Fakultät Wirtschaft
Studiengang Wirtschaftsinformatik

Seminararbeit "Integrationsseminar"

Low-Resource LLMs - Transformation BPMN in Text

Verfasser(in): **Pascal Nagel**
Matrikelnummer: **1128220**
Kurs: **WWI22B3**

Verfasser(in): **Benedikt Westphal**
Matrikelnummer: **6653031**
Kurs: **WWI22B3**

Seminargruppe: **LLM meets BPM_Process2Text**

Seminarbetreuer(in): **Tamino Fischer**

Abgabedatum: **13.01.2024**

Eigenständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Thema: „Low-Resource LLMs - Transformation BPMN in Text“ lediglich in Zusammenarbeit mit Benedikt Westphal verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Karlsruhe, 11.01.2025 

Ort, Datum Unterschrift

Inhalt

Abbildungsverzeichnis 4
1. Einleitung 5
1.1 Use-Case Beschreibung 5
1.2 Problemstellung 5
1.3 Zielsetzung 5
2. Theoretische Grundlagen 6
2.1 Begriffsklärung LLM 6
2.2 BPMN (Business Process Model and Notation) 6
2.3 Vorgehen der Nutzwertanalyse 7
3. Unterschiedliche LLMs 8
3.1 Überblick der LLMs 8
3.2 Mistral 7B 8
3.3 Gemma 8
3.4 LLaMA (Low-Rank Adaptation of Large Language Models) 8
4. Anforderungen 9
4.1 Beschreibung der Anforderungen 9
4.2 Gewichtung der Anforderungen 9
5. Nutzwertanalyse 10
5.1 Bewertung der Modelle 10
5.2 Schlussfolgerungen der Nutzwertanalyse 12
6. Handlungsempfehlung 13
7. Fazit 14
8. Literaturverzeichnis 15

Abbildungsverzeichnis

Abbildung 1: BPMN-Diagramm Beispiel https://www.inveskills.com/bpmn/bpmn-examples/ 7

Abbildung 2: Nutzwertanalyse..... 10

1. Einleitung

1.1 Use-Case Beschreibung

Die Business Process Model and Notation (BPMN) ist ein weit verbreiteter Standard zur grafischen Darstellung von Geschäftsprozessen. BPMN-Diagramme ermöglichen es Unternehmen, komplexe Abläufe übersichtlich zu visualisieren und somit die Kommunikation zwischen verschiedenen Abteilungen zu erleichtern. Dennoch können diese Diagramme für Personen, die nicht mit der Notation vertraut sind, schwer verständlich sein. Eine Transformation von BPMN-Diagrammen in natürlichsprachliche Texte kann dazu beitragen, Prozessinformationen für ein breiteres Publikum zugänglich zu machen und Missverständnisse zu reduzieren. Eine solche Umwandlung stellt WoPeD (Workflow Petri Net Designer) bereit, aktuell mit der Auswahl zwischen einem regelbasierten Ansatz und einer Schnittstelle zu Open AI GPT. Im Rahmen von WoPeD soll diese Arbeit geschrieben werden.

1.2 Problemstellung

Die manuelle Umwandlung von BPMN-Diagrammen in Text ist zeitaufwendig und fehleranfällig. Automatisierte Ansätze, die auf großen Sprachmodellen (LLMs) basieren, bieten hier Potenzial. Allerdings erfordern viele dieser Modelle erhebliche Rechenressourcen, die nicht in jedem Anwendungskontext zur Verfügung stehen. Insbesondere in ressourcenbeschränkten Umgebungen ist der Einsatz von High-End-LLMs oft nicht praktikabel. Probleme können zum einen die entstehenden Kosten für fortgeschrittene LLM-API-Endpunkte sein, weshalb die Anwendung mit dem eigenen Gerät von Vorteil sein kann. Zum anderen wird für Schnittstellen zu LLMs ein Internetzugriff benötigt. Daher besteht die Herausforderung darin, effiziente Low-Resource LLMs zu identifizieren, die trotz geringerer Anforderungen an Hardware und Rechenleistung qualitativ hochwertige Texte generieren können, sodass sie lokal laufen gelassen werden können.

1.3 Zielsetzung

Ziel dieser Arbeit ist es, verschiedene Low-Resource LLMs hinsichtlich ihrer Eignung für die Transformation von BPMN-Diagrammen in natürlichsprachliche Texte zu analysieren. Dabei sollen anhand von passenden Anforderungen die unterschiedlichen Modelle für den Use-Case “Process-to-Text” (P2T) verglichen werden. Durch eine Gegenüberstellung mittels Nutzwertanalyse sollen fundierte Empfehlungen für den praktischen Einsatz in unterschiedlichen Anwendungsszenarien abgeleitet werden.

2. Theoretische Grundlagen

2.1 Begriffsklärung LLM

Ein Large Language Model ist ein Computerprogramm, das mithilfe von künstlicher Intelligenz und maschinellem Lernen darauf trainiert wurde, natürliche Sprache zu verstehen und zu erzeugen. Diese Modelle analysieren große Mengen an Textdaten, um Muster und Zusammenhänge in der Sprache zu erkennen. Dadurch können sie Aufgaben wie Textgenerierung, Übersetzung oder Beantwortung von Fragen erfüllen. (Vaswani, Shazeer, & N., 2017)

LLMs nutzen oft die sogenannte Transformer-Architektur, die es ihnen ermöglicht, den Kontext von Wörtern in einem Satz oder Absatz zu erfassen (Vaswani, Shazeer, & N., 2017). Ermöglicht wird das durch umwandeln der textbasierten Daten in numerische Daten und je nach Wortposition in mehrdimensionale Vektoren, womit die LLMs deutlich besser arbeiten können. Dies führt zu einem tieferen Verständnis der Sprache und verbessert die Qualität der erzeugten Texte. (Klofat, 2023)

