

Experimente zur GPT-3.5-unterstützten Entwicklung von Prozessmodellen

Carlo Simon, Stefan Haag und Lara Zakfeld
Fachbereich Informatik
Hochschule Worms
Erenburgerstr. 19, 67549 Worms
Email: {simon,haag,zakfeld}@hs-worms.de

ZUSAMMENFASSUNG

Noch können Prozessmodelle nicht automatisch aus textuellen Beschreibungen generiert werden. Die Vorteile einer solchen Technologie lägen in kürzeren Entwicklungszeiten und einer kohärenteren, weil algorithmischen Interpretation des Inputs. Dazu müssten Methoden der formalen Modellierung und der Linguistik kombiniert werden. Die aktuellen GPT-Modelle eröffnen gerade hinsichtlich der linguistischen Analyse neue Möglichkeiten.

Dieser Beitrag untersucht, ob die Fähigkeiten von GPT-3.5, Texte zusammenzufassen, kausale Zusammenhänge zu erkennen oder Begriffe zu klassifizieren, auch ausreichen, um formale Prozessmodelle zu erstellen. Hierzu wurden systematisch Experimente zur Nutzung von GPT-3.5 durchgeführt. Dabei konnte beobachtet werden, dass die Resultate abhängig von den gewählten Transformer-Optionen erheblich variieren. Die Protokolle der Experimente werden in diesem Beitrag zusammengefasst und aus den Ergebnissen dann mögliche, künftige Forschungsaktivitäten abgeleitet.

SCHLÜSSELWÖRTER

GPT-3.5, Prozessmodellierung, Künstliche Intelligenz

DISCLAIMER

Dieser Beitrag ist eine übersetzte und überarbeitete Version des Papers *Experiments on GPT-3 Assisted Process Model Development* in den *Proceedings of the 37th International ECMS Conference on Modelling and Simulation* (Simon et al., 2023).

DIE NÄCHSTE DISRUPTION?

ChatGPT verwendet eine Chat-ähnliche Oberfläche, um mit Benutzern in natürlicher Sprache zu kommunizieren. Das von OpenAI entwickelte ChatGPT ist damit eine Umgebung, die maschinelles Lernen und künstliche Intelligenz einem breiten Publikum zugänglich macht. Viele Autoren schreiben ihr das Potenzial zu, bestehende Technologien zu verändern oder abzulösen. Eine von vielen Veröffentlichungen zum Thema ist <https://www.theguardian.com/commentisfree/2023/jan/07/chatgpt-bot-excel-ai-chatbot-tech>.

ChatGPT kann Fragen in natürlicher Sprache besser beantworten als bisherige Systeme. Hierzu wurde das zugrundeliegende Sprachmodell GPT-3.5, ein Large Language Model (LLM), mit über 235 Millionen Internet-Textdokumenten in Englisch und über 10

Millionen Textdokumenten in anderen Sprachen trainiert (Kublik und Saboo, 2022, S. 6). Zudem wurden Sourcecode und Dokumentationen von GitHub sowie Stack Overflow-Diskussionen genutzt. Hierdurch kann ChatGPT auch Fragen zur Softwareentwicklung und zur Erstellung formaler Spezifikationen beantworten.

Aus der prinzipiellen Fähigkeit, mit formalen Sprachen umzugehen, ergibt sich die Frage, welche Auswirkungen diese Technologie auf die Themenfelder Modellierung und Simulation haben könnte. Dieser Beitrag untersucht, ob und wie GPT-3.5 Menschen beim Entwickeln von simulierbaren Modellen assistieren kann.

In der vorliegenden Arbeit werden daher Prozesse, ihre natürlichsprachliche Beschreibung sowie deren formale Modellierung betrachtet. Das praktische Potential von GPT-3.5 wird anhand einer Reihe systematischer Experimente untersucht. Dabei haben sich die drei folgenden Herausforderungen und zugehörigen Lösungsansätze herauskristallisiert:

1. Trotz der riesigen Textmenge zum Trainieren von GPT-3.5 stehen nur vergleichsweise wenige natürlichsprachliche Prozessbeschreibungen zusammen mit ihrer formalen Spezifikation frei im Internet zur Verfügung. Die Autoren vermuten, dass insbesondere die grafische Repräsentation der Prozessmodelle den Zugriff erschwert. Sie haben daher ein Training formaler Methoden in den Dialog integriert.
2. Prozessmodelle von Organisationen basieren auf aktuellen Gegebenheiten. Das Trainingsmaterial für GPT-3.5 ist aber allgemeiner Text aus dem Internet, der bis 2021 erstellt wurde. Die Autoren haben daher versucht, aktuelle Fakten über in den Dialog integrierte Trainings in das System einzuspielen.
3. Menschen entwickeln Prozessbeschreibungen systematisch. Die Dialoge sollten daher so gestaltet werden, dass Nutzer ein „natürliches“ Modellierungserlebnis haben. Die Autoren haben hierfür die Fähigkeit von GPT genutzt, aufeinander folgende Anfragen als Dialog zu interpretieren, um das System Prozessmodelle in einer ähnlichen Reihenfolge erzeugen zu lassen, wie dies menschliche Modellierer üblicherweise tun. In diesen Dialogen werden die Modelle aufgebaut.

Die Idee des vorgestellten Ansatzes besteht darin, GPT-3.5 in einem fragengestützten Dialog schrittweise passende formale Spezifikationen aus einer gegebenen Prozessbeschreibung entwickeln zu lassen. Die Forschungsfrage lautet:

Welche Schritte im Prozess der (Prozess-)Modellentwicklung kann GPT-3.5 unterstützen?

Die vorliegende Arbeit unterliegt Einschränkungen: Da OpenAI seine Modelle permanent aktualisiert, ist es praktisch unmöglich, beim Wiederholen der Experimente auch die erzielten Ergebnisse zu reproduzieren. Dieser Beitrag ist wie folgt aufgebaut: Zunächst bettet der Abschnitt VERWANDTE ARBEITEN das Paper in einen größeren Rahmen ein und bietet grundlegende Informationen zu LLMs. Der Abschnitt EIN EXPERIMENTALLABOR beschreibt die verwendeten Systeme und gibt einen Überblick über die experimentellen Phasen. Die Experimente und ihre Ergebnisse werden in den drei folgenden Abschnitten beschrieben. Der Artikel endet mit ZUSAMMENFASSUNG UND AUSBLICK.

