und besitzt eine geringe Datenvarianz, wodurch er anfällig für Overfitting ist [86, S. 1-2]. Darüber hinaus können bei manuelle Extraktion 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 regelbasierter Verfahren: durch das persistierende Risiko an Halluzinationen ist eine vollständige Minimierung von falsch-positiver MXP-Opportunity-

Zuordnung erschwert möglich [51, S. 11-12]. Des Weiteren wurde im Rahmen dieser Arbeit je Prompting-Technik ein Prompt [Prompt 2, Prompt 3] generiert, welcher Fehler bzw. modellspezifische Bias aufweisen kann [82, S. 8-10]. Mittels „LLM-Stacking“ kann diese Beobachtung im Rahmen einer anderen Arbeit aufgegriffen werden [99].

Die Interpretation der Ergebnisse ist limitiert durch die in dieser Arbeit verwendeten Tools: der Orchestration Service des SAP GenAI Hub limitiert die einsetzbaren Modelle auf einen Anbieter [64], [72, S. 1], wodurch ein hohes Risiko von anbieterspezifischen Bias besteht und die Ergebnisse verzerren kann [100, S. 3-7]. Zur Validierung der Ergebnisse ist eine Durchführung notwendig, sobald mehr Modelle des SAP GenAI Hub „Structured Outputs“ unterstützen.

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

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. Durch die Nichterfüllung aller Vorgaben der RIG sollte sich die zukünftige Forschung auf eine weitere Optimierung des Prozesses fokussieren. Mögliche Ansätze sind hierbei:

- **Modellauswahl**: Durch die Veröffentlichung weiterer Modelle diverser Anbieter, die die Leistung der alten Modelle übertreffen, bieten diese Potential für eine weitere Optimierung in der Zukunft [101, S. 17-19].
- **Hyperparameter-Tuning**: Eine Optimierung anhand von Hyperparamtern erzielt in anderen Datenextraktionsaufgaben eine verbesserte Extraktionsgenauigkeit [36]. Der SAP GenAI Hub soll in Zukunft das Setzen von Temperatur und Tokenlimit der Modelle o1 und o3-mini unterstützen
- **Optimierung der 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 identifizieren kann, 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.

Bei der Umsetzung weiterer Optimierung 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/XtjgLMPVFW8LAIGnoROj6R/slide_viewer

[6] Z. C. Lipton, C. Elkan, und B. Narayanaswamy, „Thresholding Classifiers to Maximize F1 Score“, *arXiv preprint*, 2014, Zugegriffen: 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“. Zugegriffen: 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. Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-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“. Zugegriffen: 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. Zugegriffen: 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, Zugegriffen: 4. März 2025. [Online]. Verfügbar unter: https://arxiv.org/abs/2208.00752

[19] 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.

[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, Zugegriffen: 7. März 2025. [Online]. Verfügbar unter: https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1180962/full

[22] X. Pu, M. Gao, und X. Wan, „Summarization is (Almost) Dead“, *arXiv*, 2023, Zugegriffen: 24. Februar 2025. [Online]. Verfügbar unter: https://arxiv.org/abs/2309.09558

[23] 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.

[24] 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

[25] 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

[26] 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.

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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/ModelerCRISPDM.pdf

[32] D. Dean, „Advanced Data Analytics for Organizations“. Zugegriffen: 4. März 2025. [Online]. Verfügbar unter: https://app.myeducator.com/reader/web/1421a

[33] 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.

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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.

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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.

[49] OpenAI, „GPT-4 Technical Report“, *arXiv preprint arXiv:2303.08774*, 2023, Zugegriffen: 27. Februar 2025. [Online]. Verfügbar unter: https://arxiv.org/pdf/2303.08774

[50] 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

[51] 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

[52] 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

[53] OpenAI, „OpenAI API Pricing“. [Online]. Verfügbar unter: https://openai.com/api/pricing/

[54] 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

[55] 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/1w12fy

78ghug5/Rule-based-Information-Extraction-Advantages-Limitations-and-Perspectives

[56] H. Wang *u. a.*, „A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy“, *arXiv preprint arXiv:2501.09431*, 2024.

[57] „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

[58] 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

[59] 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.

[60] 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, Zugegriffen: 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

[61] 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.

[62] 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

[63] 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

[64] 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

[65] OpenAI, „Structured Outputs“. [Online]. Verfügbar unter: https://platform.openai.com/docs/guides/structured-outputs?api-mode=responses

[66] 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

[67] 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

[68] SAP, „Multi Experience Platform (MXP) Overview“. [Online]. Verfügbar unter: https://sap.sharepoint.com/sites/206579/SitePages/Multi%20Experience%20Platform%20(MXP)%20Overview.aspx

[69] 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

[70] 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

[71] 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

[72] 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

[73] 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

[74] 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

[75] 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

[76] 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

[77] OpenAI, „Prompt Engineering“. Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: https://platform.openai.com/docs/guides/prompt-engineering