Low-Resource LLMs sind speziell optimierte Versionen großer Sprachmodelle, die mit weniger Rechenleistung und Speicher auskommen. Sie wurden entwickelt, um effizienter zu arbeiten und dennoch gute Ergebnisse zu liefern (Alam, Chowdhury, Boughorbel, & Hasanain, 2024). Dies macht sie besonders geeignet für den Einsatz auf Geräten mit begrenzten Ressourcen, wie Smartphones oder eingebetteten Systemen. Durch Techniken wie Modellkompression und vereinfachte Architekturen benötigen sie weniger Speicherplatz und können schneller ausgeführt werden. Ein entscheidender Vorteil ist, dass sie lokal, ohne Internetverbindung, genutzt werden können, was sowohl die Datensicherheit erhöht als auch den Einsatz in Umgebungen ohne stabile Online-Verbindung ermöglicht.

2.2 BPMN (Business Process Model and Notation)

Die Business Process Model and Notation (BPMN) ist eine standardisierte grafische Sprache zur Darstellung von Geschäftsprozessen (Business Process Model and Notation (BPMN) Version 2.0.2, 2014). Sie wurde entwickelt, um Abläufe innerhalb von Unternehmen einheitlich und verständlich abzubilden. BPMN ermöglicht es, Prozesse visuell zu modellieren, sodass sowohl technische Experten als auch Fachanwender diese verstehen können.

Ein BPMN-Diagramm besteht aus verschiedenen Symbolen und Elementen, die bestimmte Aspekte eines Prozesses repräsentieren. Ereignisse werden durch Kreissymbole dargestellt und kennzeichnen den Start, das Ende oder Zwischenfälle in einem Prozess. Aktivitäten erscheinen als abgerundete Rechtecke und stellen Aufgaben oder Arbeitsschritte dar, die ausgeführt werden müssen. Gateways, dargestellt durch rautenförmige Symbole, zeigen Entscheidungspunkte oder Verzweigungen im Prozessfluss an. Sequenzflüsse sind Pfeile, die die Reihenfolge der Aktivitäten und Ereignisse verbinden und den Ablauf des Prozesses zeigen.

Ein zentrales Element in BPMN sind die sogenannten Swimlanes oder Schwimmbahnen, die den Prozess in verschiedene Bereiche unterteilen. Diese Schwimmbahnen können in Pools und Lanes unterteilt werden. Pools repräsentieren eigenständige Organisationseinheiten oder Teilnehmer, beispielsweise verschiedene Unternehmen oder Abteilungen. Lanes sind Unterteilungen innerhalb eines Pools und stellen spezifische Rollen, Personen oder Unterabteilungen dar (Mendling, Reijers, & van der Aalst, 2010). Durch diese Strukturierung wird verdeutlicht, wer für welche Aktivitäten verantwortlich ist und wie die Interaktion zwischen verschiedenen Akteuren erfolgt.

BPMN erlaubt es, komplexe Prozesse in übersichtliche Diagramme zu überführen. Dies erleichtert das Verständnis der Abläufe und unterstützt bei der Identifikation von Verbesserungsmöglichkeiten (Mendling, Reijers, & van der Aalst, 2010). Zudem fördert es die Kommunikation zwischen verschiedenen Abteilungen und Stakeholdern, da alle Beteiligten eine gemeinsame visuelle Sprache nutzen. Die Anwendung von BPMN umfasst die Prozessdokumentation, also das Festhalten und Standardisieren von Abläufen, sowie die Prozessanalyse, bei der Engpässe oder ineffiziente Schritte identifiziert werden. Darüber hinaus dient BPMN der Prozessoptimierung, indem effizientere Abläufe durch Anpassung des Prozessdesigns entwickelt werden. Auch in der Schulung findet BPMN Anwendung, da neue Mitarbeiter durch visuelle Darstellungen Prozesse schneller erlernen können.

Durch die klare Struktur und Standardisierung von BPMN können Unternehmen ihre Prozesse besser steuern und anpassen, was zu einer höheren Effizienz und Wettbewerbsfähigkeit führt.

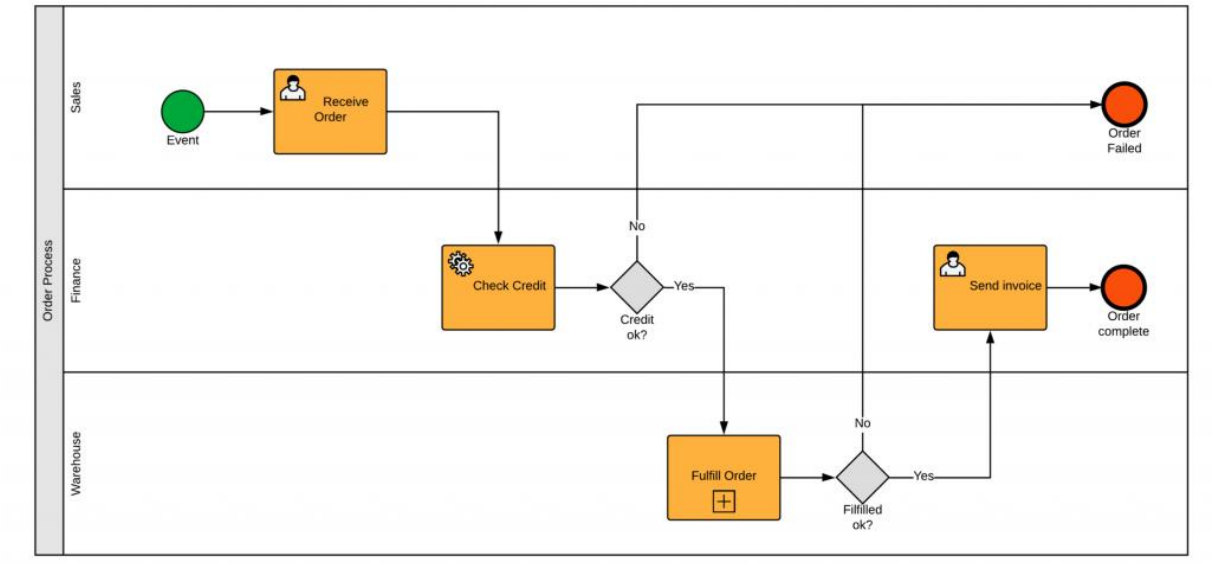

Abbildung 1: BPMN-Diagramm Beispiel https://www.inveskills.com/bpmn/bpmn-examples/

