Philipps Universität Marburg

Namen des Studenten: Gerd Siebert

Fachmodul: Marburg Modul

Prüfbericht

(Anteil Fehlersuche)

Projekt: AI-Lab: Wie können wir mit KI zusammen gut leben?

Semester: Winter 2023/24

Thema: Fundamentale KI-Modelle in der Datenverarbeitung

Prüfbericht zur Fehlersuche in Tabellen unter Verwendung mehrerer AI-Tools

Projektdauer: 1. Oktober 2023 bis 31. März 2024

Philipps-Universität Marburg
Fachbereich 12 – Mathematik und Informatik
Prof. Dr. Thorsten Papenbrock

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

Abkürzung	Bedeutung
AI	Artificial Intelligence
AI-Lab	Artificial Intelligence-Laboratory
CSV	Comma Separated Values
GDP	Gross Domestic Product
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
KI	Künstliche Intelligenz
LLaMa-2	Large Language Model Meta AI - 2
LLM	Large Language Model
MS	Microsoft

Tabelle 1: Abkürzungsverzeichnis

Glossar

AI-Lab: Das AI-Lab ist ein projektbasiertes Modul der Universität Marburg, in dem die Teilnehmenden in Teams Projekte entwickeln und dabei sowohl Programmierkenntnisse als auch Methoden der KI erlernen. *(https://ilias.uni-marburg.de/ilias.php?ref_id=3276318&cmd=view&cmdClass=ilrepositorygui&cmdNode=yz&baseClass=ilRepositoryGUI)*

AI-Tools: AI-Tools sind Werkzeuge, die auf künstlicher Intelligenz (KI/AI) und maschinellem Lernen basieren und es ermöglichen menschenähnliche Dialoge zu führen, um den Menschen über unterschiedlichste Dialoge bei der Erarbeitung und Lösung verschiedenster Aufgaben zu unterstützen, z.B. ChatGPT oder MS Copilot. *(https://www.campixx.de/tools/ki-ai-*

tools/#:~:text=AI%2DTools%20zur%20Transkription%20sind,Videoaufnahmen%20automatisch%20in%20Text%20umzuwandeln)

Artificial Intelligence (AI): Bezeichnet Systeme, mit denen versucht wird bestimmte Entscheidungsstrukturen des Menschen nachzubilden – also Wissen zu erwerben, zu verstehen und anzuwenden, um komplexe Probleme zu lösen oder sich an neue Situationen anzupassen. *(https://ilias.uni-giessen.de/goto.php?target=wiki_347086_glossar&lang=de)*

Generative Pre-trained Transformer (GPT): Bezeichnet ein Sprachmodell, welches auf einem via Deep Learning trainierten künstlichen neuronalen Netz basiert. Ein solches Sprachmodell kann auf unterschiedliche Trainingsziele ausgerichtet werden (z.B. menschenähnliche Dialoge führen). *(https://ilias.uni-giessen.de/goto.php?target=wiki_347086_glossar&lang=de)*

Large Language Models (LLM): Large Language Models sind große generative Sprachmodelle mit Künstlicher Intelligenz, die mit riesigen Mengen an Textdaten vortrainiert sind. Sie basieren auf neuronalen Netzen, in der Regel in Transformer-Architektur, und besitzen viele Milliarden Parameter. LLMs können natürliche Sprache verarbeiten, verstehen und generieren. Große Sprachmodelle eignen sich für zahlreiche Anwendungen und sind beispielsweise die Grundlage für KI-Chatbots wie ChatGPT oder Google Bard. *(https://www.bigdata-insider.de/was-ist-ein-large-language-model-a-d735d93bbc24d3c4091de8ce25aa36e8/)*

Änderungsverzeichnis

Datum	Version	Autor	Änderung	Kapitel/ Seite
10.02.2024	0.1	Gerd	Entwurf	alle
24.03.2024	1.0	Gerd	Überarbeitung	alle

Tabelle 2: Änderungsübersicht

1 Einleitung

In diesem Prüfbericht werden die Ergebnisse und Erkenntnisse aus der Durchführung eines Projekts zur Fehlersuche in Datensätzen/Tabellen mit Hilfe von AI-Tools detailliert dargestellt. Die im Rahmen des Projekts durchgeführten Prüfaktivitäten werden beschrieben, wobei ein besonderer Fokus auf den Einsatz der AI-Tools BloomZ und LLaMa-2 gelegt wird.

Die gewonnenen Erkenntnisse werden zusammengefasst und Schlussfolgerungen hinsichtlich der Zielerreichung gezogen. Darüber hinaus werden Vorschläge für mögliche weitere Vorgehensweisen gemacht.

Im Kapitel “Ergebnisse” werden die Testergebnisse für die AI-Tools dokumentiert. Dies umfasst eine allgemeine Bewertung der AI-Tools, eine Stärken/Schwächen-Analyse, Aussagen zu möglichen Verbesserungen der AI-Tools BloomZ und LLaMa-2, sofern erkennbar, sowie Aussagen zu möglichen Einsatzmöglichkeiten dieser Tools.

Dieser Bericht dient als umfassende Ressource für alle, die an der Fehlersuche von Datensätzen/Tabellen mit AI-Tools interessiert sind, und bietet wertvolle Einblicke in die Leistungsfähigkeit und das Potenzial dieser Tools.

Zudem gibt er eine Zusammenfassung der gewonnenen Erkenntnisse, zieht Schlussfolgerungen über die Erfüllung der gesetzten Ziel und macht Vorschläge über ein mögliches, weiteres Vorgehen.

2 Projektbeschreibung

In diesem Projekt werden verschiedene AI-Tools auf ihre Fähigkeit hin untersucht, Datensätze und Tabellen zu analysieren und Fehler zu identifizieren. Die Prüfverfahren für diese AI-Tools umfassen mehrere Aspekte.

Das Hauptziel dieses Projekts war es, die Fähigkeiten der AI-Tools BloomZ und LLaMa-2 zur Analyse von Tabellen mittels AI-Prompts zu bewerten und mit anderen AI-Tools zu vergleichen. Die Prüfung konzentrierte sich auf zwei Hauptbereiche:

- Fehlersuche in Datensätzen/Tabellen
- Vergleich von Datensätzen/Tabellen (separates Dokument)

Das Projekt zielte darauf ab, die Eignung von BloomZ und LLaMa-2 als Werkzeug zur Analyse und Bereinigung von Datensatz-/Tabelleninhalten zu bewerten. Andere allgemeine AI-Tools, wie z.B. ChatGPT oder Microsoft Copilot dienen als Benchmark.

Die verwendeten AI-Tools sollten in der Lage sein, Tabellen in verschiedenen Formaten (z.B. CSV) zu verarbeiten. Sie sollten eine benutzerfreundliche Oberfläche für die Eingabe von Prompts bieten und leistungsfähig genug sein, um kleinere Tabellen zu verarbeiten.

Das AI-Tool BloomZ und später LLaMa-2 wurden auf einem eigenen Server der Universität Marburg gehostet. Diese AI-Tools konnten auch über die Website von Huggingface genutzt werden. Die anderen AI-Tools wurden ebenfalls über den Webbrowser genutzt.

Die Tabellen lagen im CSV-Format vor und wurden von den Prüfern so manipuliert, dass sie kontrolliert für die Fehleranalyse inkonsistent wurden, d.h. sie enthielten dann Fehler in Bezug auf die Datenintegrität und die referenzielle Integrität.

3 Durchführung

Die AI-Tools wurden auf ihre Fähigkeit hin geprüft, Fehler in Datensätzen und Tabellen zu finden. Sie suchen nach leeren Zellen, Duplikaten, ungültigen und inkonsistenten Werten und erzeugen eine Meldung und eine Beschreibung der gefundenen Fehler.

Es gab dabei verschiedene wesentliche Funktionen, die zur Erfüllung der Prüfaufgaben vorgesehen sind. Dazu gehören die Überprüfung des Datentyps, die Erkennung fehlender Werte, die Korrektur der Daten, die Überprüfung des Datenformats, die Erkennung von Duplikaten und der Datenvergleich.

Dabei war zu beachten, dass die kostenlosen Versionen der genutzten AI-Tools oft eingeschränkte Funktionen hatten. Beispielsweise hat die kostenlose Version von MS Copilot keine Möglichkeit, Datensätze oder Tabellen als Datei hochzuladen. Es gab auch Einschränkungen bei der Anzahl der Prompts oder der Anzahl der eingebbaren Character.