[78] 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

[79] Microsoft, „Prompt Engineering techniques“. Zugegriffen: 27. März 2025. [Online]. Verfügbar unter: https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/prompt-engineering

[80] 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

[81] 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

[82] 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

[83] 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

[84] 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

[85] 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

[86] 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

[87] OpenAI, „GPT-4o“. [Online]. Verfügbar unter: https://platform.openai.com/docs/models/gpt-4o

[88] 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

[89] 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

[90] A. Brokman, X. Ai, Y. Jiang, S. Gupta, und R. Kavuluru, „A Benchmark for End-to-End Zero-Shot Biomedical Relation Extraction with LLMs: Experiments with OpenAI Models“. Zugegriffen: 17. April 2025. [Online]. Verfügbar unter: https://arxiv.org/abs/2504.04083

[91] OpenAI *u. a.*, „OpenAI o1 System Card“. Zugegriffen: 17. April 2025. [Online]. Verfügbar unter: https://arxiv.org/abs/2412.16720

[92] OpenAI, „OpenAI o3-mini“. [Online]. Verfügbar unter: https://openai.com/index/openai-o3-mini/

[93] 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.

[94] 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

[95] 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

[96] A. Arora, A. Bowe, und P. Rajpurkar, „Humans Continue to Outperform Large Language Models in Complex Clinical Decision-Making: A Study with Medical Calculators“, *arXiv preprint arXiv:2411.05897*, Nov. 2024, doi: 10.48550/arXiv.2411.05897.

[97] J. Opitz, „A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice“. Zugegriffen: 20. April 2025. [Online]. Verfügbar unter: https://arxiv.org/abs/2404.16958

[98] 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

[99] 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

[100] 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

[101] 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

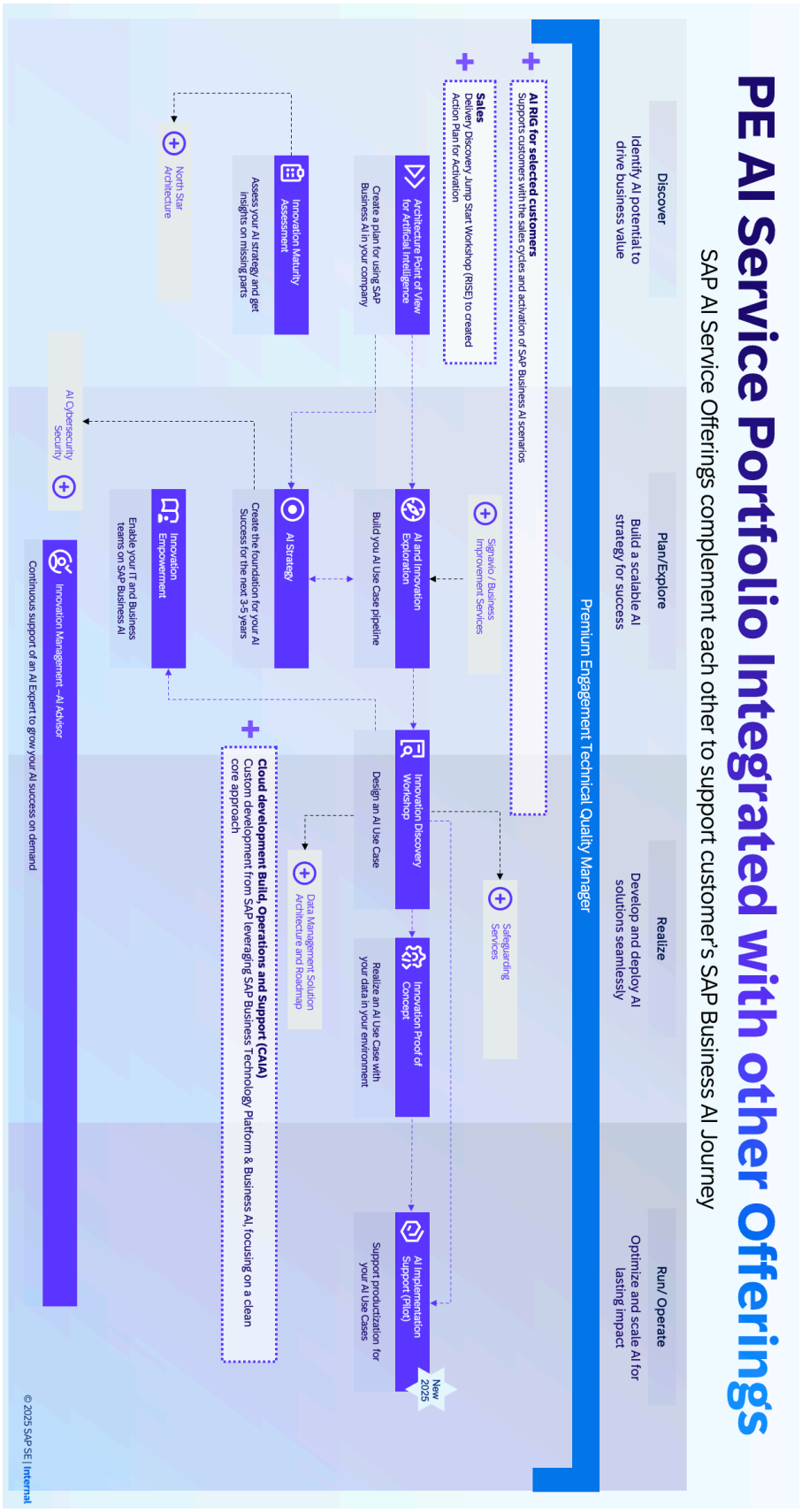

Abbildung 20: SAP AI Services - Aufgaben der RIG [5, S. 8]