2.3 Vorgehen der Nutzwertanalyse

Die Nutzwertanalyse ist ein bewährtes Verfahren zur Entscheidungsfindung, das insbesondere bei komplexen und multidimensionalen Problemen eingesetzt wird. Sie ermöglicht es, verschiedene Alternativen anhand gewichteter Kriterien zu bewerten und eine fundierte Entscheidung zu treffen. Das Vorgehen bei der Durchführung einer Nutzwertanalyse folgt einem standardisierten Prozess, der in mehreren Schritten abläuft.

Zu Beginn der Nutzwertanalyse wird das Ziel präzise definiert. Es wird festgelegt, welche Entscheidung getroffen werden soll und welchen Beitrag die Nutzwertanalyse zur Lösung des Entscheidungsproblems leisten soll. Eine klare Zieldefinition ist essenziell, um die nachfolgenden Schritte zielgerichtet durchführen zu können. Daraufhin werden die zu bewertenden Alternativen identifiziert und beschrieben. Diese Alternativen müssen sich sinnvoll voneinander unterscheiden und vergleichbar sein. Nun werden die Kriterien festgelegt, anhand derer die Alternativen bewertet werden. Diese Kriterien müssen vollständig, relevant, überschneidungsfrei und bewertbar sein. Eine strukturierte Sammlung und Kategorisierung der Kriterien erleichtert die spätere Gewichtung und Bewertung. Die Kriterien werden gewichtet, um ihre relative Bedeutung für die Zielerreichung zu bestimmen. Die Summe der Gewichte muss 100 % betragen, um eine einheitliche Basis für die Bewertung zu schaffen. Für jedes Kriterium wird eine Skala definiert, die die Bewertung der Alternativen ermöglicht. Die Skalen müssen praktikabel, repräsentativ und einheitlich sein. Zudem werden Bewertungsvorschriften festgelegt, um die Konsistenz der Bewertungen sicherzustellen. Die Alternativen werden anhand der festgelegten Kriterien und Skalen bewertet. Die Bewertungen werden dokumentiert, um die Nachvollziehbarkeit zu gewährleisten. Als nächstes werden die Bewertungen der Alternativen mit den jeweiligen Kriteriengewichten multipliziert und zu einem Gesamtnutzwert summiert. Dieser Nutzwert dient als Grundlage für die Rangfolge der Alternativen. (Kühnapfel, 2021, S. 23-80)

3. Unterschiedliche LLMs

3.1 Überblick der LLMs

In diesem Kapitel werden verschiedene LLMs beschrieben, die für den geplanten Anwendungsfall in Betracht gezogen werden. Die Evaluierung dieser Modelle erfolgt mithilfe von LM Studio, einer Desktop-Anwendung, die es ermöglicht, lokale LLMs auf dem eigenen Computer auszuführen und zu bewerten. LM Studio unterstützt eine Vielzahl von Modellen, darunter LLaMA, Mistral und Gemma, und bietet eine benutzerfreundliche Oberfläche für den Download, die Installation und die Nutzung dieser Modelle. Es ist ein Open-Source Programm, wodurch keine Lizenzkosten anfallen. Zudem kann es auf allen gängigen Betriebssystemen laufen, wie Windows, Mac und Linux.

3.2 Mistral 7B

Mistral 7B ist ein Sprachmodell mit 7 Milliarden Parametern, das für hohe Leistung und Effizienz optimiert ist. Entgegen des Trents, immer größere Modellgrößen und Parameteranzahlen zu haben, um die Modelle besser zu machen, soll das Modell eine sehr gute Performance haben und trotzdem Effizient sein. Es soll größere Modelle wie LLaMA 2 (13B) und LLaMA 1 (34B) in Bereichen wie logisches Denken, Mathematik und Codegenerierung übertreffen. Es verwendet innovative Techniken wie Grouped-Query Attention (GQA) und Sliding Window Attention (SWA), um sowohl Speicher- als auch Rechenanforderungen zu reduzieren. Dadurch eignet es sich besonders für Anwendungen in Echtzeit. Das Modell wurde für jeden zugänglich veröffentlicht und kann leicht angepasst und in verschiedene Plattformen integriert werden. (Albert Q. Jiang, 2023)

3.3 Gemma

Gemma 2 9B ist ein kompaktes Sprachmodell mit 9 Milliarden Parametern, das von Google entwickelt wurde. Es gehört zur Gemma-Familie leichter, fortschrittlicher Open-Source-Modelle, die auf der Technologie der Gemini-Modelle basieren. Gemma 2 9B wurde mit Techniken wie Wissensdistillation trainiert, bei der es von einem größeren, vorab trainierten Modell lernt, um seine Effizienz zu maximieren. Trotz seiner kompakten Größe zeigt Gemma 2 9B eine bemerkenswerte Leistung in verschiedenen Aufgaben der natürlichen Sprachverarbeitung und übertrifft in einigen Benchmarks sogar größere Modelle. Seine geringe Größe ermöglicht den Einsatz in ressourcenbeschränkten Umgebungen, was es zu einer vielseitigen Lösung für verschiedene Anwendungen macht. (GemmaTeam, 2024)