Für die Durchführung kamen zum Einsatz:

- Je Prüfaufgabe ein Datensatz/Tabelle im CSV-Format
- Die Promptvarianten Zero-Shot, Few-Shot und Template
- Vier verschiedene generative Large Language Models (LLMs)/AI-Tools
  - BloomZ (3B / 7B1-mt)
  - LLaMa-2 7B Chat
  - Microsoft 365 Copilot
  - ChatGPT (GPT 3.5)

3.1 Genutzte Datensätze/Tabellen zur Fehlersuche

Der Datensatz „Adult“ wurde, wie in nachfolgender Tabelle zu sehen, angepasst. Dabei wurden einige Tabellenspalten eliminiert und andere hinzugefügt. Zum Teil wurden auch Werte modifiziert. Aufgrund der begrenzten Eingabemöglichkeiten bei den verschiedenen AI-Tools (z.B. Copilot mit 4000 Zeichen) wurde der Datensatz auf 18 Instanzen reduziert. Ein Verkürzung oder Erweiterung ist jedoch jederzeit für die Untersuchung möglich.

Row	First Name	Family Name	Age	Birthday	Employment	Degree	Relationship	Occupation	Salary (USD)	Children	Colour	Gender	Religion	Country
101	Mei Ling	Chen	36	October 10, 1987	Private	Doctorate	Never-married	Prof-specialty	$144,470	0	Black	Male	Atheism	England
102	Carlos	Rodriguez	32	November 1, 1991	Private	Some-college	Divorced	Other-service	$62,280	2	Black	Male	Christianity	Spain
103	Aisha	Ahmed	28	August 30, 1995	State-gov	HS-grad	Married	?	$77,420	2	White	Female	Islam	USA
104	Raj	Kapoor	26	June 6, 1997	Private	Bachelors	Never-married	Prof-specialty	$113,540	2	Asian	Male	Islam	Pakistan
105	Amira	Khan	47	September 15, 1976	Private	5th-6th	Never-married	Priv-house-serv	$28,570	0	White	Female	Buddhism	England
106	Mia	Johansson	28	October 28, 1995	State-gov	11th	Separated	Adm-clerical	$67,250	4	Asian	Female	Christianity	Sweden
107	Ahmed	Abdel Nasser	47	September 9, 1976	Self-emp-not-inc	5th-6th	Married-civ-spouse	Sales	$97,440	3	White	Male	Islam	Egypt
108	Mei	Wang	34	April 16, 1989	Self-emp-not-inc	Bachelors	Married-civ-spouse	Sales	$98,680	2	Asian	Male	Atheism	China
109	Takeshi	Sato	51	May 31, 1972	Private	7th-8th	Married	Transport-moving	$48,790	1	White	Male	Buddhism	Japan
110	Fatur	Al-Farsi	35	April 8, 1990	Private	Assoc-voc	Married-civ-spouse	Sales	$85,100	0	White	Male	Islam	Egypt
111	Leo	Garmann	44	July 17, 1979	Federal-gov	Bachelors	Married-civ-spouse	Adm-clerical	$90,230	2	Black	Male	Christianity	Germany
112	Aarav	Mehta	38	November 4, 1985	Local-gov	Bachelors	Never-married	Exec-managerial	$112,500	0	White	Male	Islam	Saudi Arabia
113	Liam	Johnson	44	July 6, 1979	Private	HS-grad	Married-civ-spouse	Craft-repair	$71,640	0	White	Male	Atheism	United States
114	Kaito	Watanabe	29	January 11, 1994	Private	Doctorate	Never-married	Prof-specialty	$123,450	0	Asian	Male	Buddhism	Japan
115	Siti	Lim	34	August 2, 1989	Private	Some-college	Married	Protective-serv	$69,180	3	Black	Female	Buddhism	USA
116	Elmar B.	Ivanov	23	January 8, 1998	Federal-gov	Some-college	Never-married	?	$70,750	1	White	Male	Christianity	Bulgaria
117	Liam	O'Brien	25	March 18, 2000	?	Bachelors	Married-spouse-absent	?	$93,610	0	White	Male	Judaism	Canada
118	Priya	Desai	53	January 22, 1969	Federal-gov	Some-college	Married-civ-spouse	Exec-managerial	$65,220	0	Asian	Female	Hinduism	France

Tabelle 3: Geänderter Datensatz/Tabelle zur Analyse von Fehlern in Datensätzen/Tabellen

Nachfolgend eine Liste der Spalten und ihre Bedeutung:

- **Row:** Diese Spalte gibt die Nummer der Zeile an, in der die Daten enthalten sind.
- **First Name:** Diese Spalte enthält den Vornamen der Person.
- **Family Name**: Diese Spalte enthält den Familiennamen der Person.
- **Age:** Diese Spalte enthält das Alter der Person.
- **Birthday:** Diese Spalte enthält das Geburtsdatum der Person.
- **Employment:** Diese Spalte beschreibt den Beschäftigungsstatus der Person.
- **Degree:** Diese Spalte enthält den höchsten Bildungsabschluss der Person.
- **Relationship:** Diese Spalte beschreibt den Familienstand der Person.
- **Occupation:** Diese Spalte beschreibt den Beruf der Person.
- **Salary (USD):** Diese Spalte enthält das Jahresgehalt der Person in US-Dollar.
- **Children:** Diese Spalte gibt an, wie viele Kinder die Person hat.
- **Colour:** Diese Spalte enthält die Hautfarbe der Person.
- **Gender:** Diese Spalte enthält das Geschlecht der Person.
- **Religion:** Diese Spalte enthält die Religionszugehörigkeit der Person.
- **Country:** Diese Spalte enthält das Heimatland der Person.

3.2 Eingesetzte Prompts

Nachfolgend jeweils eine kurze Beschreibung der eingesetzten Prompts auf die im Weiteren jedoch nicht detaillierter eingegangen wird.

3.2.1 Zero-Shot Prompting

Zero-Shot Prompting ist eine Technik, bei der die AI-Modells Aufgaben ohne explizites Training lösen. Um das Modell zu steuern, werden natürliche Spracheingaben in Form von Anweisungen oder Fragen genutzt. Zero-Shot Prompting ermöglicht die Nutzung von Sprachmodellen für verschiedene Aufgaben, z.B. Textübersetzung, Beantwortung von Fragen und Textgenerierung.

3.2.2 Few-Shot Prompting

Bei einem Few-Shot Prompting lernen die AI-Modells genaue Vorhersagen zu treffen, indem sie auf einer sehr geringen Anzahl von markierten Beispielen trainiert werden. Es wird typischerweise verwendet, um Modelle für Klassifikationsaufgaben zu trainieren, wenn geeignete Trainingsdaten knapp sind.

3.2.3 Template Prompting

Template Prompting (als Variante des One-Shot Prompting) ist ein maschinelles Lernparadigma, bei dem Modelle lernen, neue Objekte oder Muster zu erkennen, basierend auf einem einzigen Beispiel oder einer minimalen Anzahl von Beispielen. Template Prompting zielt darauf ab aus einer einzigen Instanz verallgemeinern können. Die Hauptaufgabe beim Template Prompting besteht darin, Prompts zu entwerfen, so dass das AI-Tools effektiv relevante Merkmale aus den begrenzt verfügbaren Daten extrahieren und genaue Vorhersagen oder Klassifikationen treffen können, wenn sie auf neue, bisher nicht gesehene Beispiele treffen.

3.3 Genutzte LLMs

Nachfolgend eine kurze Beschreibung der bei der Durchführung des Projekts genutzten, generative Large Language Models (LLMs)/AI-Tools:

Feature	BloomZ 7b1-mt	ChatGPT	MS Copilot	LLaMa-2 7B
	BLOOMZ & mT0	ChatGPT	Copilot	
Anzahl der Parameter (10^9)	~7 Milliarden	~175 Milliarden	~175 Milliarden	~7 Milliarden
Trainingsdaten	Text und Code	Text und Code	Text und Code	Text und Code
Zugriffsmöglichkeit	Open-source	Closed-source	Closed-source	Open-source
Sprachunterstützung	54	41	63	52
Entwickelt von	BloomZ Foundation	OpenAI	Microsoft	META
Release Datum	Mai 2023	Oktober 2022	März 2023	Juli 2023
Transformer-Layers	30	96	78	32

Tabelle 4: Features der eingesetzten LLMs

3.3.1 BloomZ

BloomZ ist ein fortschrittliches AI-Tool, das sich durch seine Fähigkeit zur Analyse und Verarbeitung von Daten auszeichnet. Es kann in verschiedenen Anwendungsbereichen eingesetzt werden, von der Datenanalyse bis hin zur Fehlererkennung. BloomZ kann komplexe Datenstrukturen verarbeiten und bietet eine intuitive Benutzeroberfläche für die Interaktion mit den Daten. Es kann auch mit anderen Tools und Plattformen integriert werden, um eine nahtlose Datenverarbeitung zu ermöglichen. BloomZ ist ein großes Sprachmodell (LLM), das auf einem riesigen Datensatz von Text und Code trainiert wurde. Es kann Text generieren, Sprachen übersetzen, verschiedene Arten von kreativen Inhalten schreiben und Ihre Fragen auf informative Weise beantworten.

