HAW HAMBURG

BACHELORARBEIT

LLM-basierte Vereinfachung von Verwaltungstexten am Department Informatik

vorgelegt am 18. März 2025

Narges Shafieyoun

Erstprüferin: Prof. Dr. Marina Tropmann-Frick
Zweitprüfer: Prof. Dr. Stefan Sarstedt

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Department Informatik
Haus B, Berliner Tor 7
20099 Hamburg

Abstract

Der Zugang zu öffentlichen Dienstleistungen und Informationen in einer verständlichen und barrierefreien Form ist eine zentrale Herausforderung moderner Gesellschaften. Besonders betroffen sind Menschen mit begrenzten Sprachkenntnissen oder kognitiven Beeinträchtigungen wie Demenz, für die komplexe amtliche Texte und Formulare erhebliche Hürden darstellen. Diese Barrieren erschweren nicht nur den Zugang zu wichtigen Informationen, sondern auch die gesellschaftliche Teilhabe.

Diese Arbeit untersucht, wie Künstliche Intelligenz (KI), insbesondere Large Language Models (LLMs), dazu beitragen kann, englische Verwaltungstexte verständlicher zu machen, ohne ihre inhaltliche Genauigkeit zu verfälschen. Im theoretischen Teil wird die Problematik schwer verständlicher Verwaltungssprache erläutert und der aktuelle Stand der Forschung im Bereich KI-gestützter Sprachvereinfachung analysiert. Dabei liegt der Fokus auf modernen Natural Language Processing (NLP)-Modellen, insbesondere auf Transfer-Learning-Ansätzen, die komplexe Sprachstrukturen gezielt transformieren können. Ein zentraler Beitrag dieser Arbeit ist die Erstellung eines spezifischen Datensatzes auf Basis amtlicher Formulare, der für das Training und die Evaluation der Modelle genutzt wird.

Im experimentellen Teil werden drei führende Sprachmodelle – LLaMA 3 8B, Phi-3 mini und Mistral 7B– feinabgestimmt und evaluiert. Ziel ist es, ihre Fähigkeit zur Vereinfachung englischer Verwaltungssprache zu bewerten, indem sie Texte leichter verständlich machen, während ihre inhaltliche Präzision erhalten bleibt. Die Ergebnisse zeigen, dass diese Modelle die Lesbarkeit und Verständlichkeit amtlicher Texte erheblich verbessern können, ohne Informationen zu verfälschen.

Abschließend werden die Herausforderungen und Grenzen der eingesetzten Methoden diskutiert, insbesondere im Hinblick auf die Anpassung an sprachliche und kulturelle Kontexte. Die Ergebnisse verdeutlichen das Potenzial von KI zur Förderung von Barrierefreiheit und Inklusion, indem sie zeigen, wie Technologie dazu beitragen kann, Verwaltungskommunikation nicht nur effizienter, sondern auch menschlicher zu gestalten. Abschließend werden Ansätze für zukünftige Forschungen aufgezeigt, um diese Technologien weiterzuentwickeln und ihre praktische Anwendbarkeit zu optimieren.

Inhaltsverzeichnis

I. Abkürzungsverzeichnis

APE	Automatic Prompt Engineering
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CoT	Chain-of-Thought
CRF	Conditional Random Fields
GPT	Generative Pre-trained Transformer
GQA	Grouped-Query Attention
HF	Hugging Face
HMM	Hidden Markov Models
ICL	In-Context Learning
KI	Künstliche Intelligenz
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
ML	Machine Learning
MMLU	Massive Multitask Language Understanding
NER	Named Entity Recognition
NLP	Natural Language Processing
PEFT	Parameter-Efficient Fine-Tuning
SARI	System Output Against References and Input
SLM	Smal Language Model
SWA	Sliding Window Attention