Für die Umwandlung von BPMN-Diagrammen in natürlichsprachliche Texte könnte Gemma 2 9B aufgrund seiner kompakten Architektur und der damit verbundenen Ressourceneffizienz von Interesse sein. Seine Fähigkeit, präzise und verständliche Texte zu generieren, könnte die effektive Kommunikation von Prozessinformationen unterstützen. Zudem könnte die kompakte Architektur eine Integration in Systeme mit begrenzten Ressourcen erleichtern, was für die praktische Anwendung in verschiedenen Unternehmensumgebungen vorteilhaft wäre. (GemmaTeam, 2024)

3.4 LLaMA (Low-Rank Adaptation of Large Language Models)

LLaMA (Large Language Model Meta AI) ist eine Familie von großen Sprachmodellen, die von Meta AI entwickelt wurden. Ziel dieser Modelle ist es, leistungsstarke und effiziente Sprachmodelle bereitzustellen, die sowohl für Forschung als auch für praktische Anwendungen geeignet sind. Die LLaMA-Modelle zeichnen sich durch ihre Fähigkeit aus, mit weniger Ressourcen als andere große Sprachmodelle zu arbeiten, während sie dennoch eine hohe Leistung in verschiedenen NLP-Aufgaben bieten. In diesem Paper sollen die LLaMA-Modelle 3.2 1B sowie 3.1 8B untersucht und mit den anderen Modellen verglichen werden. Das Modell 3.2 1B ist ein neueres Modell, das mit einer Milliarde Parametern trainiert wurde, während das Modell 3.1 8B ein etwas älteres Modell ist, welches allerdings mit einer Parameteranzahl von acht Milliarden trainiert wurde. Durch diese Vergleiche soll zum einen untersucht werden, wie gut das LLaMA im Vergleich zu den anderen Modellen ist. Zum anderen kann dieser Vergleich darstellen, was für einen Unterschied eine andere Parameteranzahl ausmachen kann. (Hugo Touvron, 2023)

4. Anforderungen

4.1 Beschreibung der Anforderungen

Die Auswahl geeigneter Low-Resource-LLMs für die Transformation von BPMN-Diagrammen in natürlichsprachliche Texte erfordert eine klare Definition der Anforderungen. Diese Anforderungen wurden durch eine umfassende Literaturrecherche und die Analyse der spezifischen Bedürfnisse des Anwendungsfalls ermittelt. Im Folgenden werden die identifizierten Anforderungen detailliert beschrieben

Sprachliche Qualität und Stil:
Die generierten Texte sollten eine hohe sprachliche Qualität aufweisen, einschließlich korrekter Grammatik, klarer Syntax und eines flüssigen Stils. Ein professionelles Sprachniveau ist besonders wichtig, um die Texte für verschiedene Zielgruppen ansprechend und leicht verständlich zu gestalten. Der Fokus liegt hierbei auf der sprachlichen Ausführung und der Lesbarkeit des Textes.

Inhaltliche Präzision:
Die inhaltliche Präzision bezieht sich darauf, dass alle Prozessschritte, Entscheidungen und Abhängigkeiten aus den BPMN-Diagrammen korrekt und vollständig im Text wiedergegeben werden. Dies stellt sicher, dass die generierten Texte eine authentische und zuverlässige Abbildung der ursprünglichen Prozessdaten darstellen und keine wichtigen Informationen ausgelassen oder verfälscht werden.

Ressourceneffizienz:
In ressourcenbeschränkten Umgebungen ist die Effizienz des Modells hinsichtlich Speicher- und Rechenkapazität essenziell. Low-Resource-LLMs sollten mit begrenzten Hardware-Ressourcen effizient arbeiten können, um eine breite Einsatzfähigkeit zu gewährleisten. Dies schließt die Möglichkeit ein, die Modelle auch auf weniger leistungsstarker Hardware wie Laptops oder mobilen Geräten auszuführen.

Verarbeitungszeit:
Eine kurze Verarbeitungszeit ist entscheidend, um eine effiziente Arbeitsweise zu ermöglichen und die Produktivität zu steigern. Modelle mit schnellen Antwortzeiten sind besonders vorteilhaft in dynamischen Umgebungen, in denen Prozesse häufig aktualisiert werden müssen. Zudem trägt eine schnelle Verarbeitung dazu bei, dass die Modelle auch in zeitkritischen Anwendungsfällen eingesetzt werden können.

4.2 Gewichtung der Anforderungen

Die Gewichtung der Anforderungen orientiert sich an den zentralen Kriterien, die für die Transformation von BPMN-Diagrammen in natürlichsprachliche Texte relevant sind.

Sprachliche Qualität und Stil:
Dieser Aspekt wird mit 35 % am höchsten gewichtet, da eine klare und verständliche Sprache entscheidend für die Akzeptanz und das Verständnis der generierten Texte ist. Eine hohe sprachliche Qualität fördert die Benutzerfreundlichkeit und die effektive Kommunikation der Prozessinformationen.

Inhaltliche Präzision:
Mit einer Gewichtung von 30 % ist die korrekte und vollständige Wiedergabe der Prozessinformationen unerlässlich, um die Zuverlässigkeit der generierten Texte sicherzustellen. Dies minimiert das Risiko von Missverständnissen und gewährleistet, dass alle relevanten Details akkurat wiedergegeben werden.

Ressourceneffizienz:
Dieser Faktor wird mit 20 % bewertet, da die Fähigkeit, mit begrenzten Hardware-Ressourcen effizient zu arbeiten, die Praktikabilität und Einsatzfähigkeit der Modelle in verschiedenen Umgebungen erhöht. Effiziente Modelle ermöglichen den Einsatz auch auf Geräten mit geringerer Leistung.