VERWANDTE ARBEITEN

Komplexe Tätigkeiten wie das Ableiten formaler Prozessspezifikationen aus narrativen Prozessbeschreibungen werden üblicherweise in Teilschritte gegliedert. Abbildung 1 zeigt, wie menschliche Modellierer typischerweise eine gegebene Prozessbeschreibung in ein Prozessmodell überführen. Zunächst werden die Prozesselemente identifiziert (Rollen und Akteure, Ereignisse und Auslöser, Aktivitäten, Geschäfts- und Informationsobjekte). Darauf aufbauend wird der Prozessfluss gegliedert (in Sequenzen, Alternativen, Iterationen und Nebenläufigkeiten). Schließlich wird das Modell um quantitative Informationen ergänzt. Hierfür muss ein Mensch den zu modellierenden Prozess verstehen.

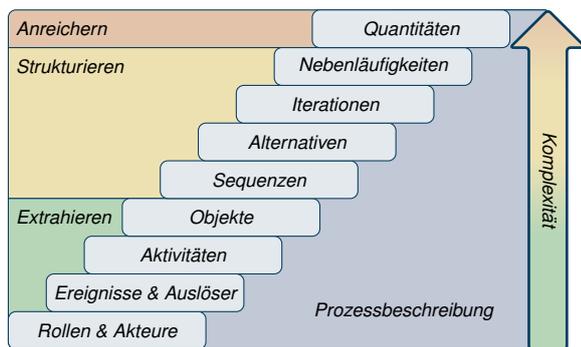


Abbildung 1: Ein Vorgehensmodell für die formale (Prozess-)Modellierung

GPT extrahiert hingegen aus einer Prozessbeschreibung zuvor gelernte Muster inklusive deren Transformation in Modelle, um diese zu übersetzen. Diese werden dann zu einem Ganzen verbunden. Benutzer können bei Bedarf zwischen alternativen Mustern wechseln, und so unterschiedliche Ergebnisse erzielen.

(Kecht et al., 2023) untersuchen anhand von Kundengesprächen auf Twitter die Fähigkeit von Chatbots, Geschäftsprozesse aus natürlicher Sprache zu extrahieren. Sie stoßen aber aufgrund der begrenzten Fähigkeiten von Chatbots, Antworten zu generieren, an die Grenzen solcher Systeme.

Da das Transformer-Modell von GPT-3.5 leistungsfähiger und weniger deterministisch als Chatbots ist, ist der in diesem Beitrag beschriebene Ansatz ein völlig anderer. Um dies zu verdeutlichen, werden zunächst die Begriffe Sprachbasierter Agent, Transformer-

Sprachmodell und insbesondere GPT-3.5 erläutert. Schließlich wird der hier vorgestellte Ansatz in einen größeren Forschungskontext eingebettet.

Zum Zeitpunkt der ursprünglichen Einreichung waren weder GPT-4 noch ein endgültiger Zeitplan für dessen Veröffentlichung verfügbar. In der Zwischenzeit durchgeführte initiale Tests mit ChatGPT Plus (das einen etwas eingeschränkten Zugang zu GPT-4 ermöglicht) zeigen jedoch, dass die hier gesammelten Erfahrungen ihre allgemeine Gültigkeit behalten.

Sprachbasierte Agenten

Sprachbasierte Agenten (Conversational Agents) simulieren menschenähnlich Unterhaltungen mit Benutzern. Diese Unterhaltungen können dabei als Dialoge ablaufen, die frühere Gesprächsverläufe mit berücksichtigen (Adamopoulou und Moussiades, 2022). Diese Systeme lassen sich irgendwo im folgenden Spektrum verorten:

- Regelbasierte Chatbots reagieren auf bestimmte Stichworte (vgl. Weizenbaum (1966)).
- Sprachassistenten nutzen maschinelles Lernen, um Eingaben in natürlicher Sprache zu „verstehen“. So vermeiden sie, nur auf bestimmte Stichworte anzusprechen. Sie reagieren somit menschenähnlicher, sind aber immer noch regelbasiert (vgl. Lopez et al. (2017)).

Transformer-Sprachmodelle

Transformer-Sprachmodelle verwenden Deep Learning, um Eingabesequenzen in Ausgabesequenzen umzuwandeln. Ein solches Modell wird probabilistisch trainiert, um Wörter und ihre Kontexte mit anderen Wörtern und deren Kontexten in Beziehung zu setzen. Das Modell kann dann entscheiden, welche dieser Kontexte mit höherer Wahrscheinlichkeit mit einem bestimmten Thema verbunden sind als andere. Bei einer Folge von Eingabewörtern berücksichtigt ein Transformer zudem den Kontext, um so das wahrscheinlichste Wort zu finden, das als nächstes folgt (Vaswani et al., 2017).

Jede Benutzereingabe und Systemausgabe wird in numerischer Form in einem Kontextvektor gespeichert. Dieser Vektor stellt das aktuelle Gedächtnis des Modells dar und hält es während eines Dialogs im Kontext. Durch Benutzereingaben kann er Informationen aufnehmen, die in den ursprünglichen Trainingsdaten nicht enthalten waren (Vaswani et al., 2017).

Die probabilistische Natur dieser Sprachmodelle macht sie zu natürlichen Feinden deterministischer Prozessbeschreibungen. Selbst vom Benutzer eingegebene Prozessfakten können von anderen zuvor gelernten Mustern überlagert werden. Das Problem, Faktenwissen verbindlich zu integrieren, ist noch ungelöst.

GPT-3.5

GPT-3.5 ist ein hochmoderner Transformer der Non-Profit-Organisation OpenAI. GPT-3.5 wurde mit einer großen Menge von Textdaten trainiert und ist in der Lage, natürlich anmutende, flüssige und kohärente Texte zu erzeugen. Aufgrund des großen Trainingskorpus kann ein spezifisches Feintuning durch „learning by doing“ zu überraschend guten Ergebnissen führen.