3.3.2 LLaMa-2

LLaMa-2 ist ein leistungsstarkes AI-Tool, das für seine Fähigkeiten zur Datenverarbeitung und -analyse bekannt ist. Es kann eine Vielzahl von Datenformaten verarbeiten und bietet eine Reihe von Funktionen zur Datenmanipulation und -analyse. LLaMa-2 ist besonders nützlich für Aufgaben, die eine gründliche Analyse und Verarbeitung von Daten erfordern. Meta Llama-2 ist eine Weiterentwicklung der LLaMa-Modelle von Meta AI. Dieses große Sprachmodell (LLM) hat ca. 7 Milliarden Parameter. Es wurde als effizientere und fortschrittlichere Version entwickelt und unterstützt Codegenerierung. LLaMa-2-Modelle sind auf 2 Billionen Tokens trainiert und haben die doppelte Kontextlänge von LLaMa-1.

3.3.3 Copilot

Copilot ist ein AI-gestütztes Tool, das entwickelt wurde, um Entwicklern bei der Codierung zu helfen. Es kann Code in verschiedenen Programmiersprachen generieren und bietet eine Reihe von Funktionen, die das Codieren einfacher und effizienter machen. Copilot kann auch Code überprüfen und Fehler identifizieren, was es zu einem wertvollen Werkzeug für die Codeüberprüfung und -optimierung macht. Microsoft 365 Copilot ist ein fortschrittliches KI-Produktivitätstool, das in Microsoft 365 integriert ist. Als leistungsfähige Verarbeitungsmaschine nutzt es OpenAI's ChatGPT, um natürliche Sprachverarbeitung mit linguistischen Modellen und Daten zu kombinieren. Es bietet Zugang zu GPT-4 und GPT-4 Turbo während der Nicht-Spitzenzeiten, ermöglicht die Verwendung von Text, Sprache und Bildern in der konversationellen Suche.

3.3.4 ChatGPT

ChatGPT ist ein generatives, vortrainiertes Transformer-Modell von OpenAI. Es wurde auf einem riesigen Datensatz von Text und Code trainiert und kann für eine Vielzahl von Aufgaben verwendet werden, einschließlich der Generierung von Text, der Übersetzung von Sprachen und der Beantwortung von Fragen. ChatGPT ist als Web-Applikation verfügbar und daher auch ohne technische Vorkenntnisse leicht zu bedienen. Es ist weniger darauf ausgerichtet, qualitativ hochwertigen Text zu generieren, der sachlich korrekt und für die Anfrage des Nutzers relevant ist. Es bietet erweitertes Verständnis für natürliche Sprache, nahtlose Integration und die Möglichkeit, Bilder zu zeigen und zu erstellen, sowie die Möglichkeit, mit Sprache zu chatten.

3.4 Prüfer

Die beiden Prüfumfänge (Volumes) wurden aufgeteilt und die einzelnen Prüfungen je Volume wurden von den folgenden Teammitgliedern durchgeführt.

Name	Rolle	Prüfanteil/Volume
Taylan Bapur	Prüfer	Vergleich von zwei Datensätzen/Tabellen (separates Dokument)
Gerd Siebert	Prüfer	Fehlersuche in Datensätzen/Tabellen

Tabelle 5: Prüfer

3.5 Umfang der Prüfungen

Das AI-Tool sucht in den Datensätzen/Tabellen nach Fehler und stellt diese dar.

1. Das AI-Tool sucht nach leeren Zellen, Duplikaten, ungültigen und inkonsistenten Werten.
2. Das AI-Tool erzeugt eine Meldung und eine Beschreibung der gefundenen Fehler.
3. Das AI-Tool soll folgende Arten von Fehlern finden:
   - Leere Zellen, Ungültige Werte, Inkonsistente Werte
   - Korrigieren der Fehlerstellen mit dem LLM. Also gegeben dieses Tupel mit einem markierten Fehler, welchen Korrekturvorschlag bietet das einzelne LLM an?
   - Duplikate
   - Abbildung des Vergleichs zweier Records auf das Modell, also die konkrete Frage "Sind (Thorsten, Papenbrock, Uni Marburg, Data Science)" und "(T., Papenbrock, HPI, Database Systems)" dieselben Entitäten?

3.5.1 Nicht zu prüfende Funktionalitäten

Nicht zu prüfende Funktionalitäten waren u.a.:

1. Datenzusammenfassung
   Zusammenfassung der wichtigsten Informationen aus den Datensätzen/Tabellen, z.B. die Anzahl der Datensätze, den Durchschnittswert jeder Spalte sowie die Höchst- und Mindestwerte, sofern diese nicht für die Analyse selbst erforderlich waren.
2. Konsistenzprüfungen über Summen hinweg
   Überprüft wurde die Konsistenz der Daten über verschiedene Spalten und Zeilen hinweg, z.B., ob die Summe der Werte in einer Spalte mit den in der Tabelle ausgewiesenen Summen übereinstimmte.
3. Datenvisualisierung
   Eine Erstellung von Visualisierungen der Daten, um Trends, Muster und Beziehungen besser zu verstehen, erfolgte nicht.
4. Prüfungen der Datenintegrität mit externen Quellen
   Eine Überprüfung der Richtigkeit und Vollständigkeit der Daten durch den Vergleich der Daten mit externen Quellen oder mit Hilfe statistischer Methoden erfolgte nicht.

5. Prüfungen über mehrere Dateiformate hinweg
   Für jede Tabelle wurde nur ein Dateiformat, nämlich CSV überprüft. Eine Prüfung einzelner Datensätze/Tabellen unter Zugrundelegung aller in der Untersuchung genutzten Dateiformate erfolgte nicht.
6. Benutzerfreundlichkeit
   Die Benutzerfreundlichkeit, insbesondere der Ein- und Ausgabemöglichkeiten der einzelnen AI-Tools, wurde in die Analyse ihrer Leistungsfähigkeit nicht mit einbezogen.

3.6 Prüfkriterien für das Bestehen/Nichtbestehen

Ein Prüfung bei der Fehlersuche galt als bestanden, wenn das AI-Tool alle Anforderungen vollständig erfüllte. Eine Prüfung galt als nicht oder teilweise bestanden, wenn das AI-Tool eine oder mehrere Anforderungen nicht erfüllte. Zur Bewertung der Erfüllung wurden die Merkmale „erfüllt“, „(bedingt) erfüllt“ und „nicht erfüllt“ vergeben.

Die Nichterfüllung von Anforderungen bedeutete jedoch nicht, dass eine abschließende Aussage über die Nutzbarkeit des jeweiligen AI-Tools für die geprüfte Anforderung getroffen werden konnte. Deswegen wurde bei einer teilweisen Erfüllung der Anforderungen eine prozentuale Bewertung der Antwort (Score) vorgenommen. Damit gibt der Score einen besseren Überblick über die tatsächliche Nutzbarkeit der Antwort eines jeden AI-Tools.

3.7 Abbruchkriterium

Prüfungen wurden abgebrochen, wenn vom eingesetzten AI-Tool 50 % oder mehr der implementierten Fehler nicht gefunden wurden. Über einen Abbruch hat jeder Prüfer selbständig entschieden.

3.8 Prüfphasen

Im Wesentlichen erfolgten die Prüfungen in zwei Phasen. In der Vorbereitungsphase wurde sich mit dem Projekt und den AI-Tools vertraut gemacht. In der nachfolgenden operativen Durchführungsphase erfolgte dann die eigentliche Prüfung nach ein paar Vortests.