Verarbeitungszeit:
Mit einer Gewichtung von 15 % ist eine kurze Verarbeitungszeit wichtig, um eine zügige Generierung der Texte zu ermöglichen. Schnelle Antwortzeiten sind besonders in dynamischen und zeitkritischen Anwendungen von Vorteil.

Diese Gewichtungen spiegeln die Prioritäten wider, die bei der Auswahl des geeignetsten Low-Resource-LLMs für die Transformation von BPMN-Diagrammen in natürlichsprachliche Texte berücksichtigt werden sollten.

5. Nutzwertanalyse

5.1 Bewertung der Modelle

In diesem Kapitel werden die in dieser Arbeit beschriebenen LLMs anhand der definierten Kriterien untersucht und systematisch verglichen. Für die Evaluation wurde LM Studio verwendet, um die Modelle unter identischen Bedingungen zu testen. Die Ergebnisse werden in einer Tabelle zusammengefasst, um die Modelle zu bewerten und zu gewichten. Diese strukturierte Darstellung erleichtert den Vergleich und die Identifizierung des am besten geeigneten LLMs für den spezifischen Anwendungsfall.

Für das Testen der verschiedenen Modelle wurden 7 verschiedene BPMN Diagramme in XML Form genutzt. Primär wurde ein Windows Computer mit 32 GB RAM ohne zusätzliche Grafikkarte mit 2133 MHz maximaler Speichergeschwindigkeit verwendet. Der Prozessor des Computers ist ein Intel Core i5-7260U Dual-Core-Prozessor. Zudem wurde als Referenz noch ein besserer Computer benutzt zur Überprüfung, wie viel Zeitunterschied in der Antwortgenerierung das macht. Hier wurden testweise einige Modelle in Kombination mit einigen Diagrammen stichprobenartig getestet. Dadurch sollen die Zeiten besser eingeordnet werden, um eine fundiertere Empfehlung geben zu können. Dies ist ein MacBook Pro mit 64 GB RAM und einer Speicherbandbreite von 400 GB/s. Der Computer hat einen M2 Max Chip mit 12 CPU-Kernen.

Kriterien	Mistral 7B	Gemma 9B	LLaMA 1B	LLaMA 8B	Gewichtung
Sprachliche Qualität und Verständlichkeit	6	8	7	6	2,5
Korrektheit der Inhalte	6	9	7	5	3,5
Ressourceneffizienz	6	4	9	5	2
Verarbeitungszeit	5	2	8	4	2
Gesamtpunkte	**58**	**63,5**	**76**	**50,5**	

Abbildung 2: Nutzwertanalyse

Für die Nutzwertanalyse wurden die vier zu untersuchenden Modelle auf die zuvor beschriebenen Anforderungen überprüft und untereinander verglichen. Entsprechend der jeweiligen Eigenschaften und getesteten Ergebnisse, wurde jeweils eine Punktevergabe zwischen 1 und 10 vorgenommen, die daraufhin mit der Gewichtung multipliziert wurde. Schlussendlich wurden diese Produkte der unterschiedlichen Anforderungen für jedes Modell addiert, sodass pro Low-Ressource LLM eine Gesamtpunktzahl entstanden ist. Das Modell mit der höchsten Gesamtpunktzahl ist am besten für den Use-Case geeignet und lieferte die besten Testergebnisse.

Sprachlichen Qualität und der Verständlichkeit:

Die Evaluierung der LLMs hinsichtlich der Sprachlichen Qualität und der Verständlichkeit erfolgte anhand der Analyse ihrer generierten Texte zu sieben BPMN-Diagrammen. Die Bewertung konzentrierte sich auf Klarheit, Lesbarkeit und narrative Struktur. Während einige Modelle die Diagramme präzise und flüssig beschrieben, produzierten andere lediglich unklare Nummernfolgen oder technische Listen, die den Prozesszusammenhang kaum erkennen ließen.

Gemma 2 9B erzielte die beste sprachliche Leistung. Die generierten Texte waren gut strukturiert, formal ansprechend und frei von grammatikalischen Fehlern. Prozesse und deren Abläufe wurden detailliert beschrieben, einschließlich der Entscheidungslogik und der Rollen der Beteiligten. Diese Qualität machte die Ausgaben leicht verständlich und professionell. LLaMA 3.2 1B lieferte ebenfalls solide Ergebnisse. Die Texte waren klar und sprachlich korrekt, wirkten jedoch gelegentlich stilistisch weniger ausgereift. In einigen Fällen waren die Beschreibungen stellenweise verkürzt, was die Lesbarkeit leicht beeinträchtigte.

Im Gegensatz dazu waren die Ergebnisse von LLaMA 3.1 8B und Mistral 7B deutlich schwächer. Besonders bei komplexeren Prozessen wurden von beiden Modellen stark vereinfachte und reduzierte Beschreibungen geliefert, die oft auf technische Stichpunkte beschränkt waren. Ein typisches Beispiel von LLaMA 3.1 8B lautet: "1. Start - Register (Activity_18bvx7s) - Flow to Gateway_0pmf7i8". Solche Ausgaben bieten kaum Kontext und

erschweren das Verständnis des Prozesses. Mistral 7B zeigte ähnliche Schwächen und produzierte fragmentarische Texte, die durch unklare Struktur und Technizität schwer zugänglich waren. Häufig fehlten klare Beschreibungen von Genehmigungsprozessen und weiteren relevanten Details.

Korrektheit der Inhalte:

Die Kategorie Korrektheit der Inhalte wurde anhand der Genauigkeit und Vollständigkeit der generierten Beschreibungen bewertet. Hierbei lag der Fokus darauf, ob die Modelle die BPMN-Prozesse umfassend und korrekt darstellten. Unterschiede in der Detailtiefe und Vollständigkeit waren zwischen den Modellen deutlich sichtbar.