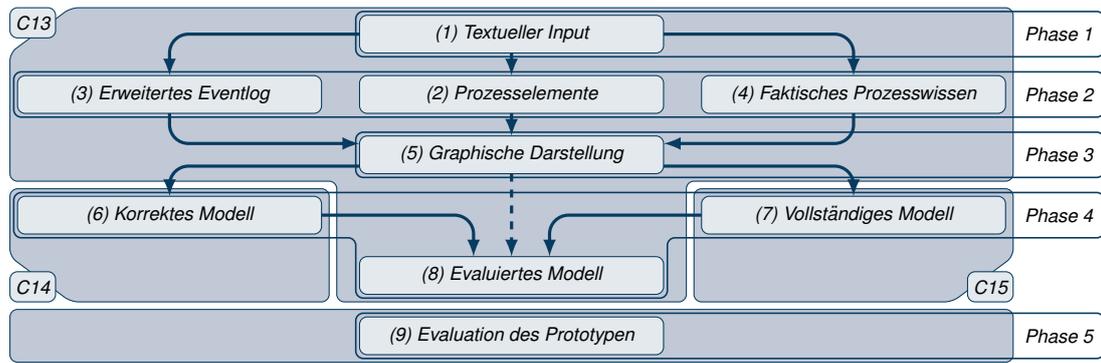


Abbildung 2: Phasen und Artefakte eines Forschungsplans zur Extraktion von Modellen aus natürlichsprachlichen Texten, siehe (Haag und Simon, 2022)

Eine der größten Stärken von GPT-3.5 ist seine Vielseitigkeit in einem breiten Spektrum von Anwendungen. Es kann für Aufgaben wie Übersetzung, Textvervollständigung, Schreiben von Zusammenfassungen, aber auch dem Erstellen von Chatbots und virtuellen Assistenten eingesetzt werden. GPT-3.5 zeichnet sich auch durch die Fähigkeit aus, kontextabhängige Texte zu generieren, die sich an die Präferenzen und den Sprachstil des Benutzers anpassen.

Transformation von Sprache in ein Modell

(Mendling et al., 2014) benennen 25 Herausforderungen der semantischen Prozessmodellierung, unter anderem als 13. Herausforderung diejenige, natürlichsprachliche Textbeschreibungen in Modelle zu transformieren. Abbildung 2 stellt diese Herausforderung in einen Kontext und nennt die verschiedenen Artefakte, die zur Lösung erstellt werden müssen. Zudem wird gezeigt, wie die Ergebnisse mit zwei der anderen Herausforderungen zusammenhängen, nämlich die Korrektheit und die Vollständigkeit von Modellen zu verifizieren (Haag und Simon, 2022).

Die meisten Beiträge, die sich mit solchen Transformationen beschäftigen, verwenden vorverarbeitete Daten aus Texten (z.B. Fliedl et al. (2005); Nolte (2020)). Aktuelle NLP-Technologien wie GPT-3.5 oder andere Transformer wurden bisher noch nicht untersucht.

Die Aufgabenstellung in ihrer allgemeinen Form ist aus zwei Gründen noch ungelöst:

1. Regelbasierte Automaten handhaben die Komplexität formaler Modellierung bislang unzuverlässig.
2. Methoden, die auf maschinellem Lernen basieren, benötigen in der Trainingsphase große Mengen an korrekt annotierten Beispielen. Diese sind jedoch schlicht nicht in ausreichender Menge verfügbar.

GPT-3.5 scheint diese Einschränkungen (zumindest in Ansätzen) zu überwinden.

EIN EXPERIMENTALLABOR

Die Experimente wurden mit den von OpenAI als „Labor“ zur Verfügung gestellten Werkzeugen durchgeführt. Jedes Experiment wurde in einem Laborjournal dokumentiert, das auf der Website der Arbeitsgruppe veröffentlicht ist. Es enthält alle Voreinstellungen, beschreibt die Experimente und umfasst die vollständigen Gesprächsverläufe (Haag, 2023).

GPT-3.5 und ChatGPT

Die umgangssprachlich als ChatGPT bezeichnete Web-Oberfläche (<https://chat.openai.com/>) ist genau genommen eine natürlichsprachliche Schnittstelle zu den Sprachmodellen GPT-3.5 und GPT-4. Auf deren Basis erstellt die Applikation kontextrelevante Antworten, die denen eines Menschen ähnlich sind (OpenAI, 2022). ChatGPT ist in der Lage Dialoge zu führen, um Antworten aufgrund bereits durchgeführter Interaktionen zu verbessern.

Ein an das System gesendeter Text, der als „Programmiersprache in normalem Englisch“ (oder einer anderen unterstützten Sprache) fungiert, sowie ein von ihm generierter Text wird als Prompt bezeichnet. Drei Arten von Input-Prompts werden unterschieden (Tingiris, 2021, S. 6-10):

Zero-shot prompts geben eine Aufgabenstellung oder einen Text vor, auf denen weitere Prompts aufgebaut werden sollen.

One-shot prompts sind Beispiele, anhand derer sich ableiten lässt, wie eine Aufgabe zu lösen ist.

Few-shot prompts liefern mit Hilfe mehrerer Beispiele Muster, denen gefolgt werden soll.

OpenAI Playground

Auch der OpenAI Playground (<https://platform.openai.com/playground>) ist eine webbasierte Schnittstelle, die benutzt werden kann, ohne eine Zeile Code zu schreiben (Tingiris, 2021, S. 20). Er bietet Zugang zu verschiedenen GPT-basierten Sprachmodellen, wobei *text-davinci-003* das zum Zeitpunkt der ursprünglichen Einreichung aktuelle Standardmodell war. Mehrere Parameter beeinflussen die Ausgabe des Systems (Brown et al., 2020), unter anderen die folgenden:

Temperatur steuert die Determiniertheit der Ausgabe. Je niedriger der Wert im Intervall $[0,1]$ liegt, desto deterministischer und repetitiver verhält sich das System. *Maximale Länge* begrenzt die Länge des Prompts, der in Token - einer numerischen Darstellung von Wortteilen - unterteilt wird. Für die meisten Sprachmodelle liegt die maximale Länge bei 2.048 Token, was etwa 1.500 Wörtern entspricht (Tingiris, 2021, S. 12). Bei Erreichen dieser Grenze stoppt die Ausgabe, d.h. es wird weder gekürzt noch umgeschrieben.