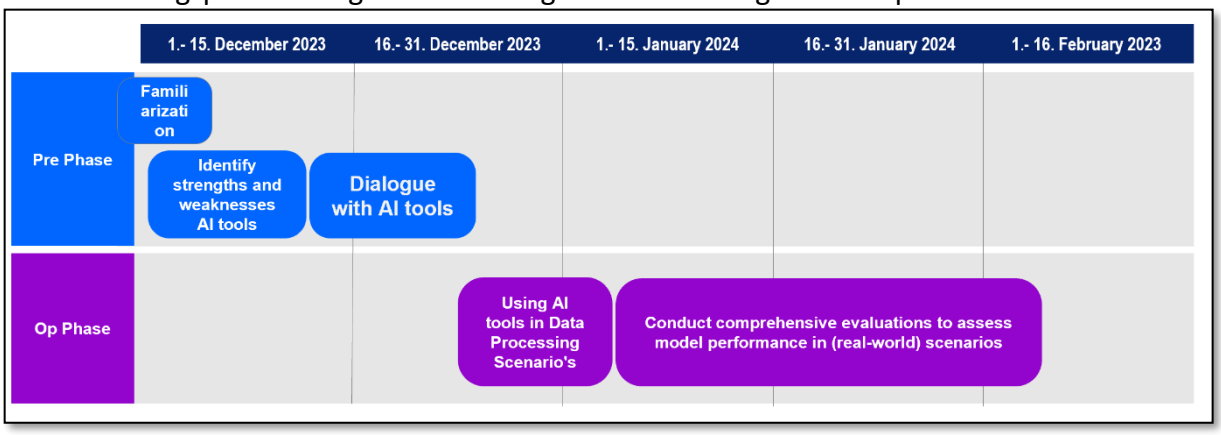

Abbildung 1: Projektphasen (englische Version)

3.9 Restriktionen

Die Durchführung der Prüfungen war mit einer Reihe von Einschränkungen verbunden, die im Folgenden kurz beschrieben werden.

3.9.1 Stichprobenumfang

Die Free-Versionen der AI-Tools erlaubten nur das Hochladen einer begrenzten Anzahl von Datensätzen/Tabelleninstanzen. Dies erforderte manuelle Anpassungen der genutzten Datensätze/Tabellen, was den Prozess zeitaufwendig und weniger effizient machte. Darüber hinaus waren die Anzahl der Instanzen und Attribute in den Datensätze/Tabellenumfänge zu klein für grundlegende statistische Auswertungen/Statistiken, was die Aussagekraft der Ergebnisse begrenzt.

Schließlich wurden je Volume weniger als 20 spezifische Datensatz-/Tabellenmanipulationen vorgenommen, was ebenfalls die Aussagekraft der Untersuchung limitiert und man daher nur tendenziell etwas zu den Einsatzmöglichkeiten der verwendeten AI-Tools zur Datensatz/Tabellenbereinigung aussagen kann.

3.9.2 Genutzte Datenformate

Als Datenformat **i**n dieser Untersuchung wurde ausschließlich das CSV-Format für die Datensätze/Tabellen verwendet. Dies könnte die Aussagekraft und Übertragbarkeit der Ergebnisse auf andere Datenformate einschränken.

3.9.3 Anzahl der Versuche

Es wurden grundsätzlich keine Mehrfachversuche durchgeführt, um etwaige “Lerneffekte” auszuschließen und eine Gleichbehandlung aller AI-Tools zu gewährleisten. Wenn Mehrfachversuche durchgeführt wurden, dann im Allgemeinen, weil es sich um einen Abbruch der Antwort des AI-Tools handelte. In einem Fall wurde durch „Nachfragen“ beim Einsatz von Copilot eruiert, ob sich die Antwort des AI-Tools noch weiter verbessert.

3.9.4 Mangelnde Kenntnisse der Prüfer beim Prompting

Die Prüfer verfügten über keine spezifische Ausbildung im Bereich des Prompting. Dies könnte die Qualität der Eingabeaufforderungen und damit die Qualität der von den AI-Tools generierten Antworten beeinflusst haben.

3.9.5 Projektdauer

Die zur Verfügung stehende Zeit für eine umfassende Analyse war begrenzt (nur etwa 6 Wochen effektive Prüfzeit). Dies hat die Tiefe und Breite der durchgeführten Tests und Analysen eingeschränkt. Während der Laufzeit, am 14.12.2023, wurde das lokal deployte AI-Tool BloomZ durch das lokal deployte AI-Tool LLaMa-2 auf dem Server des FB12 ersetzt.

Zusammenfassend führten diese Restriktionen dazu, dass eine umfassende Analyse der Fähigkeiten der AI-Tools zur Datenanalyse, Fehlererkennung und Bereinigung der Daten deutlich limitiert war. Die diesbezügliche Bewertung der Leistungsfähigkeit der AI-Tools ist somit im Rahmen dieses Projekts mit all seinen Restriktionen nur als tendenziell zu betrachten.

4 Ergebnisse der Fehlersuche in Datensätzen/Tabellen

In der Welt von Big Data Analytics ist die Fehlererkennung ein entscheidendes Merkmal. Die Fehlererkennung in Datensätzen und Tabellen ist vergleichbar mit dem Lösen eines komplexen Puzzles. Jeder Fehler repräsentiert ein fehlendes Puzzleteil, das die Gesamtbildqualität beeinträchtigt.

Die Fähigkeit von AI-Tools auf ihre Fähigkeiten zur Fehlererkennung und -behebung ist entscheidend, um fundierte Entscheidungen auf genauen und zuverlässigen Daten zu ermöglichen. Der Prozess der Fehleridentifikation und -korrektur trägt damit zur Steigerung der Effizienz und Präzision bei.

4.1 Prüfszenarien

Leerstellen, Duplikate, ungültige Werte und inkonsistente Werte stellen verschiedene Arten von Fehlern dar, die die Datenqualität beeinflussen. Diese Fehler sind die Herausforderungen, denen sich die genutzten AI-Tools auf dem Weg zur perfekten Datenanalyse stellen mussten.

Die nachfolgende Abbildung 2 illustriert beispielhaft verschiedene Fehlerarten, die in die Ausgangstabelle als Test-Setup eingebracht wurden, um die Effizienz der AI-Tools bei der Fehlererkennung zu testen. Dabei gab es vier Hauptkategorien von Fehlern: leere Zellen, Duplikate, ungültige Werte und inkonsistente Werte.

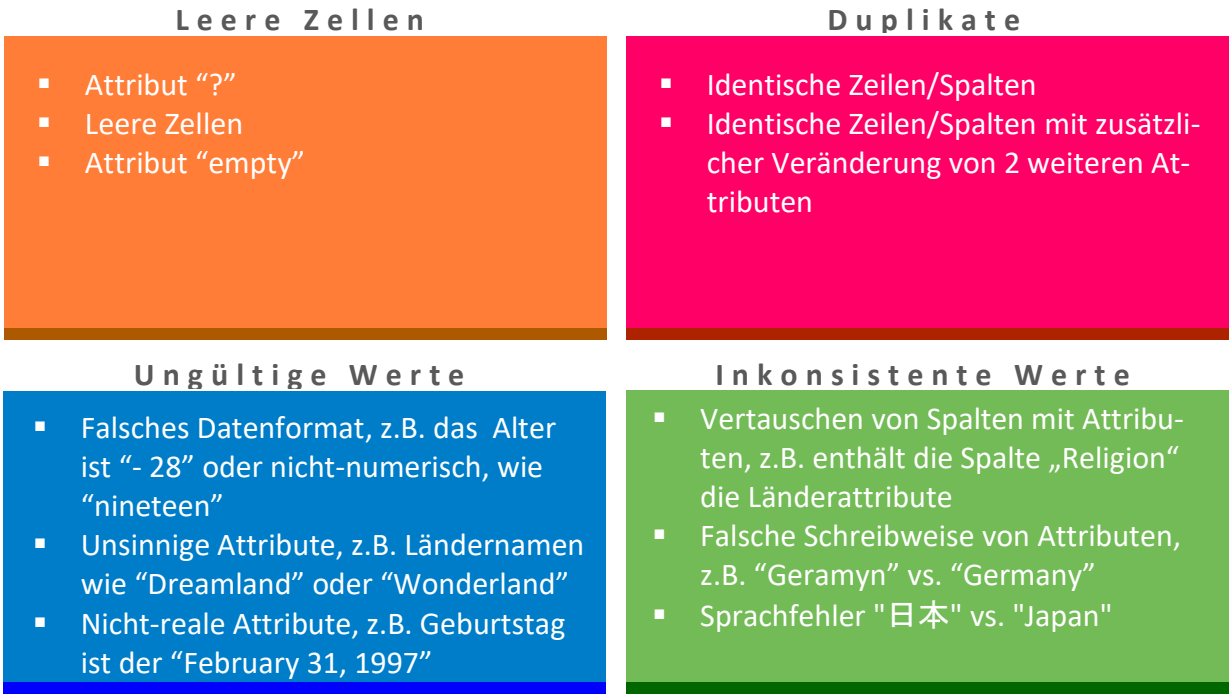

Abbildung 2: Datensatz/Tabellen-Manipulationen (Auszug)

Leere Zellen können z.B. durch das Attribut „?“, tatsächlich leere Zellen oder das Attribut „empty“ repräsentiert werden. Duplikate sind identische Zeilen oder Spalten sowie solche mit geringfügigen Änderungen in weiteren Attributen. Ungültige Werte beinhalten unter anderem falsche Datenformate wie nicht-numerische Altersangaben oder unsinnige Attribute wie unrealistische Ländernamen und nicht existierende Geburtsdaten. Inkonsistente Werte treten auf, wenn Spalten oder Zeilen u.a. mit Attributen vertauscht werden oder falsche Schreibweisen und Sprachfehler vorhanden sind.