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

Produktname	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 verwendete 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 verwendete 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

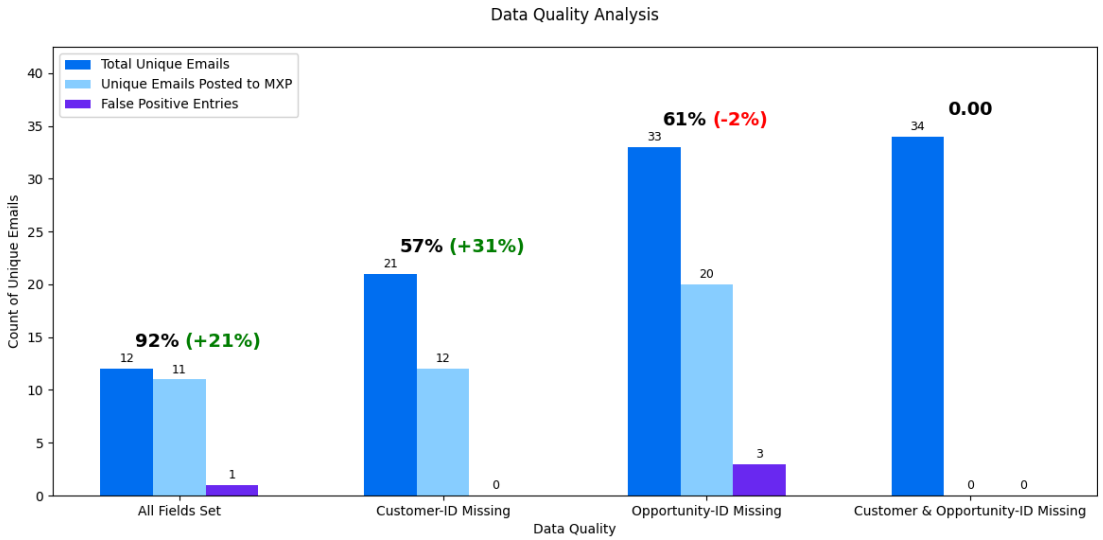

Abbildung 21: Identifikationsraten [beste o1 Konfig.]. Eigene Darstellung.

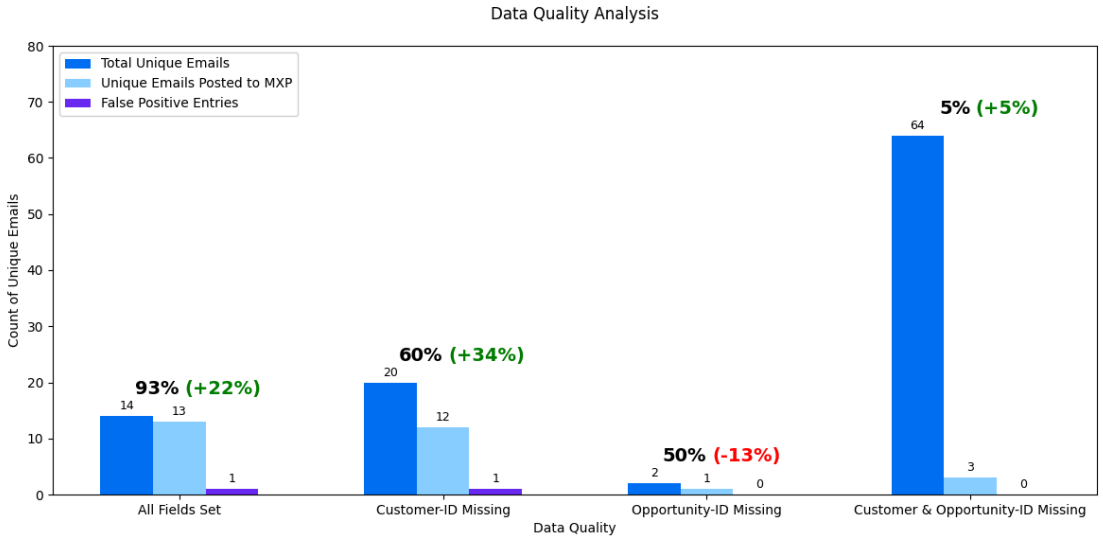

Abbildung 22: Identifikationsraten [beste o3-mini Konfig.]. Eigene Darstellung.