Token Highlighting färbt Token gemäß der Wahrscheinlichkeit, mit die sie generiert wurden.

Bei den Experimenten wurde die Werte für diese Parameter variiert:

- Teilweise wurde die Temperatur auf 0 statt der üblichen 0,7 gesetzt, um aufgrund der erzielten Determiniertheit Referenzprompts zu bilden.
- Die maximale Länge wurde erhöht. Dies beeinflusst die tatsächlich gegebene Antwort nicht, ermöglicht aber längere Gesprächsverläufe.
- Volles Token Highlighting wurde erlaubt, um Token mit vergleichsweise geringer Wahrscheinlichkeit zu erkennen.

Alle anderen Parameter wurden auf ihren Standard-einstellungen belassen.

Phasen der Experimente

Die Experimente wurden in drei Phasen durchgeführt:

1. *Erweitern einer Beschreibung*: GPT-3.5 wurde eingesetzt, um eine gegebene Prozessbeschreibung mit weiteren Details anzureichern.

2. *Informationsextraktion*: Anschließend wurde GPT-3.5 aufgefordert, prozessrelevante Informationen aus dieser erweiterten Beschreibung zu extrahieren.

3. *Formale Transformation*: Schließlich sollte GPT-3.5 diese Antworten in formalisierter Syntax darstellen.

Nach dieser kurzen Einführung in die technische Umgebung werden in den folgenden Abschnitten der experimentelle Aufbau und die beobachteten Ergebnisse beschrieben.

ERWEITERN EINER BESCHREIBUNG

In einem vorgelagerten Schritt wurde ein einfacher Prozess aus der Literatur übernommen (Simon et al., 2022a). Seine textuelle Beschreibung und eine Visualisierung sind in Abbildung 3 dargestellt.

The process is the handling of personalized wine gifts in a winery. Unlabeled bottles of wine are taken from the storage and delivered to the input inventory of a workbench where they receive a personalized label according to customer wishes.



In parallel, cartons are dispatched in form of sheets to another workplace to be assembled there. Both items are taken to the packing station where bottle and newly created supporting material like greeting cards are put into the carton. Afterwards, the packed gift box (for short: box) is taken to another workplace where the shipping label is attached, and the box is sealed. The completed box is deposited in the outgoing goods area.

Abbildung 3: Der Beispielprozess und seine Visualisierung

Aus diesem Text sollte GPT-3.5 Aktivitäten, Ereignisse und Prozessstrukturen extrahieren. Aufgrund der vielversprechenden Ergebnisse war das erste Ziel, die Komplexität mit Hilfe von GPT-3.5 zu erhöhen.

Konfiguration des Experiments

Dem OpenAI Playground wurde die ursprüngliche Prozessbeschreibung und die Liste der in Abbildung 4 dar-

gestellten Modifikationen vorgelegt. Verbunden damit war die Anweisung, hieraus eine erweiterte Prozessbeschreibung zu erstellen. Es wurden zehn Durchläufe mit einer Temperatur von 0 und weitere zehn Durchläufe mit einer Temperatur von 0,7 durchgeführt.

1. *The process starts with a customer order*
2. *From this order, amount and type of wine bottles are used to determine the bottles to be packed*
3. *The process not only handles gifts, but also standard orders. If an order is a gift, the personalized label for the bottle needs to be printed at an own workplace. If it is not a gift, this step isn't necessary as a stock label is used*
4. *The wine bottles are not necessarily packaged in size one, but there are different package sizes: 1, 3, 6, 12 bottles per package*
5. *Before sealing the box, a quality check is conducted for breakage, leakages, package contents, and overall looks*
6. *If during quality check an issue is discovered, the process should continue with the corresponding process step. The possible issues are: damaged bottle, damaged label on bottle, missing or damaged gift card, insufficient stuffing material, damaged cartons, missing or damaged shipping label*
7. *The process doesn't start immediately after receiving an order, but all orders are processed once a day*

Abbildung 4: Modifikationen zur Erweiterung der Beschreibung

Die so entwickelte Beschreibung ist in Abbildung 5 angegeben. Es handelt sich dabei um eine leicht bearbeitete, konsolidierte Version der Antworten von GPT-3.5. Die Gründe für die Bearbeitung werden nachfolgend erläutert. Trotz der Bearbeitung werden die beeindruckenden Potenziale großer Sprachmodelle deutlich.

Ergebnisse der ersten Phase

Ziel der ersten Phase war das Erstellen einer Prozessbeschreibung als Basis für die folgenden Phasen.

Die linguistischen Fähigkeiten ermöglichen es GPT-3.5, der Beschreibung Details hinzuzufügen und Prozesselemente zu modifizieren. Auf die Frage nach der Güte der generierten Prozessbeschreibungen bemängelte das Modell jeweils zu Recht das Fehlen integraler Konzepte des Geschäftsprozessmanagements wie etwa Leistungsmessung, Prozessziele oder die Anbindung an einen größeren organisatorischen Kontext.

The daily process of handling customer orders starts with the determination of the amount and type of wine bottles required for each order. Customers can order packages of one, three, six or twelve bottles each. The bottles are taken from storage to the input inventory at the workbench. If the order is a gift, the corresponding bottles receive a personalized label according to customer specifications that is printed at a small workstation. If it is a standard order, the bottles receive a stock label instead.

In parallel, cartons are dispatched to the assembly workplace, where they are assembled. At the packing station, the bottles and materials are put into the carton. This includes printing of these materials, such as accompanying documents and, for gift orders, greeting cards. Next, a quality check is conducted to ensure that there are no breakages, leakages, or discrepancies in the package contents and, if it is a gift package, to verify that the overall appearance is suitable. If any issues are found during the quality check, the process continues with the corresponding step to rectify the issue.