Diese Fehlerarten stellen eine Herausforderung für AI-Tools dar, da sie die Datenqualität beeinträchtigen und zu falschen Analysen führen können. Durch das Testen der AI-Tools auf diese Fehler können wir ihre Effizienz bei der Fehlererkennung zumindest tendenziell im vorhandenen Projektszenario bewerten. Dies trägt mit dazu bei, die Genauigkeit der Datenanalyse und damit die Genauigkeit und Brauchbarkeit des jeweiligen AI-Tools zu beurteilen, um fundierte Entscheidungen hinsichtlich der Nutzung solcher Werkzeuge treffen zu können.

4.2 Einzelauswertungen

Das Abbildung 3 zeigt die Ergebnisse der Untersuchung der Leistung der verwendeten AI-Tools hinsichtlich ihres Scores unter Berücksichtigung des eingesetzten Prompting im Detail. Die Analyse umfasste die vier bereits genannten Kategorien:

- Leere Zellen (LZCSVG) mit 3 Einzelszenarien
- Duplikate (DUCSVG) mit 2 Einzelszenarien
- Ungültige Werte (UWCSVG) mit 3 Einzelszenarien
- Inkonsistente Werte (IVCSVG) mit 10 Einzelszenarien

Analyseumfang	Empty Cells (LZCSVG)				Duplicates (DUCSVG)			
AI-Tool	Zero-Shot	Few-Shot	Template	Summe	Zero-Shot	Few-Shot	Template	Summe
BloomZ	27%	0%	10%	12%	5%	0%	5%	3%
ChatGPT	73%	90%	90%	84%	95%	50%	90%	78%
Copilot	83%	100%	100%	94%	100%	100%	100%	100%
LLaMa-2	0%	33%	0%	11%	20%	5%	35%	20%

Analyseumfang	Invalid Values (UWCSVG)				Inconsistent Values (IVCSVG)			
AI-Tool	Zero-Shot	Few-Shot	Template	Summe	Zero-Shot	Few-Shot	Template	Summe
BloomZ	23%	23%	3%	17%	0%	abgebrochen	0%	0%
ChatGPT	93%	73%	87%	84%	70%	70%	56%	65%
Copilot	100%	100%	100%	100%	100%	94%	100%	98%
LLaMa-2	37%	0%	0%	12%	0%	abgebrochen	0%	1%

Abbildung 3: Auswertung der einzelnen Fehlervarianten

Jede Kategorie wurde in mehreren Szenarien mit unterschiedlichen Komplexitätsgraden untersucht. Jedes AI-Tool wurde mit drei Methoden Zero-Shot, Few-Shot und Template bewertet.

Die folgenden sieben Punkte fassen die wichtigsten Ergebnisse der Untersuchung bezogen auf ihre einzelne Ausprägung zusammen:

- BloomZ und LLaMa-2 zeigten in den meisten Kategorien und Szenarien eine signifikant geringere Leistung als ChatGPT und Copilot.
- ChatGPT und Copilot erreichten in den Kategorien "Leere Zellen" und "Ungültige Werte" eine hohe bis sehr hohe Genauigkeit (jeweils über 80%).
- Die Genauigkeit von ChatGPT und Copilot nahm in den Kategorien "Duplikate" und "Inkonsistente Werte" nur ganz leicht ab.
- Copilot erreichte in allen Kategorien die höchste Gesamtleistung (zwischen 94% und 100%).
- BloomZ und LLaMa-2 zeigten in allen Kategorien die geringste Leistungsfähigkeit (meist deutlich unter 20%), was letztlich zum Abbruch der Untersuchung für diese AI-Tools (Nichterreichen der 50% Schwelle) führte.
- ChatGPT und Copilot lagen immer über 60% und damit über dem Abbruchkriterium der 50% Schwelle.
- Die Leistung der AI-Tools nahm mit der Komplexität und der Anzahl der Szenarien ab, wobei Copilot allerdings nur einen marginalen Leistungsabfall zeigte.

Insgesamt zeigte sich bei der vergleichenden Analyse der vier AI-Tools in Bezug auf ihre Leistung bei den vier verschiedenen Test-Szenarien zur Datenvalidierung, dass BloomZ und LLaMa-2 die 50% Schwelle nicht überschreiten konnten. Sie wurden deshalb wurden ab den Tests 220 - 222 fortfolgend abgebrochen. Wohingegen ChatGPT und Copilot durchweg überlegene Leistungen in allen Kategorien zeigten, wobei Copilot in den meisten Fällen eine Erfolgsquote von 100% erreichte. Weitere Details sind den Anhängen „A – Einzelperformance“ und „B - Einzel-Scores Fehlersuche“ zu entnehmen.

4.3 Gesamtergebnis

Die nachfolgende Analyse vergleicht die Leistung von vier AI-Tools: BloomZ, ChatGPT, Copilot und LLama-2. Dabei wurde die Leistung wieder in den drei Kategorien “nicht erfüllt”, “(bedingt) erfüllt” und “erfüllt” bewertet.

Die Abbildung 4 zeigt die Anzahl der Testfälle, die jedes Tool nicht erfüllt, bedingt erfüllt oder erfüllt hat.

AI-Tool	nicht erfüllt	(bedingt) erfüllt	erfüllt
BloomZ	27	5	1
ChatGPT	4	24	26
Copilot	0	3	51
LLaMa-2	25	6	2

Abbildung 4: Vollständig erfüllte Datensatz-/Tabellenaufgaben

BloomZ hat mit 27 die höchste Anzahl an nicht erfüllten Testfällen. Auch LLaMa-2 kämpft mit einer hohen Anzahl an nicht erfüllten Testfällen. ChatGPT zeigt eine durchwachsene Leistung mit einer Vielzahl an erfüllten Testfällen, aber auch mit einer doch beträchtlichen Anzahl an nur bedingt erfüllten Testfällen. Im Gegensatz dazu kann Copilot mit 51 erfüllten Testfällen alle anderen AI-Tools deutlich übertreffen.

Die nebenstehende Perfomance-Grafik zeigt nochmal das Gesamtergebnis jedes AI-Tools hinsichtlich ihres Anteils an nicht erfüllten, (bedingt) erfüllten und erfüllten Aufgaben.

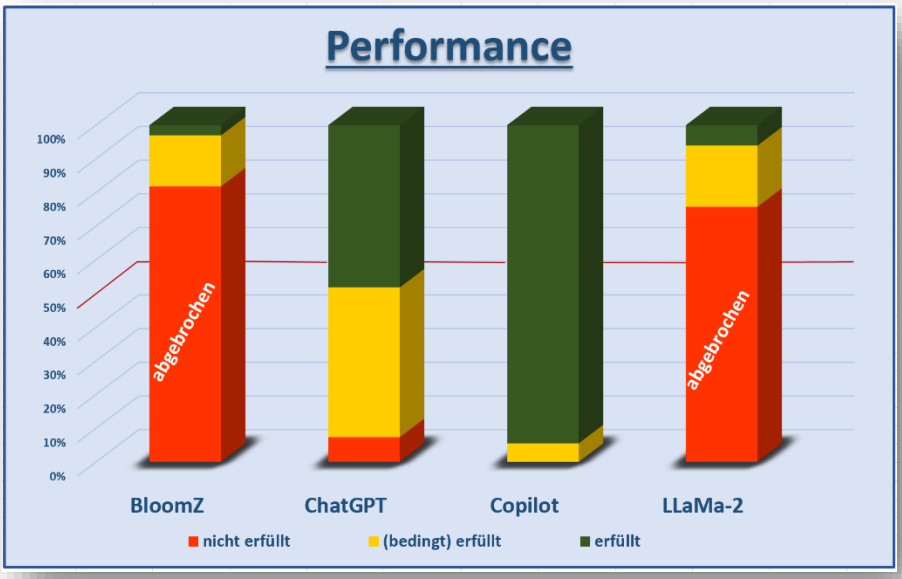

Abbildung 5: Performance Fehlersuche

Unter Beachtung des 50% Abbruchkriteriums ist deutlich zu erkennen, dass BloomZ und LLaMa-2 signifikant diese Grenze (rote Fläche), aufgrund ihrer vielen nicht erfüllte Aufgaben, überschreiten.

ChatGPT dagegen hat nur wenige nicht erfüllte Aufgaben und einen großen Teil vollständig erfüllte Aufgaben. Allerdings hat ChatGPT auch einen signifikanten Anteil von nur (bedingt) erfüllten Aufgaben, was zeigt, dass mit den Antworten des AI-Tools, die Aufgaben teilweise nur unvollständig erfüllt werden konnten.