II. Abbildungsverzeichnis

III. Tabellenverzeichnis

1 Einleitung und historischer Hintergrund

In einer zunehmend informationsgetriebenen Gesellschaft ist die Verständlichkeit von Texten für alle Bevölkerungsgruppen von zentraler Bedeutung. Besonders im behördlichen und administrativen Bereich kann eine komplexe Sprache erhebliche Barrieren für Bürger*innen darstellen. Die automatische Textvereinfachung (Text Simplification) hat sich daher als ein bedeutendes Forschungsgebiet im Bereich des Natural Language Processing (NLP) etabliert. Ihr Ziel besteht darin, die sprachliche Komplexität von Texten zu reduzieren, ohne deren ursprüngliche Bedeutung oder den Informationsgehalt zu verfälschen [1] [2] [3]. Besonders profitieren davon Menschen mit Leseschwierigkeiten, Fremdsprachenlernende, Personen mit kognitiven Einschränkungen sowie all jene, die komplexe Fachinformationen in verständlicher Form benötigen.

1.1 Historische Entwicklung der Textvereinfachung

Die Idee der Textvereinfachung ist keineswegs neu. Erste Konzepte lassen sich bis in die 1930er Jahre zurückverfolgen, als Charles Kay Ogden mit *Basic English* eine reduzierte Form des Englischen entwickelte, die durch einen begrenzten Wortschatz von 850 Wörtern die Verständlichkeit erleichtern sollt [4]. Ein weiterer bedeutender Meilenstein war die Einführung von Special English durch Voice of America im Jahr 1959 [5]. Dieses sprachlich vereinfachte Nachrichtensystem zielte darauf ab, internationale Zuhörer*innen durch den Einsatz eines begrenzten Wortschatzes und vereinfachter Grammatik besser zu erreichen.

Ein Durchbruch in der behördlichen Anwendung erfolgte 2010 in den USA mit dem Plain Writing Act, der öffentliche Institutionen dazu verpflichtete, eine klare und verständliche Sprache zu verwenden [6]. Auch im deutschsprachigen Raum gewann die Idee der Leichten und Einfachen Sprache zunehmend an Bedeutung. Ein entscheidender rechtlicher Rahmen wurde mit dem Behindertengleichstellungsgesetz [7] geschaffen. Die zugehörige Barrierefreie-Informationstechnik-Verordnung (BITV) verpflichtet öffentliche Einrichtungen, Verwaltungsdokumente in verständlicher Sprache bereitzustellen [7]. Diese gesetzlichen Initiativen verdeutlichen den wachsenden Bedarf an effizienten Methoden zur Textvereinfachung, insbesondere im behördlichen Umfeld.

1.2 Technologische Entwicklung der Textvereinfachung

Die technologische Entwicklung der Textvereinfachung verlief parallel zu Fortschritten im maschinellen Lernen. Erste computerlinguistische Ansätze in den 1990er Jahren verwendeten regelbasierte Systeme zur Vereinfachung bestimmter Textaspekte, wie etwa die syntaktische

Vereinfachung durch Satzteilung [8] oder die lexikalische Vereinfachung durch Wortersetzung [9]. Später, mit dem Aufkommen statistischer Maschinenlernverfahren, entstanden erste datengetriebene Modelle, die durch das Lernen aus parallelen Textkorpora Vereinfachungsmuster automatisch übernehmen konnten.

Ab 2017 revolutionierten Transformer-Architekturen [10] das Feld der Textvereinfachung. Diese neuronalen Netzwerke ermöglichen durch den Einsatz des Aufmerksamkeitsmechanismus (Self-Attention) eine präzisere Erfassung und Verarbeitung sprachlicher Zusammenhänge. Moderne Transformer-Modelle behandeln die Textvereinfachung als Sequenz-zu-Sequenz-Aufgabe, bei der sie anhand paralleler Textkorpora – bestehend aus Original- und vereinfachten Versionen – lernen, wie Texte optimal umformuliert werden können. Die aktuelle Forschung konzentriert sich dabei auf die Verbesserung der Modelle, um die sprachliche Komplexitätsreduktion noch präziser umzusetzen und die ursprüngliche Bedeutung gleichzeitig exakt zu bewahren.