The shipping label is attached to the box at another workplace, where it is also sealed. Lastly, the package is deposited in the outgoing goods area.

Abbildung 5: (Teil-)automatisch erweiterte Beschreibung

Tabelle 1: Berechnete Wahrscheinlichkeiten für das Auftreten verschiedener Token in 10 OpenAI Playground Prompts bei Temperatur 0

Token	Lauf 1	Lauf 2	Lauf 3	Lauf 4	Lauf 5	Lauf 6	Lauf 7	Lauf 8	Lauf 9	Lauf 10
<i>corresponding</i>	27,41	28,46	28,18	27,07	27,83	28,15	28,06	27,89	27,96	28,73
<i>personalized</i>	23,31	22,18	22,67	22,68	22,80	22,43	22,69	22,96	23,19	21,90
<i>appropriate</i>	19,55	19,74	19,47	19,56	19,54	19,67	19,57	19,44	19,36	19,70
<i>label</i>	5,62	5,62	5,79	5,72	5,71	5,81	5,73	5,78	5,74	5,64
<i>necessary</i>	5,34	5,22	5,25	5,25	5,30	5,21	5,22	5,18		5,26
<i>respective</i>									5,20	
Summe	81,23	81,22	81,36	81,28	81,18	81,27	81,27	81,25	81,45	81,23

Alle zehn Durchläufe bei einer Temperatur von 0 führten zur genau gleichen Formulierung. Die zehn Durchläufe bei einer Temperatur von 0,7 hingegen lieferten jeweils unterschiedliche Ergebnisse hinsichtlich Formulierung und Detaillierungsgrad. Für eine verlässlichere Wiederholbarkeit sollte daher stets eine Temperatur von 0 verwendet werden. Eine Wiederholbarkeit ist aber selbst dann nicht garantiert, wie das Beispiel des Begriffs „corresponding“ im vierten Satz der in Abbildung 5 gezeigten Ausgabe zeigt:

Tabelle 1 listet die von GPT-3.5 berechneten Wahrscheinlichkeiten für das Auftreten dieses Tokens in zehn Durchläufen. Obwohl die Temperatur auf 0 gesetzt wurde, variieren die Werte zwischen den verschiedenen Läufen um etwa 4,6%. Die alternativen Terme besitzen immer niedrigere Wahrscheinlichkeiten und in allen Durchläufen sind die 3 folgenden Terme immer die gleichen. Der geringste Abstand zwischen den beiden obersten Wahrscheinlichkeiten beträgt jedoch vergleichsweise geringe 4,1%. Im neunten Durchlauf wechseln die fünftplatzierten Token, wenn dies auch keinen Einfluss auf die tatsächliche Ausgabe hat. Aufgrund dieser Beobachtungen kann aber nicht ausgeschlossen werden, dass ein solcher Wechsel auch auf der ersten Position vorkommen könnte. Die daraus folgende Änderung des Kontextvektors könnte Auswirkungen auf den weiteren Verlauf der Ausgabe haben.

Die Prozessbeschreibung in Abbildung 5 ist nicht die Kopie eines einzelnen Prompts, sondern eine leicht bearbeitete Kombination aus mehreren. Keine einzelne Version war aus menschlicher Sicht vollkommen überzeugend. So war beabsichtigt, die zuletzt hinzugefügte Information (Nr. 7: „The process doesn’t start immediately after receiving an order, but all orders are processed once a day“) als ersten Satz in die Beschreibung aufzunehmen. Stattdessen wurde diese Aussage nahezu wörtlich immer am Ende platziert. Mit Ausnahme des ersten stammen alle anderen Sätze der finalen Version aus den verschiedenen Prompts. Dies sollte die Auswirkungen des Eingriffs möglichst gering halten. Jedoch fehlen in der Beschreibung noch wichtige Informationen. So werden keine Akteure, Rollen oder Geschäfts- und Informationsobjekte definiert. Auch wird der Kontrollfluss ohne die explizite Nennung von Kontrollstrukturen beschrieben. Dies motivierte die Entscheidung, GPT-3.5 auf die Fähigkeit zu untersuchen, implizite Informationen aufzudecken und fehlende Informationen abzuleiten.

INFORMATIONSEXTRAKTION

In der zweiten Phase des Experiments wurde untersucht, inwiefern GPT-3.5 prozessrelevante Informationen aus einer Prozessbeschreibung extrahieren kann. Hierfür wurden 13 Fragen formuliert, die dem System gestellt wurden.

Konfiguration des Experiments

Dem System wurde der Prozess aus Abbildung 5 und die Aufforderungen aus Abbildung 6 bei einer maximalen Länge von 2.048 Token eingegeben. Es wurden vier verschiedene Settings betrachtet.

1. Provide an ordered list of the workplaces that are used in this process.
2. Provide an ordered list of business roles or actors in this process.
3. Provide an ordered list of events that occur in this process.
4. Provide an ordered list of activities that need to be conducted in this process.
5. Provide an ordered list of business objects in this process.
6. Provide an ordered list of information objects in this process.
7. Is there a sequence in this process? If so, which?
8. Is there a decision in this process? If so, which?
9. Is there an alternative in this process? If so, which?
10. Is there an exclusivity in this process? If so, which?
11. Is there a concurrency in this process? If so, which?
12. Is there an iteration in this process? If so, which?
13. Is there a loop in this process? If so, which?

Abbildung 6: Aufforderungen zur Analyse der Prozessbeschreibung

1. Zunächst wurden alle Aufrufe bei einer Temperatur von 0 durchgeführt. Der Prozess wurde zusammen mit jeder Frage in einem separaten Prompt an das System gesendet. Dadurch konnte das System keinen fragenübergreifenden Kontextvektor aufbauen.
2. Die Temperatur für den zweiten Versuch wurde ebenfalls auf 0 gesetzt. Allerdings wurden die Fragen im selben Prompt nach der jeweils vorhergehenden Antwort an das System gesendet, so dass das Modell dieses Mal einen Kontextvektor erstellen konnte.
3. Für den dritten Versuch wurde die Temperatur auf 0,7 gesetzt. Die Fragen wurden wie im zweiten Versuch an das System gesendet. Zur Prüfung der Wiederholbarkeit wurde dieser Versuch fünfmal wiederholt.