Die fast vollständig grüne Fläche von Copilot zeigt die herausragende Performance, was auch durch die hohe Anzahl von vollständig erfüllten Aufgaben bestätigt wird.

Das Gesagte spiegelt sich auch im Gesamtscore (Abbildung 6) der einzelnen AI-Tools wider. Die Gesamtpunktzahl zeigt, dass Copilot mit fast 100% Effizienz am besten abschneidet, gefolgt von ChatGPT mit etwa 78%. LLama-2 und BloomZ liegen mit etwa 11% bzw. 8% deutlich zurück.

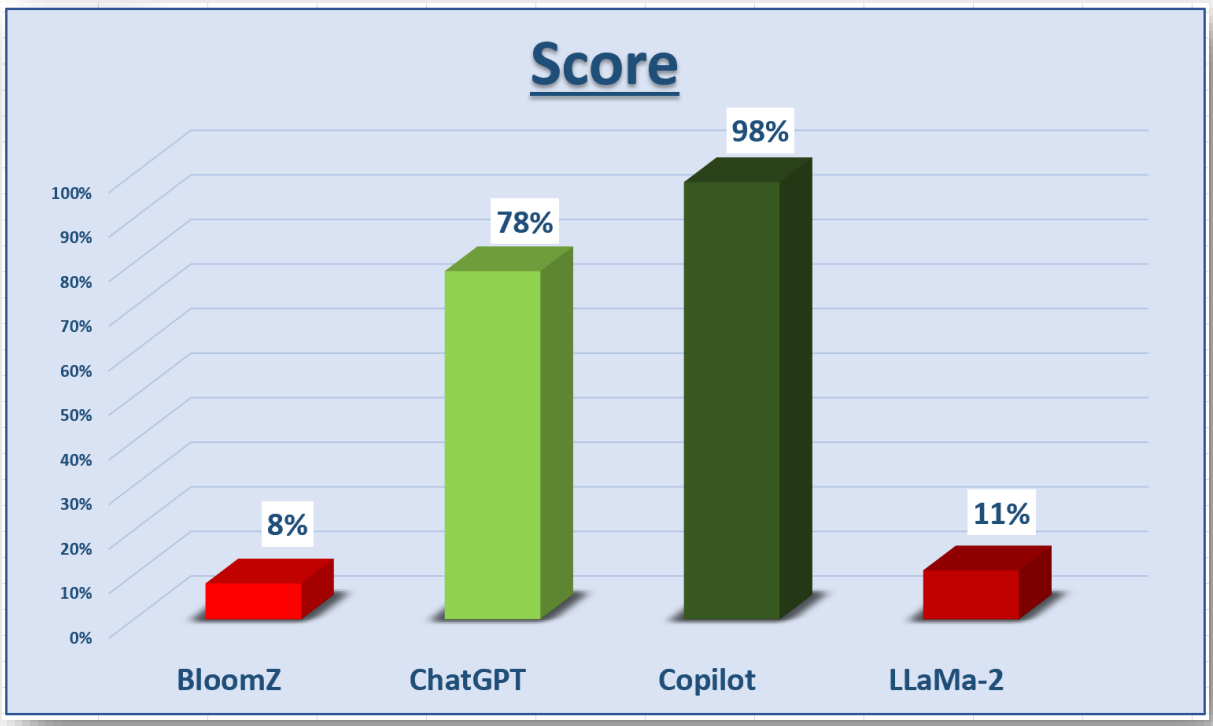

Abbildung 6: Score Fehlersuche

Diese Ergebnisse zeigen, dass Copilot das effizienteste Tool war, während BloomZ und LLaMa-2 deutliche Schwächen zeigten. BloomZ und LLaMa-2 sollten daher tendenziell für die Analyse von Tabellen mit Vorsicht verwendet werden, da ihre Leistung in den meisten Kategorien und Szenarien außerordentlich gering ist.

Trotz des hohen Anteils an (bedingt) erfüllten Aufgaben, erreicht ChatGPT einen diesbezüglich überdurchschnittlichen Score. Dies zeigt, dass die Antwort von ChatGPT zwar in vielen Fällen nicht ausreichten um die Aufgabe vollständig zu lösen, jedoch einen hinreichenden Informationsgehalt hatten, um zumindest einen Großteil der Aufgabenstellung zu lösen.

Wie die Einzeltests schon gezeigt haben, deuten die Daten auf unterschiedliche Effizienz- und Effektivitätsniveaus der einzelnen AI-Tools bei der Bewältigung spezifischer Aufgaben oder Probleme zur Datensatz-/Tabellenbereinigung hin.

Schlussendlich lässt sich sagen, dass BloomZ mit einem erzielten Score von nur 8% und einer Bewertung, die in den meisten Kategorien ein “nicht erfüllt” erhalten hat, nicht für die Datensatz-/Tabellenreinigung geeignet erscheint. Diese schlechte Performance führte auch dazu, dass die Untersuchung mit BloomZ frühzeitig abgebrochen wurde. Die Ergebnisse deuten also darauf hin, dass BloomZ möglicherweise nicht die beste Wahl für Aufgaben ist, die eine so hohe Genauigkeit und Vollständigkeit erfordern.

Auch LLaMa-2 erzielte einen Score von nur 11% und zeigte eine Leistung, die in etwa im Bereich von BloomZ lag. Deshalb wurde auch bei diesem AI-Tool die Nutzung im Laufe der Untersuchungen abgebrochen. Ebenso, wie bei BloomZ, zeigte sich, dass LLaMa-2 wohl nicht die beste Wahl für solche Aufgaben ist.

ChatGPT hingegen erzielte einen Score von 78% und zeigte eine solide Leistung über die verschiedenen Kategorien hinweg. Obwohl es einige Bereiche gab, in denen es nur bedingt erfüllte, war es in der Lage, hilfreiche Antworten zu generieren, die die Bereinigung der Datensätze/Tabellen erheblich erleichterten. Da ChatGPT in der Lage war in vielen Bereichen zu erfüllen, deutet dies auf eine robuste und vielseitige Leistungsfähigkeit und Brauchbarkeit hin.

Copilot übertraf die anderen AI-Tools mit einem Score von 98% bei weitem und erfüllte fast alle Kategorien zu 100%. Auf Basis dieser hervorragenden Leistung und Zuverlässigkeit scheint Copilot ein effektives AI-Tool für eine Vielzahl von Aufgaben in diesem Bereich zu sein, insbesondere wenn hohe Genauigkeit und Vollständigkeit erforderlich sind.

4.4 Erfahrungen mit den AI-Tools

In der Untersuchung der verschiedenen AI-Tools wurden im Bereich der Fehlersuche folgende Beobachtungen gemacht:

BloomZ: Dieses Tool hat oft verwirrende Antworten geliefert, die durch Halluzinationseffekte gekennzeichnet schienen. Darüber hinaus passten die Antworten oft nicht zur gestellten Frage, was wohl auf eine mangelnde Kontextsensitivität hinweist.

LLaMa-2: Bei der Verwendung von LLaMa-2 traten häufig sehr verwirrende Antworten auf, die ebenfalls durch Halluzinationseffekte gekennzeichnet waren. Mehrfache Anläufe über Huggingface waren oft notwendig, und es gab immer wieder Fehlermeldungen aufgrund der Nichtverfügbarkeit einer GPU. Manchmal waren die Antworten nicht nachvollziehbar oder passten nicht zur Frage. Es gab auch Fälle, in denen tagelang kein Zugriff über Huggingface möglich war oder die Antwort mitten im Prozess abbrach.

ChatGPT: ChatGPT zeigte eine teilweise langsame Reaktionszeit. Manchmal waren die Antworten sehr ausschweifend und unübersichtlich, was die Interpretation erschwerte. Es gab auch mehrere Abbrüche, die durch das Auftreten eines “object Object” gekennzeichnet waren. In einigen Fällen waren deshalb weitere Versuche erforderlich, um eine Antwort zu erhalten. Insgesamt waren die Antworten aber sehr hilfreich, wenn auch nicht immer vollständig.

Copilot: Im Vergleich zu den anderen Tools war Copilot meist sehr schnell. Es gab jedoch eine Verlangsamung, wenn eine korrigierte Tabelle ausgegeben wurde. Die Antworten waren meist kompakt und fast ausschließlich punktgenau, was auf eine hohe Genauigkeit und Effizienz hinweist. Die Interaktion mit Copilot war äußerst erfreulich, da es hinsichtlich der Genauigkeit und Vollständigkeit der bereitgestellten Antworten signifikant den anderen AI-Tools überlegen war. Darüber hinaus wurde beobachtet, dass in den Fällen, in denen das Copilot die Anfragen nicht vollständig oder korrekt beantwortete, eine einmalige Nachfrage den Mangel aus der ersten Antwort abstellte und eine signifikante Verbesserung des Ergebnisses bewirkte.