1.3 Problemstellung

Die Verständlichkeit von Verwaltungstexten ist ein zentraler Faktor für die Bürgerbeteiligung und die Transparenz öffentlicher Prozesse. Wenn amtliche Dokumente oder behördliche Mitteilungen sprachlich zu komplex oder unklar formuliert sind, kann dies den Zugang zu wichtigen Informationen erheblich erschweren und das Vertrauen der Bürgerinnen und Bürger in Verwaltungsstrukturen beeinträchtigen. Besonders betroffen sind Menschen, die Englisch als Fremd- oder Zweitsprache erlernt haben, Personen mit kognitiven oder sprachlichen Einschränkungen sowie Bevölkerungsgruppen mit unterschiedlichen Bildungs- und Erfahrungshintergründen. Eine verständliche, barrierefreie Kommunikation ist daher essenziell, um eine gleichberechtigte gesellschaftliche Teilhabe zu gewährleisten [11].

Die manuelle Vereinfachung von Verwaltungstexten stellt jedoch keine praktikable Lösung dar, da sie mit einem hohen Ressourcenaufwand verbunden ist und aufgrund fehlender standardisierter Verfahren zu Qualitätsschwankungen führen kann. Zudem existieren bislang nur begrenzte objektive Bewertungsmethoden, was die Qualitätssicherung zusätzlich erschwert. Da herkömmliche Ansätze kaum skalierbar sind, sind sie ungeeignet, um große Mengen an Texten effizient zu vereinfachen oder sich flexibel an steigende Anforderungen anzupassen. Ohne eine geeignete digitale Infrastruktur bleibt eine umfassende und konsistente Automatisierung dieser Prozesse eine Herausforderung.

Vor diesem Hintergrund eröffnet der Einsatz moderner NLP-Technologien, insbesondere Transformer-basierter Large Language Models (LLMs), neue Möglichkeiten. Diese Modelle ermöglichen eine automatisierte, konsistente und qualitativ hochwertige Vereinfachung komplexer Verwaltungstexte und

tragen dazu bei, die Verständlichkeit und Barrierefreiheit in der behördlichen Kommunikation nachhaltig zu verbessern [12].

1.4 Zielsetzung der Arbeit

Die vorliegende Bachelorarbeit verfolgt das Ziel, ein KI-gestütztes System zur automatisierten Vereinfachung behördlicher Texte zu entwickeln. Damit soll allen Bürgerinnen und Bürgern der Zugang zu wichtigen Informationen erleichtert werden. Im Fokus steht dabei die Vereinfachung der häufig komplexen Verwaltungssprache, um das Verständnis und die Nutzung administrativer Informationen zu fördern.

Um dieses Ziel zu erreichen, wird in dieser Arbeit ein spezifischer Ansatz zur Anpassung moderner transformer-basierter LLMs erarbeitet. Durch gezieltes Fine-Tuning lernen die Sprachmodelle, komplexe Verwaltungstexte so zu vereinfachen, dass der Inhalt vollständig und präzise erhalten bleibt, während die sprachliche Komplexität reduziert wird. Ein eigens erstellter Trainingsdatensatz bildet die Grundlage, bestehend aus Originaltexten und ihren vereinfachten Versionen, um die Bedürfnisse der Verwaltungskommunikation zu adressieren.

Da vortrainierte Modelle auf Englisch eine höhere Leistung zeigen, konzentriert sich diese Arbeit zunächst auf die Vereinfachung englischsprachiger Verwaltungstexte. Diese Entscheidung basiert darauf, dass die zugrunde liegenden Sprachmodelle auf großen, überwiegend englischsprachigen Datensätzen trainiert wurden und somit bei englischen Texten eine bessere Qualität und Genauigkeit liefern können. Langfristig bietet dieser Ansatz jedoch eine Grundlage, um ähnliche Systeme für andere Sprachen, einschließlich Deutsch, zu entwickeln.