4. Schließlich wurde die Reihenfolge der Fragen randomisiert, die Temperatur jedoch auf 0,7 belassen. Diese Prozedur wurde zwanzigmal wiederholt.

Eine vollständige Untersuchung aller 13! möglichen Permutationen scheitert an der Anzahl der notwendigen Auswertungen.

Die Antworten von GPT-3.5 wurden, wie nachstehend erläutert, aus einer menschlichen Perspektive auf Vollständigkeit und Korrektheit hin bewertet.

Ergebnisse der zweiten Phase

Folgende weitere Beobachtungen wurden gemacht:

- Explizite Informationen über Arbeitsplätze, Tätigkeiten und einige Informationen über den Kontrollfluss wurden erfolgreich extrahiert.

- Implizite Informationen wurden geraten:

- Die Antworten zu Rollen und Akteuren reichten von insgesamt zwei sehr generischen Rollen (Kunden und Arbeiter) bis hin zu zehn Rollen (ein dedizierter Mitarbeiter je identifizierter Aktivität). Allerdings wurden in drei Durchläufen überhaupt keine Rollen identifiziert, sondern stattdessen Arbeitsplätze ausgegeben.

- Geschäfts- und Informationsobjekte wurden ebenfalls geraten und beinhalteten fälschlicherweise u. a. Materialien, Arbeitsplätze oder Kontrollstrukturen.

- Die Aktivitäten wurden nahezu korrekt erkannt.

- Die Extraktion von Ereignissen schlug faktisch fehl. Die Versuche lieferten eher Listen von Aktivitäten, die teilweise nicht einmal der Beschreibung entsprachen.

- Die Kontrollstrukturen zeigen ein unklares Bild:

- Sequenzen wurden in allen Läufen identifiziert. Sie sind jedoch in der Reihenfolge angeordnet, in der die Aktivitäten in der Beschreibung erscheinen, was nicht der beabsichtigten Reihenfolge innerhalb des Prozesses entspricht. Auch wurden einige Aktivitäten, die nebenläufig stattfinden sollten, in eine Sequenz gegliedert.

- Die Frage nach der Existenz von Nebenläufigkeiten wurde in 22 der 27 Läufe bejaht, auch wenn zuvor nur eine einzige, durchgehende Sequenz identifiziert wurde.

- In 23 Läufen wurde die Alternative korrekt erkannt.

- In 15 Läufen wurde die beabsichtigte Iteration nach der Qualitätskontrolle identifiziert. In 3 Läufen wurden Angaben zu Flaschenmengen, Kartonmontage oder Etikettendruck als Iterationen fehlinterpretiert. In 9 Durchläufen wurde keinerlei Iteration erkannt.

Aus Sicht der Autoren ist dies ein bemerkenswertes Ergebnis, wenn man bedenkt, dass es sich um erste Versuche mit dieser neuen Technologie handelt. Weitere Verbesserungen der Ergebnisse scheinen möglich, wenn der Kontextvektor systematisch durch passend formulierte Fragen in der richtigen Reihenfolge aufgebaut wird.

FORMALE TRANSFORMATION

Die bisher besprochenen Transformationen haben Reintext erzeugt. In diesem Beitrag geht es jedoch um die GPT-3.5-unterstützte Entwicklung von Prozessmodellen. Solche Modelle müssen in einer formalen Spezifikationsprache beschrieben werden, die speziellen syntaktischen Regeln folgt, wie z.B. Fluss- oder BPMN-Diagramme. Anspruchsvoller wäre eine Sprache mit

Ausführungssemantik wie Petri-Netze. Die extrahierten Ergebnisse müssen dann zusammengeführt werden, um eine Darstellung des Prozesses zu erhalten. Der Gesamtprozess besteht aus Aktivitäten, Ereignissen, Prozessstrukturen und anderen typischen Prozessmodellkomponenten. Diese Elemente müssen schließlich in einer Form abgebildet werden, die von einer Prozessmodellierungsumgebung verarbeitet werden kann.

Konfiguration des Experiments

Die Überführung natürlichsprachlicher Aussagen durch die Aufforderungen aus Abbildung 6 in eine formale Darstellung erfordert eine Modellierungsumgebung mit einer offenen API oder einem Klartext-Dateiformat.

Die Autoren haben sich für die Spezifikationsprache des Process-Simulation.Center (P-S.C) entschieden. Das P-S.C ist ein integriertes Managementsystem, das Modellierung, Simulation und Dokumentation von Prozessen mit Hilfe von Petri-Netzen ermöglicht. Das Tool bietet zudem Spezifikationsprachen für Organigramme und Swimlanes und unterstützt die Verknüpfung von Prozessen mittels Prozesslandkarten (Simon et al., 2022b).

Diese Versuche wurden in zwei Abschnitten durchgeführt. Der erste Abschnitt war wie folgt strukturiert:

1. Aufruf der Prozessbeschreibung bei Temperatur 0.
2. Abfrage der Arbeitsplätze, da aufgrund der vorherigen Versuche eine bessere Extraktion der anderen Prozesselemente erwartet wurde.

3. Extrahieren der Aktivitäten in einer Verb-Substantiv-Form und Formatieren der Ausgabe für das P-S.C.
4. Extrahieren der Ereignisse und formatieren. Da das Wort „event“ nicht verlässlich funktionierte, wurde stattdessen das Wort „trigger“ verwendet.

5. Extrahieren der Strukturen „sequence“, „branching“, „merging“, „iteration“ und „concurrency“ in der angegebenen Reihenfolge und formatieren der Ausgabe entsprechend der jeweiligen Struktur.

Dieses Experiment wurde einmal bei einer Temperatur von 0 und fünfmal in ChatGPT mit unterschiedlicher Reihenfolge der Aufforderungen bzgl. der Kontrollstruktur wiederholt.