4.5 Prompting Strategien

Die Wahl der Prompt-Technik kann einen Einfluss auf die Ergebnisqualität bei der Tabellenanalyse haben. Daher war es wichtig, verschiedene Prompt-Techniken auszuprobieren, um die beste Lösung für die jeweilige Aufgabenstellung zu finden.

Die Tabelle 5 zeigt eine vergleichende Darstellung der Leistung der vier AI-Tools (BloomZ, ChatGPT, Copilot, LLama-2) bei der Bewältigung verschiedener Datenprobleme (Leere Zellen (LZCSVG), Duplikate (DUCSVG), Ungültige Werte (UWCSVG), Inkonsistente Werte (IVCSVG)) unter Verwendung der eingesetzten Prompt-Techniken:

- **Zero-Shot:** Keine explizite Anleitung, nur allgemeine Beschreibung der Aufgabe.
- **Few-Shot:** Wenige Beispiele zur Veranschaulichung der gewünschten Ausgabe.
- **Template:** Vordefinierte Vorlage, die mit spezifischen Informationen gefüllt wird.

Die Betrachtung zeigt aber, dass die Wahl der Prompt-Technik in dieser Untersuchung keinen signifikanten Einfluss auf die Ergebnisqualität bei der Tabellenanalyse hat. Jedoch basiert diese Analyse nur auf einer begrenzten Anzahl von Aufgaben und Szenarien, so dass weitere Untersuchungen mit größeren Datenmengen und komplexeren Szenarien notwendig wären, um bessere Aussagen zum Einfluss von Prompt-Techniken auf das Ergebnis treffen zu können.

Darüber hinaus ist anzumerken, dass die Prüfer keine ausgebildeten Prompt Engineers waren. Die Erstellung von effektiven Prompts erfordert Expertise und Erfahrung. Es ist daher wahrscheinlich, dass die Ergebnisse durch den Einsatz von professionellen Prompt Engineers hätten bei einigen AI-Tools und Aufgabenstellungen noch verbessert werden könnten.

Zudem ist zu beachten, dass diese Untersuchung nur einen kleinen Ausschnitt der möglichen Prompt-Techniken abdeckt. Es gibt eine Vielzahl weiterer Techniken, die in zukünftigen Studien untersucht werden könnten.

Weitere Forschungen erscheinen notwendig, um den Einfluss von Prompt-Techniken auf die Ergebnisqualität besser zu verstehen. In zukünftigen Untersuchungen sollten größere Datenmengen und komplexere Szenarien verwendet werden, um die Ergebnisse zu verallgemeinern.

4.6 Abbruch der Tests bei BloomZ und LLaMA-2

Die Tests mit BloomZ und LLaMa-2 wurden nach jeweils 11 von 18 Test abgebrochen, da die Leistung beider Tools unter 50% korrekter Antworten sank. Deshalb wurden insgesamt nur 174 von 216 geplanten Tests ausgeführt.

Dies weist auf eine Reihe von Problemen bei diesen AI-Tools hin, darunter möglicherweise technische Schwierigkeiten, unzureichende Anpassungsfähigkeit der AI-Tools an die Testbedingungen oder eine unzureichende Leistung bei komplexeren Aufgaben. Mögliche Gründe für die schlechte Leistung von BloomZ und LLaMa-2 könnten sein:

Die Architektur der AI-Modelle: BloomZ und LLaMa-2 basieren auf anderen Architekturen als die von ChatGPT und Copilot, die sich möglicherweise besser für die Tabellenanalyse eignen. Auch liegt z.B. die Anzahl der Parameter mit nur ca. 7 Milliarden Parametern auf der Seite der von BloomZ und LLaMa-2 signifikant unter der Parameteranzahl von ChatGPT und Copilot mit jeweils etwa 175 Milliarden Parametern.

Die Datensätze, mit denen die AI-Tools trainiert wurden: BloomZ und LLaMa-2 wurden möglicherweise mit Datensätzen trainiert, die nicht für die Tabellenanalyse geeignet sind.

Die verwendeten Prompt-Techniken: Die verwendeten Prompt-Techniken waren möglicherweise für BloomZ und LLaMa-2 nicht optimal für das Lösen der jeweiligen Aufgaben.

Aber auch hier ist es wichtig zu beachten, dass die Tests dieser Untersuchung nur auf einer begrenzten Anzahl von Tabellen und Szenarien basieren. Es ist daher möglich, dass die Leistung von BloomZ und LLaMa-2 bei anderen Aufgaben deutlich besser ist.

Weitere Untersuchungen wären notwendig, um die Gründe für die schlechte Leistung von BloomZ und LLaMa-2 bei der Tabellenanalyse zu verstehen. Nichtsdestotrotz haben AI-Tools ein großes Potenzial für die Analyse von Tabellen, aber es ist wichtig, das richtige AI-Tool für die jeweilige Aufgabe auszuwählen.

5 Zusammenfassung

Zwei der eingesetzten AI-Tools haben sich als effektive Instrumente für die Analyse von Datensätzen und Tabellen erwiesen. So haben die LLMs ChatGPT und speziell Copilot in diesem Bereich bemerkenswerte Leistungen erbracht.

5.1 Erkenntnisse

Es ist jedoch wichtig zu beachten, dass die vorliegenden Erkenntnisse auf einer begrenzten Anzahl von Tabelleninstanzen und Attributen basieren und ausschließlich das CSV-Format verwendet wurde. Daher können diese Ergebnisse lediglich als Indikatoren für eine allgemeine Tendenz betrachtet werden.

Die Frage, welche Leistungsfähigkeit diese AI-Tools bei einer größeren Datenmenge, beispielsweise 10.000 Instanzen mit möglicherweise 300.000 und mehr Attributen, oder bei der Verwendung anderer Formate wie JSON aufweisen würden, bleibt offen und erfordert weitere Untersuchungen. Es ist daher nicht möglich, eine abschließende Antwort auf diese Frage zu geben.

AI-Tools, die frei verfügbar sind, weisen oft bestimmte Einschränkungen auf, insbesondere in Bezug auf die Eingabemöglichkeiten. Diese Begrenzungen können sich auf die Anzahl der Zeichen oder die Anzahl der täglich möglichen Versuche beziehen. Es ist daher wichtig, diese Faktoren bei der Verwendung dieser LLMs zu berücksichtigen.

Unsere Untersuchungen haben gezeigt, dass die Strategien zur Eingabeaufforderung (Prompts) nur einen geringen Einfluss auf das Gesamtergebnis eines AI-Tools zu haben scheinen. Dies deutet darauf hin, dass andere Faktoren, wie die spezifischen Algorithmen des Tools und das speziell trainierte neuronale Netz, wohl eine größere Rolle spielen.

In unseren Tests haben sich die getesteten Versionen der AI-Tools BloomZ und LLaMa-2 als nicht sehr effektiv für die Analyse von Datensätzen und Tabellen erwiesen. Tatsächlich scheinen diese Versionen für die Analyse von Datensätzen und Tabellen grundsätzlich ungeeignet zu sein. Dies unterstreicht die Notwendigkeit, die Auswahl des richtigen AI-Tools für spezifische Aufgaben sorgfältig zu prüfen.

Ein weiteres Problem, das wir festgestellt haben, ist das Fehlen von Upload-Möglichkeiten für große Datensätze und Tabellen in den frei verfügbaren AI-Tools. Dies hat verhindert, dass Datensätze und Tabellen mit mehreren Hundert oder mehreren Tausend Instanzen geprüft werden konnten. Somit war es uns nicht möglich die Fähigkeit dieser AI-Tools, mit großen Datenmengen umzugehen, vollständig zu testen.

Ebenso ist es schwierig zu quantifizieren, inwieweit die mangelnden oder eingeschränkten Fähigkeiten der Prüfer, zum Beispiel im Bereich des Prompt-Engineering, das Gesamtergebnis beeinflusst haben. Dies ist ein wichtiger Aspekt, der in zukünftigen Untersuchungen berücksichtigt werden sollte. Es ist daher unerlässlich, die Fähigkeiten und das Wissen der Prü-

fer kontinuierlich zu verbessern und zu erweitern, um die Qualität der Ergebnisse zu optimieren. Schließlich verhinderte fehlendes Budget die Analyse von speziellen AI-Tools, die besonders für die Tabellenanalyse trainiert sind (wie z.B. Jellyfish, MS Power BI, ...).