Mit der Entwicklung dieses Systems leistet die Arbeit einen Beitrag zur Demokratisierung des Zugangs zu behördlichen Informationen. Sie unterstützt das Ziel einer inklusiven und bürgernahen Verwaltungskommunikation und trägt zu einer langfristigen Verbesserung des Verständnisses öffentlicher Informationen bei.

1.5 Methodisches Vorgehen und Strukturierung der Forschungsarbeit

Die vorliegende Bachelorarbeit untersucht die Entwicklung eines innovativen Ansatzes zur automatisierten Vereinfachung administrativer Texte unter Einsatz moderner Sprachmodellen. Der methodische Fokus liegt auf drei folgenden linguistischen Ebenen: lexikalisch, syntaktisch und semantisch. Auf der lexikalischen Ebene wird der Schwerpunkt auf die Ersetzung komplexer Fachtermini durch allgemein verständliche Begriffe gelegt, um die Verständlichkeit der

Verwaltungskommunikation zu verbessern. Die syntaktische Ebene zielt darauf ab, verschachtelte Satzstrukturen in klarere, prägnantere Formulierungen umzuwandeln, um die kognitive Belastung der Leserschaft zu reduzieren. Auf der semantischen Ebene wird sichergestellt, dass die ursprüngliche Bedeutung und juristische Präzision der Texte erhalten bleibt, um die inhaltliche Integrität zu bewahren.

Die praktische Umsetzung erfolgt in mehreren Schritten, beginnend mit der Erstellung eines spezialisierten Korpus aus Verwaltungstexten, das sowohl Originaltexte als auch deren vereinfachte Versionen umfasst. Darauf folgt das Fine-Tuning der vortrainierten Sprachmodelle, bei dem spezifische Anforderungen der Textvereinfachung berücksichtigt und relevante Modellparameter optimiert werden. Die Evaluierung der Modelle erfolgt sowohl auf Grundlage quantitativer Metriken wie BLEU und SARI, die technische Aspekte der Textvereinfachung bewerten, als auch durch qualitative Nutzerbewertungen, die die Verständlichkeit und Lesbarkeit der generierten Texte aus der Perspektive der Zielgruppe analysieren.

Nach der Einleitung und der Darstellung des historischen Hintergrunds der Textvereinfachung im administrativen Kontext konzentriert sich diese Arbeit in den folgenden Kapiteln auf die methodische und technische Umsetzung. Im zweiten Kapitel werden die theoretischen Grundlagen moderner Transformer-Architekturen erläutert, die die technische Basis der eingesetzten Sprachmodelle bilden. Das dritte Kapitel widmet sich dem Implementierungsprozess und dem Fine-Tuning der Modelle, einschließlich der spezifischen Anpassungen an die Anforderungen der Textvereinfachung. Im vierten Kapitel werden die Ergebnisse präsentiert und sowohl objektiv anhand quantitativer Metriken als auch subjektiv aus der Perspektive der Nutzer analysiert. Abschließend behandelt das fünfte Kapitel die Diskussion der Ergebnisse, gibt eine zusammenfassende Bewertung der Forschungsergebnisse und skizziert mögliche zukünftige Forschungsansätze.

Diese vielschichtige Methodik ermöglicht eine gründliche Untersuchung des Themas aus unterschiedlichen Perspektiven. Sie verbindet theoretische Grundlagen der Textvereinfachung mit praktischer Modellanwendung und einer fundierten Evaluierung. Die so entwickelte Herangehensweise bildet die wissenschaftliche Basis zur Bewertung der Effektivität von vortrainierten Sprachmodellen in der Verwaltungstextvereinfachung und für die darauf basierenden Schlussfolgerungen dieser Bachelorarbeit.

Die zuvor beschriebene detaillierte Gliederung wird durch Abbildung 1 visualisiert. Diese Darstellung bietet eine klare Übersicht über den strukturellen Aufbau der Arbeit und verdeutlicht die logische Abfolge der einzelnen Kapitel und methodischen Schritte.