Im zweiten Abschnitt wurden die an das System gestellten Aufforderungen umformuliert und mit Beispielen angereichert. Die ersten Fragen wurden in der selben Reihenfolge gestellt, jedoch wurden diejenigen nach den Kontrollstrukturen in „branching“, „merging“, „concurrency“, „iteration“ und „sequence“ geändert. Diese Anweisungen wurden einmal bei Temperatur 0 und fünfmal bei Temperatur 0,7 durchgeführt.

Ergebnisse der dritten Phase

Der erste Abschnitt lieferte passable Ergebnisse, wenn GPT-3.5 mit genauen Anweisungen und Beispielen für die korrekte Formatierung versorgt wurde. Im weiteren Verlauf der Konversationen entfernte sich das System jedoch immer mehr von den bereits (mutmaßlich) identifizierten Prozesselementen und -strukturen.

- Eine korrekte Spezifikation verlangt eine definierte Syntax und Schlüsselwörter. Daneben erhalten die einzelnen Elemente (sprechende) Namen. Bei einer Temperatur von 0,7 nummerierte ChatGPT die Elemente in mehreren Läufen jedoch durch, statt sich an den vorgegebenen Beispielen zu orientieren. Syntaktisch wichtige Zeichen wie Semikolons wurden nur in Ausnahmefällen gesetzt. Bei einer Temperatur von 0 arbeitete das System insgesamt besser, aber immer noch mit den grundlegend gleichen Fehlern.
- Das Extrahieren von Ereignissen (bzw. Triggern) funktionierte schlechter als das von Aktivitäten. Eine Umformulierung der Aufforderung und die Angabe von weiteren Beispielen verbesserten die Ergebnisse.
- Kontrollstrukturen wurden nicht erfolgreich erkannt. Die Resultate ließen sich mit einer veränderten Reihenfolge der Abfragen leicht verbessern. Trotzdem übersah das System weiterhin beabsichtigte Strukturen oder integrierte nicht vorhandene. Die Aktivitäten wurden so serialisiert, wie sie in der Beschreibung erscheinen, und explizite Korrekturen wurden größtenteils ignoriert. GPT-3.5 kann rudimentäre Prozessmodelle für Aktivitäten und Ereignisse erstellen. Allerdings sind die Resultate bzgl. der Kontrollstrukturen faktisch unbrauchbar. Unter diesem Gesichtspunkt können die Probleme von GPT-3.5, eine gegebene formale Syntax korrekt umzusetzen, vorerst zurückgestellt werden.

Zusammenfassung und Ausblick

Die in diesem Beitrag beschriebenen Experimente geben eine erste Antwort auf die Frage:

Welche Schritte im Prozess der (Prozess-)Modellentwicklung kann GPT-3.5 unterstützen?

Die Experimente wurden in die Phasen *Erweitern einer Beschreibung*, *Informationsextraktion* und *Formale Transformation* unterteilt, die unterschiedliche linguistische Fähigkeiten von GPT-3.5 adressieren: Generieren von Text, Zusammenfassen und Übersetzen (in eine formale Sprache).

Bedenkt man, wie neu diese Technologie ist, sind die erzielten Ergebnisse beeindruckend. Im Vergleich zu früheren Arbeiten an Chatsystemen im Bereich Prozessmodellierung stellt sie eine vielversprechende Entwicklung dar. Durch die riesige Menge an Trainingsdaten eröffnen Transformer neue Horizonte. Die derzeitigen Einschränkungen verhindern ein unbeaufsichtigtes Erstellen von Prozessmodellen. Eine assistierte Durchführung ist möglich, mehr aber noch nicht.

Die Gründe dafür sind in jeder Phase der Experimente zu beobachten:

Erweitern einer Beschreibung: Auch wenn die GPT-Sprachmodelle auf einem sehr großen Datensatz trainiert werden, verfügen sie nicht über Allgemeinwissen oder Menschenverstand. Sie sind nicht in der Lage, Wissen abzuleiten, sondern sind beschränkt auf das, was sie während ihres Trainings gelernt haben. Dies behindert eine sinnvolle Erweiterung bestehender Beschreibungen. Ein Feintuning kann schwierig sein, da spezifisches Domänenwissen nicht im gleichen Umfang zur Verfügung steht wie die allgemeinen Trainingsdaten.

Eine weitere Einschränkung ist von Bedeutung: Auch wenn es den Anschein hat, ist GPT-3.5 nicht in der Lage logisch zu denken und zu rechnen. Was wie logisches Denken aussieht, ist lediglich die Ausgabe der wahrscheinlichsten Wort- und Zahlenkombinationen unter Berücksichtigung des Kontextvektors.

Informationsextraktion: Die Beschreibungen müssen so präzise und klar wie möglich sein, um genaue Ergebnisse zu erhalten. Der Aufwand, der hierzu betrieben werden muss, könnte aber bereits in das Erstellen eines Modells investiert werden.

Eine Wissensextraktion ist möglich, wenn dem System die richtigen Fragen in der richtigen Formulierung und in der richtigen Reihenfolge gestellt werden. In den hier beschriebenen Experimenten konnten jedoch nicht immer alle relevanten Informationen erfasst werden.

Formale Transformation: GPT-3.5 kann auch bei der Erstellung formaler Modelle helfen. Prozesselemente und (zumindest teilweise) Strukturen lassen sich extrahieren und serialisieren. Das Resultat muss jedoch sorgfältig aufbereitet werden, um es auch für Modellierungssoftware verwenden zu können.

In Zukunft sind Arbeiten in verschiedenen Bereichen notwendig, um LLMs zur Geschäftsprozessmodellierung in der Praxis einsetzen zu können. Es existieren weitere Transformer-Sprachmodelle unterschiedlicher Implementierung ebenso wie andere, ältere Ansätze, die noch nicht detailliert untersucht wurden. Dies erfordert eine große Anzahl gut dokumentierter und evaluierter Experimente. Dazu zählt auch das Bereitstellen von geeigneten Trainingsdaten, in denen textuelle und formale Prozessbeschreibungen miteinander verknüpft sind.