Gemma 2 9B zeigte auch hier die besten Ergebnisse. Die Beschreibungen waren inhaltlich präzise und deckten die zentralen Prozessschritte sowie wichtige Details wie Entscheidungslogik und Teilnehmerrollen ab. Die Abhängigkeiten zwischen den Aufgaben wurden korrekt dargestellt, jedoch fehlten bei sehr komplexen Prozessen gelegentlich kleinere, aber relevante Details. LLaMA 3.2 1B zeigte ordentliche Ergebnisse, blieb jedoch hinter Gemma zurück. Die Texte waren zwar inhaltlich korrekt, aber nicht immer ausreichend detailliert. In einigen Fällen fehlten Verknüpfungen zwischen parallelen Prozessen, wodurch der Gesamtzusammenhang weniger klar wurde.

LLaMA 3.1 8B und Mistral 7B wiesen erhebliche Schwächen auf. LLaMA 3.1 8B vereinfachte komplexe Prozesse stark und lieferte oft nur oberflächliche Informationen. Der Umgang mit unvollständigen Formularen oder die Entscheidungsprozesse an Management-Gateways wurden häufig ausgelassen. Mistral 7B zeigte die schwächsten Ergebnisse und ließ zentrale Prozessinformationen aus. Grundlegende Schritte wurden zwar genannt, jedoch fehlten wichtige Details wie Zahlungsarten und Genehmigungsmechanismen vollständig.

Für die Bewertung der Ressourceneffizienz wurde berücksichtigt, was für Umgebungen die Modelle benötigen, damit sie laufen können. Dies soll auch auf die potenziellen Nutzer der Modelle in Betracht ziehen, was für Umgebungen diesen in der Regel zur Verfügung stehen. Die benötigten Ressourcen verändern sich nahezu parallel zu den enthaltenen Parametern eines Modells. Mit wie vielen Parametern ein Modell trainiert wurde, besagt die Zahl vor dem „B“ (für Billion auf Englisch oder Milliarde auf Deutsch), beispielsweise bei dem Mistral 7B wurden sieben Milliarde Parameter für das Training verwendet. Das LLaMA 1B Modell ist in diesem Vergleich folglich am ressourcenschonendsten, während das Gemma 9B mit den meisten Parametern trainiert wurde. Beispielsweise kann das Mistral 7B Modell noch auf einem Windows Surface mit acht GB RAM laufen gelassen werden, während das LLaMA 8B Modell bereits zu groß ist. Folglich wurden diese Punkte vergeben: Das LLaMA 1B Modell hat neun Punkte bekommen, da es auf nahezu jedem Computer oder Laptop und sogar auch auf einigen Handys ausgeführt werden kann. Das Modell Mistral 7B hat eine Punktzahl von sechs bekommen, da es schon deutlich mehr Ressourcen benötigt, allerdings trotzdem noch auf den meisten Geräten funktioniert. Etwas bessere Umgebungen benötigt das LLaMA 8B Modell, weshalb es mit fünf Punkten in der Kategorie Ressourceneffizienz abgeschlossen hat. Am wenigsten Punkte der Low-Ressource LLMs hat das Gemma 9B bekommen, da es mit neun Milliarde Parametern die größte Grundlage hat und somit die meisten Ressourcen benötigt. Daher wurde es hier eine Punktzahl von vier.

Die Bewertung der Verarbeitungszeit erfolgte auf Basis der durchführten Tests der Beispielmodelle durch die verschiedenen Modelle auf dem zuvor beschriebenen Windows Computer. Dafür wurden drei Werte verglichen: die Summe aller Zeiten, der kürzeste und der längste Wert für das Erstellen der sieben Diagrammbeschreibungen. Dadurch soll ausgeschlossen werden können, dass einzelne Ausreißer, die bei LLMs durchaus passieren können, die Ergebnisse verfälschen.

Das LLaMA 1B Modell hat daher eine Punktzahl von acht bekommen, da es im Vergleich zu den anderen getesteten Modellen mit Abstand am schnellsten war. Die kürzeste Zeit bei diesem Modell betrag ungefähr eine Minute, die längste Zeit 5,5 Minuten. In Summe waren alle Zeiten bei 1373 Sekunden, gute 20 Minuten. Obwohl es das deutlich schnellste Modell war, hat es trotzdem nur eine Punktzahl von acht bekommen, da im Vergleich zum beispielsweise beim regelbasierten Generieren der Texte aus den BPMN-Diagrammen die Zeiten trotzdem länger sind. Das Mistral 7B Modell hat fünf Punkte für die Verarbeitungszeit bekommen, da hier die Zeiten deutlich länger sind. Die minimale Zeit war hier etwas mehr als vier Minuten, während die längste Zeit bei knapp 30 Minuten lag. In Summe hat dieses Modell für alle sieben Diagramminterpretationen knapp 7500 Sekunden, also 125 Minuten benötigt. Im Vergleich dazu hat das LLaMA 8B Modell für das kleinste ca. drei Minuten und für das größte Diagramm ca. 28 Minuten benötigt. In Summe hat es für alle sieben Diagramme ca. 7800 Sekunden, also 130 Minuten, weshalb es eine Punktzahl von vier bekommen hat. Obwohl hier die minimale und maximale Antwortzeit jeweils kürzer war als bei dem Mistral 7B Modell, hat dieses Modell für die Anforderung der Zeit einen Punkt weniger bekommen, da die Summe aller Zeiten im Vergleich länger war. Das langsamste Modell in nahezu allen Kategorien und Diagrammen war das Gemma 9B, sodass es eine Punktzahl von zwei bekommen hat. Die kürzeste Verarbeitungszeit betrug hier ähnlich wie beim Mistral 7B ca. vier Minuten, wobei die längste Zeit

gute 37 Minuten waren. Auch in Summe aller Diagramme hat es am längsten benötigt mit einer Zeit von 9668 Sekunden, also guten 160 Minuten.