5.2 Empfehlungen

Basierend auf den vorliegenden Erkenntnissen können folgende Empfehlungen abgeleitet werden.

Es besteht ein deutlicher Bedarf an weiteren Tests und Analysen. Insbesondere sollten mehr Testfälle mit größeren Datensätzen und Tabellen sowie mit verschiedenen Datenformaten, wie zum Beispiel JSON, durchgeführt werden. Dies würde dazu beitragen, ein umfassenderes Verständnis der Leistungsfähigkeit dieser AI-Tools zu erlangen.

Die getesteten Versionen der AI-Tools BloomZ und LLaMa-2 haben sich als nicht sehr effektiv für die Analyse von Datensätzen und Tabellen erwiesen. Dies unterstreicht die Notwendigkeit, die Auswahl des richtigen AI-Tools für spezifische Aufgaben sorgfältig zu prüfen.

Ein weiteres Problem ist das Fehlen von Upload-Möglichkeiten für große Datensätze und Tabellen in den frei verfügbaren AI-Tools. Für den wirtschaftlichen Erfolg ist es jedoch unerlässlich, dass große Datenmengen mit mehreren 10.000 Instanzen und Hunderttausenden von Attributen in den verschiedensten Datenformaten bereinigt werden können. Zukünftige Untersuchungen sollten daher möglichst lokal deployte AI-Tools berücksichtigen, die diese Funktionen unterstützen, um ein vollständigeres Bild der Fähigkeiten dieser AI-Tools zu erhalten.

Fehlendes Budget verhinderte die Analyse von speziellen AI-Tools, die besonders für die Tabellenanalyse trainiert sind. Daher sollten zukünftige Untersuchungen, sofern das Budget es zulässt, diese speziellen AI-Tools in die Analyse einbeziehen.

Schließlich sollte in zukünftigen Untersuchungen berücksichtigt werden, inwieweit die Fähigkeiten der Prüfer das Gesamtergebnis beeinflussen. Es ist daher unerlässlich, die Fähigkeiten und das Wissen der Prüfer kontinuierlich zu verbessern und zu erweitern, um die Qualität der Ergebnisse zu optimieren.

Zusammenfassend lässt sich sagen, dass weitere Untersuchungen erforderlich sind, um ein vollständigeres Bild der Leistungsfähigkeit dieser AI-Tools zu erhalten. Dabei sollten verschiedene Aspekte berücksichtigt werden, darunter die Größe und das Format der Daten, die spezifischen Algorithmen und neuronalen Netze der AI-Tools, die Fähigkeiten der Prüfer und die Verfügbarkeit von speziellen AI-Tools für die Tabellenanalyse und Datenreinigung.

Diese Analyse bietet einen wertvollen Überblick über die Leistung von AI-Tools für die Tabellenanalyse und -bereinigung. Wenn auf die Untersuchung zahlreiche Einschränkungen hatte, können die Ergebnisse Unternehmen und Organisationen bei der Auswahl des richtigen AI-Tools für ihre Bedürfnisse unterstützen.

Anhang A – Individuelle Performance Fehlersuche

Nachfolgende Tabellen zeigen die einzelnen Performanceausprägungen, die jedes AI-Tool individuell bei jedem Test unter Nutzung der drei verschiedenen Promptingvarianten Zero-Shot, Few-Shot und Template erreicht hat.

Zero-Shot

Analyseumfang	LZCSVG			DUCSVG		UWCSVG			IVCSVG									
Testfallnummer	1	20	40	60	80	100	120	140	160	180	200	220	240	260	280	300	320	340
BloomZ	erfüllt	nicht erfüllt	nicht erfüllt	(bedingt) erfüllt	nicht erfüllt	nicht erfüllt	(bedingt) erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	abgebrochen, wurde am 14.12. ersetzt durch LLaMa-2						
ChatGPT	erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	(bedingt) erfüllt	nicht erfüllt	erfüllt	erfüllt	(bedingt) erfüllt	erfüllt	erfüllt	(bedingt) erfüllt	erfüllt	(bedingt) erfüllt	(bedingt) erfüllt
Copilot	erfüllt	erfüllt	(bedingt) erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt
LLaMa-2	nicht erfüllt	nicht erfüllt	nicht erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	nicht erfüllt	erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	abgebrochen						

Few-Shot

Analyseumfang	LZCSVG			DUCSVG		UWCSVG			IVCSVG									
Testfallnummer	2	21	41	61	81	101	121	141	161	181	201	221	241	261	281	301	321	341
BloomZ	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	(bedingt) erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	abgebrochen, wurde am 14.12. ersetzt durch LLaMa-2						
ChatGPT	erfüllt	(bedingt) erfüllt	erfüllt	erfüllt	nicht erfüllt	erfüllt	erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	erfüllt	(bedingt) erfüllt	erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	erfüllt	(bedingt) erfüllt
Copilot	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	(bedingt) erfüllt	erfüllt	erfüllt	(bedingt) erfüllt
LLaMa-2	erfüllt	nicht erfüllt	nicht erfüllt	(bedingt) erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	(bedingt) erfüllt	nicht erfüllt	nicht erfüllt	abgebrochen						

Template

Analyseumfang	LZCSVG			DUCSVG		UWCSVG			IVCSVG									
Testfallnummer	3	22	42	62	82	102	122	142	162	182	202	222	242	262	282	302	322	342
BloomZ	(bedingt) erfüllt	nicht erfüllt	nicht erfüllt	(bedingt) erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	abgebrochen, wurde am 14.12. ersetzt durch LLaMa-2						
ChatGPT	erfüllt	(bedingt) erfüllt	erfüllt	erfüllt	(bedingt) erfüllt	erfüllt	erfüllt	(bedingt) erfüllt	nicht erfüllt	erfüllt	erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	nicht erfüllt	erfüllt	(bedingt) erfüllt	(bedingt) erfüllt
Copilot	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt	erfüllt
LLaMa-2	nicht erfüllt	nicht erfüllt	nicht erfüllt	(bedingt) erfüllt	(bedingt) erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	nicht erfüllt	abgebrochen						

Abbildung 7: Individuelle Performance Fehlersuche

Anhang B – Individuelle Scores Fehlersuche

Nachfolgende Tabellen zeigen die einzelnen Scores, die jedes AI-Tool individuell bei jedem Test unter Nutzung der drei verschiedenen Prompting-varianten Zero-Shot, Few-Shot und Template erreicht hat.

Zero-Shot

Analyseumfang	LZCSVG			DUCSVG		UWCSVG			IVCSVG									
Testfallnummer	1	20	40	60	80	100	120	140	160	180	200	220	240	260	280	300	320	340
BloomZ	80%	0%	0%	10%	0%	0%	70%	0%	0%	0%	0%		abgebrochen, wurde am 14.12. ersetzt durch LLaMa-2					
ChatGPT	100%	70%	50%	100%	90%	100%	100%	80%	0%	100%	100%	50%	100%	90%	70%	90%	30%	70%
Copilot	100%	100%	50%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
LLaMa-2	0%	0%	0%	10%	30%	10%	100%	0%	0%	0%	0%				abgebrochen			

Few-Shot

Analyseumfang	LZCSVG			DUCSVG		UWCSVG			IVCSVG									
Testfallnummer	2	21	41	61	81	101	121	141	161	181	201	221	241	261	281	301	321	341
BloomZ	0%	0%	0%	0%	0%	0%	70%	0%	0%	0%	0%		abgebrochen, wurde am 14.12. ersetzt durch LLaMa-2					
ChatGPT	100%	80%	90%	100%	0%	100%	100%	20%	70%	70%	100%	50%	100%	60%	40%	40%	100%	70%
Copilot	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	60%	100%	100%	80%
LLaMa-2	100%	0%	0%	10%	0%	0%	0%	0%	20%	0%	0%				abgebrochen			

Template

Analyseumfang	LZCSVG			DUCSVG		UWCSVG			IVCSVG									
Testfallnummer	3	22	42	62	82	102	122	142	162	182	202	222	242	262	282	302	322	342
BloomZ	30%	0%	0%	10%	0%	0%	10%	0%	0%	0%	0%		abgebrochen, wurde am 14.12. ersetzt durch LLaMa-2					
ChatGPT	100%	80%	90%	100%	80%	100%	100%	60%	0%	90%	90%	50%	40%	60%	0%	80%	80%	70%
Copilot	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
LLaMa-2	0%	0%	0%	20%	50%	0%	0%	0%	0%	0%	0%				abgebrochen			

Abbildung 8: Individuelle Scores der Fehlersuche

Marburg Modul

AI-Lab: Wie können wir gemeinsam mit KI gut leben?

Fundamentale KI-Modelle in der Datenverarbeitung