Prompt Design und Engineering, das Erstellen „guter“ Anweisungen an die Modelle, könnte sich als weites, sich stetig veränderndes Forschungsfeld erweisen. Die systematische Entwicklung von Prompts wird großen Einfluss auf die Qualität der Prozessmodelle haben.

Die wichtigste Aufgabe besteht nach Ansicht der Autoren jedoch darin, die inhärent probabilistische Wissensbasis um Faktenwissen zu erweitern.

Auch die zwischenzeitliche Veröffentlichung von GPT-4 mitsamt der Integration in ChatGPT Plus ändert nur wenig an diesen Arbeitsfeldern.

Einerseits ist der Zugang zu GPT-4 kostenpflichtig und teilweise beschränkt und somit der allgemeinen Öffentlichkeit nicht ohne weiteres zugänglich. Aus diesen Gründen wurden durch die Autoren bisher nur vorläufige Tests mit ChatGPT Plus durchgeführt. Hierbei fällt negativ ins Gewicht, dass ChatGPT Plus keine Möglichkeiten zur Konfiguration des Systemverhaltens oder gar zur Anpassung des Modells besitzt.

Andererseits zeigt ChatGPT Plus im beschriebenen Anwendungsfall kein nennenswert verbessertes Verhalten. Dies ist unerwartet, da GPT-4 seinen Vorgängern gerade linguistisch deutlich überlegen ist. Jedoch konnten keine Änderungen bei der Extraktion von Ereignissen und Kontrollstrukturen oder der Formalisierung des Outputs festgestellt werden. Zu den Verbesserungen von GPT-4 gegenüber GPT-3 und den unterschiedlichen Fähigkeiten siehe (Bubeck et al., 2023).

DANKSAGUNG

Diese Forschung wird gefördert durch „ProFIL - Programm zur Förderung von Forschungspersonal, Infrastruktur und forschendem Lernen der Hochschule Worms“.

LITERATUR

- Adamopoulou, E. und L. Moussiades (2022): An Overview of Chatbot Technology. In: *Artificial Intelligence Applications and Innovations*, Hg. I. Maglogiannis; L. Iliadis und E. Pimenidis. Springer, Cham, S. 373–383.
- Brown, T.; B. Mann; N. Ryder; M. Subbiah; J. D. Kaplan; P. Dhariwal; A. Neelakantan; P. Shyam; G. Sastry; A. Askell; S. Agarwal; A. Herbert-Voss; G. Krueger; T. Henighan; R. Child; A. Ramesh; D. Ziegler; J. Wu; C. Winter; C. Hesse; M. Chen; E. Sigler; M. Litwin; S. Gray; B. Chess; J. Clark; C. Berner; S. McCandlish; A. Radford; I. Sutskever und D. Amodei (2020): Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems*, Hg. H. Larochelle; M. Ranzato; R. Hadsell; M.F. Balcan und H. Lin. Curran Associates, Inc., Bd. 33, S. 1877–1901.
- Bubeck, S.; V. Chandrasekaran; R. Eldan; J. Gehrke; E. Horvitz; E. Kamar; P. Lee; Y. T. Lee; Y. Li; S. Lundberg; H. Nori; H. Palangi; M. T. Ribeiro und Y. Zhang (2023): Sparks of Artificial General Intelligence: Early experiments with GPT-4. <https://arxiv.org/abs/2303.12712/> (last accessed: 10.04.2023).
- Fliedl, G.; C. Kop und H. C. Mayr (2005): From textual scenarios to a conceptual schema. *Data and Knowledge Engineering*, 55(1):20–37.
- Haag, S. (2023): *Protocol of Experiments on GPT-3 for Process Model Generation*. Techn. Ber., HS Worms.
- Haag, S. und C. Simon (2022): Extracting Models from Spoken Text - A Research Proposal. In: *EMISA 2022*. S. in Veröffentlichung.
- Kecht, C.; A. Egger; W. Kratsch und M. Röglinger (2023): Quantifying chatbots' ability to learn business processes. *Information Systems*, 113.
- Kublik, S. und S. Saboo (2022): *GPT-3 - Building Innovative NLP Products Using Large Language Models*. O'Reilly, Sebastopol, CA.
- Lopez, G.; L. Quesada und L. A. Guerrero (2017): Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In: *Advances in Human Factors and Systems Interaction*, Hg. I. L. Nunes. Springer, Cham, S. 241–250.
- Mendling, J.; H. Leopold und F. Pittke (2014): 25 Challenges of Semantic Process Modeling. *International Journal of Information Systems and Software Engineering for Big Companies*, 1(1):78–94.
- Nolte, F. R. (2020): *Text to Process Model: Automating Process Model Creation from Text*. Dissertation, WWU Münster.
- OpenAI (2022): ChatGPT: Optimizing Language Models for Dialogue. <https://openai.com/blog/chatgpt/> (last accessed: 10.04.2023).
- Simon, C.; S. Haag und L. Zakfeld (2022a): Stratification of Timed Petri Nets at the Example of a Production Process. In: *ECMS 2022: 36th International ECMS Conference on Modelling and Simulation*. S. 128–134.
- Simon, C.; S. Haag und L. Zakfeld (2022b): The Process-Simulation.Center. In: *SIM-SC: Special Track at SIMUL 2022: The Fourteenth International Conference on Advances in System Simulation*, Hg. F. Herrmann. Lissabon (Portugal), S. 74–77.
- Simon, C.; S. Haag und L. Zakfeld (2023): Experiments on GPT-3 Assisted Process Model Development. In: *ECMS 2023: 37th International ECMS Conference on Modelling and Simulation*. S. 270–276.
- Tingiris, S. (2021): *GPT-3 - An unofficial first look at the general-purpose language processing API from OpenAI*. Packt, Birmingham, UK.
- Vaswani, A.; N. Shazeer; N. Parmar; J. Uszkoreit; L. Jones; A. N. Gomez; L. Kaiser und I. Polosukhin (2017): Attention Is All You Need. In: *31st Conference on Neural Information Processing System*. Curran Associates.
- Weizenbaum, J. (1966): ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine. *Communications of the ACM*, 9(1):36–45.