Im Vergleich dazu waren die gemessenen Zeiten auf dem beschriebenen MacBook deutlich kürzer. Hier konnte die textuelle Beschreibung des BPMN-Diagramms der Modelle um ein Vielfaches schneller erstellt werden. Beispielsweise betrug die Zeit zum Erstellen des Diagramms „capacity_planning“ mit dem Gemma 9B Modell auf dem Mac 33 Sekunden, während mit dem Windows Computer 1824 Sekunden gewartet werden müsste, bis der Text erschien. Hier ist der bessere Computer also mehr als 50-mal so schnell. Auch bei kleineren Modellen oder kleineren Diagrammen ist der Mac deutlich schneller. Zum Beispiel benötigt er für das Diagramm „mostSimple“, ein sehr kleines und einfaches Diagramm hat er unter einer Sekunde benötigt, während der andere Computer 224 Sekunden benötigt hat. Als Beispiel, dass der Mac auch bei kleineren Modellen deutlich schneller ist, kann das Modell LLaMA 1B erneut mit dem Diagramm „capacity_planning“ genommen werden. Der Mac hat für das Diagramm mit dem genannten Modell 12,5 Sekunden benötigt, während der Windows Computer 337 Sekunden benötigt hat.

Es kann gesagt werden, dass die benutze Hardware für das Analysieren der Diagramme einen starken Einfluss auf die Zeit zur Antwortgenerierung hat. Mit den gezeigten Computern war der eine um ein Vielfaches, bis zu ungefähr 200 Mal, schneller.

5.2 Schlussfolgerungen der Nutzwertanalyse

Die durchgeführte Nutzwertanalyse hat gezeigt, dass die Wahl des geeigneten Low-Resource-LLMs stark von den spezifischen Anforderungen des Anwendungsfalls abhängt. Die Ergebnisse der Analyse verdeutlichen, dass jedes Modell seine Stärken und Schwächen hat, jedoch ein Modell besonders hervorsticht. Das Modell LLaMA 3.2 1B hat sich als die beste Lösung zur Integration in WoPeD erwiesen. Es kombiniert eine solide sprachliche Qualität und inhaltliche Präzision mit einer außergewöhnlichen Ressourceneffizienz und kurzen Verarbeitungszeiten. Diese Eigenschaften machen es besonders geeignet für ressourcenbeschränkte Umgebungen, in denen Hardwarekapazitäten begrenzt sind und dennoch qualitativ hochwertige Ergebnisse erwartet werden. Im Vergleich zu den anderen Modellen bietet LLaMA 3.2 1B folgende Vorteile:

Ressourceneffizienz: Mit nur einer Milliarde Parametern ist es das ressourcenschonendste Modell im Test. Es kann problemlos auf Geräten mit begrenzter Hardware wie Laptops oder sogar Smartphones ausgeführt werden, was es für eine breite Nutzerbasis zugänglich macht.

Verarbeitungszeit: LLaMA 3.2 1B war das schnellste Modell im Test und konnte die BPMN-Diagramme in deutlich kürzerer Zeit analysieren und beschreiben als die anderen Modelle. Dies ist besonders in dynamischen und zeitkritischen Umgebungen von Vorteil.

Sprachliche Qualität: Obwohl es nicht die höchste Punktzahl in der sprachlichen Qualität erreichte, lieferte es dennoch klare und verständliche Texte, die für den Anwendungsfall ausreichend präzise und lesbar waren.

Die anderen Modelle zeigten in bestimmten Bereichen ebenfalls Stärken, konnten jedoch nicht die gleiche Ausgewogenheit wie LLaMA 3.2 1B bieten:
Gemma 9B erzielte die besten Ergebnisse in der sprachlichen Qualität und inhaltlichen Präzision, war jedoch aufgrund seiner hohen Hardwareanforderungen und langen Verarbeitungszeiten weniger praktikabel für ressourcenbeschränkte Umgebungen.
Mistral 7B und LLaMA 3.1 8B zeigten Schwächen in der sprachlichen Qualität und inhaltlichen Präzision, was sie für den Anwendungsfall weniger geeignet macht.

Zusammenfassend lässt sich sagen, dass LLaMA 3.2 1B die beste Lösung für den Use-Case "Process-to-Text" darstellt. Es bietet eine optimale Balance zwischen Effizienz und Qualität und erfüllt die Anforderungen des Anwendungsfalls am besten. Für Szenarien, in denen die sprachliche Qualität oberste Priorität hat und leistungsstarke Hardware verfügbar ist, könnte jedoch auch Gemma 9B eine geeignete Alternative sein.

6. Handlungsempfehlung

Basierend auf den Ergebnissen der Nutzwertanalyse wird folgende Handlungsempfehlung ausgesprochen:

Das Modell LLaMA 1B sollte als Standardlösung direkt in WoPeD integriert werden. Aufgrund seiner hohen Ressourceneffizienz und im Vergleich kurzen Verarbeitungszeiten eignet sich dieses Modell besonders gut für ressourcenbeschränkte Umgebungen und lokale Anwendungen. Mit LLaMA 1B kann der Benutzer BPMN-Diagramme ohne Internetverbindung in natürlichsprachliche Texte umwandeln. Dies bietet eine kostengünstige und datenschutzfreundliche Lösung, da keine sensiblen Daten an externe Server gesendet werden. LLaMA 1B liefert zudem eine ausreichende sprachliche Qualität und Präzision für die meisten Anwendungsfälle.

Für Szenarien, in denen eine höhere sprachliche Qualität und Präzision erforderlich sind, sollte zusätzlich die Möglichkeit bestehen, über eine API Anfragen an leistungsstarke Modelle wie OpenAI GPT zu senden. Diese Option ist bereits in WoPeD integriert und sollte daher als Option weiterhin vorhanden sein. Sie hat die Vorteile, dass die Qualität der Ergebnisse in der Regel höher ist, ohne dass lokal gute Hardware benötigt wird. Zudem gibt es mehr Skalierbarkeit und Flexibilität.

Auch die regelbasierten Übersetzungen der Modelle in Text sollten weiterhin in WoPeD zur Verfügung stehen, da diese einen deterministischen Output liefern und vollständig unabhängig von Rechenleistung sind.

Um die Benutzerfreundlichkeit und Flexibilität von WoPeD zu maximieren, sollte eine Auswahlfunktion integriert werden, die es dem Benutzer ermöglicht, zwischen verschiedenen Methoden zur Transformation von BPMN-Diagrammen in Text zu wählen. Dabei könnte LLaMA 1B als Standardlösung für ressourcenbeschränkte und datenschutzkritische Szenarien dienen, da es lokal ausgeführt werden kann und keine Internetverbindung benötigt. Für anspruchsvollere Anwendungen, bei denen eine höhere sprachliche Qualität erforderlich ist, sollte die Möglichkeit bestehen, über eine API externe Modelle wie OpenAI GPT zu nutzen. Zusätzlich sollte der regelbasierte Ansatz als deterministische Alternative verfügbar bleiben. Für Nutzer mit leistungsstarker Hardware könnte auch die Auswahl zwischen weiteren lokalen Modellen wie Gemma 9B oder Mistral 7B angeboten werden, um maximale Flexibilität zu gewährleisten.

7. Fazit

Die Ergebnisse dieser Arbeit zeigen, dass Low-Resource-LLMs eine praktikable und effiziente Lösung für die Transformation von BPMN-Diagrammen in natürlichsprachliche Texte darstellen. Insbesondere das Modell LLaMA 1B hat sich als die beste Wahl für den spezifischen Use-Case erwiesen. Es bietet eine optimale Balance zwischen Ressourceneffizienz, Verarbeitungszeit und einer ausreichenden sprachlichen Qualität, wodurch es sich ideal für ressourcenbeschränkte Umgebungen eignet. Die Möglichkeit, dieses Modell lokal auszuführen, gewährleistet zudem eine hohe Datensicherheit und Unabhängigkeit von externen Diensten. Für anspruchsvollere Szenarien, bei denen eine höhere sprachliche Qualität und Präzision erforderlich sind, bleibt jedoch die Option eines API-Aufrufs an leistungsstarke Modelle wie OpenAI GPT eine sinnvolle Ergänzung. Diese Kombination ermöglicht es, sowohl einfache als auch komplexe Anforderungen abzudecken. Der regelbasierte Ansatz bietet darüber hinaus eine deterministische Alternative, die unabhängig von KI-Modellen funktioniert und in standardisierten Szenarien weiterhin von Nutzen sein kann. Ein vielversprechender Ausblick besteht in der Nutzung von Retrieval-Augmented Generation (RAG), um die Modelle gezielt auf den spezifischen Anwendungsfall zu trainieren. Durch die Integration von domänenspezifischen Daten, wie BPMN-Dokumentationen oder branchenspezifischen Prozessbeschreibungen, könnten die Modelle weiter optimiert werden, um noch präzisere und kontextbezogenere Ergebnisse zu liefern. Dies würde nicht nur die Qualität der generierten Texte verbessern, sondern auch die Anpassungsfähigkeit der Modelle an individuelle Anforderungen erhöhen. Zusammenfassend lässt sich sagen, dass eine flexible Architektur, die dem Benutzer die Wahl zwischen verschiedenen Ansätzen und Modellen ermöglicht, die beste Lösung darstellt. Die Kombination aus lokal ausführbaren Modellen wie LLaMA 1B, API-Aufrufen für komplexe Anforderungen und der Möglichkeit, Modelle durch RAG auf spezifische Use-Cases zu trainieren, bietet eine umfassende und zukunftssichere Grundlage für die Transformation von BPMN-Diagrammen in Text.

8. Literaturverzeichnis

Alam, F., Chowdhury, S. A., Boughorbel, S., & Hasanain, M. (2024). *LLMs for Low Resource Languages.* Von https://aclanthology.org/2024.eacl-tutorials.5.pdf abgerufen

Albert Q. Jiang, A. S. (10. 10 2023). *Mistral 7B.* Von https://arxiv.org/pdf/2310.06825 abgerufen

(2014). *Business Process Model and Notation (BPMN) Version 2.0.2.* OMG (Object Management Group).

GemmaTeam. (16. 04 2024). *Gemma: Open Models Based on Gemini.* Von https://arxiv.org/pdf/2403.08295 abgerufen

Hugo Touvron, T. L. (27. 02 2023). *LLaMA: Open and Efficient Foundation Language Models.* Von https://arxiv.org/pdf/2302.13971 abgerufen

Klofat, D. A. (16. 05 2023). *Wie funktionieren Transformer? Definition und Praxis*. Von https://www.informatik-aktuell.de/betrieb/kuenstliche-intelligenz/wie-funktionieren-transformer-definition-und-praxis.html abgerufen

Kühnapfel, J. B. (2021). *Scoring und Nutzwertanalysen.* Ludwigshafen am Rhein: Springer.

Mendling, J., Reijers, H., & van der Aalst, W. (2010). *Seven Process Modeling Guidelines (7PMG).*

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge University Press.

Vaswani, A., Shazeer, N., & N., P. (2017). *Attention is All You Need. In Advances in Neural Information Processing Systems.*