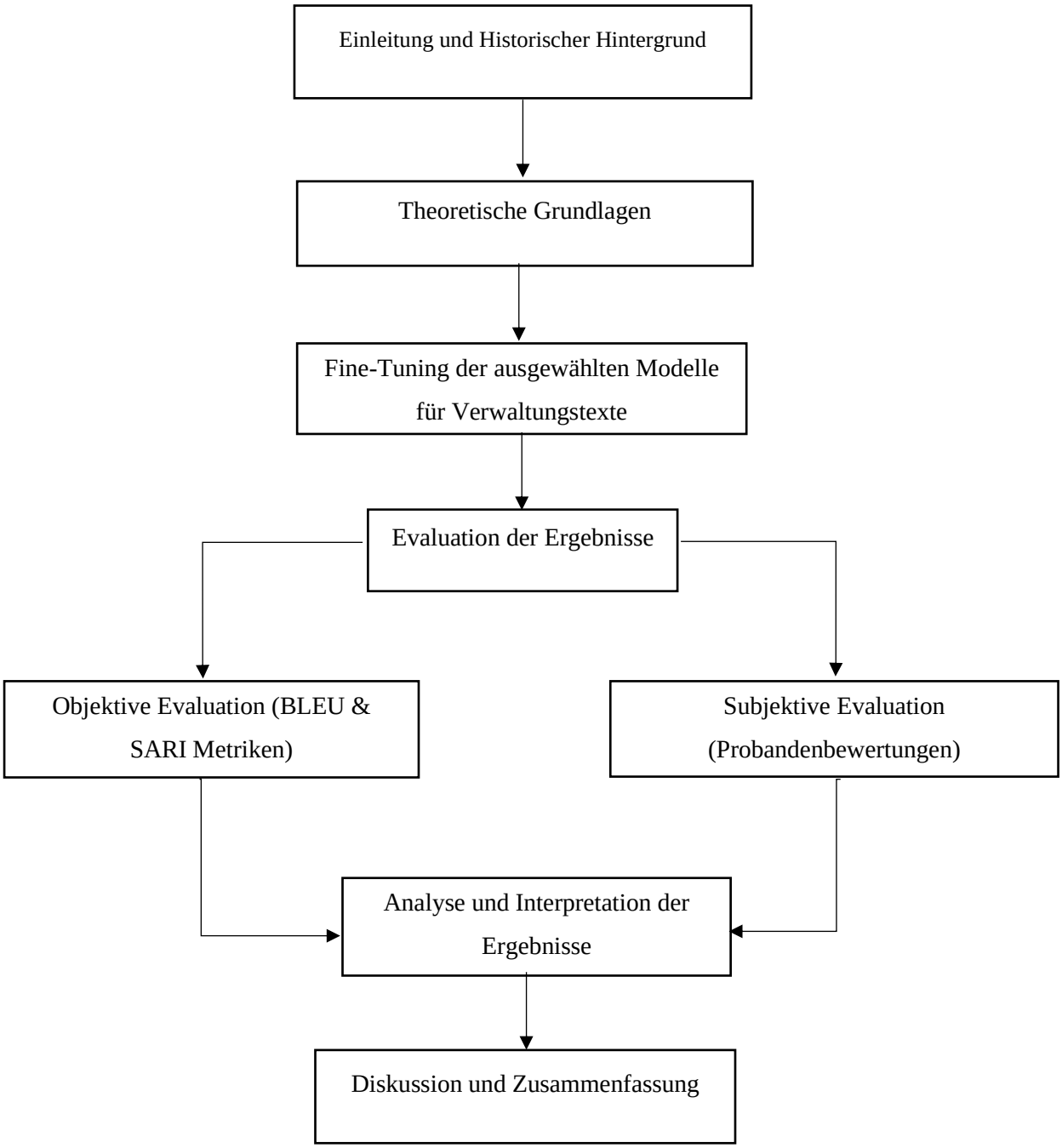

Abbildung 1: Struktur der Bachelorarbeit

2 Theoretische Grundlagen

Die Verständlichkeit von Texten spielt eine zentrale Rolle in der schriftlichen Kommunikation, insbesondere in Verwaltung, Recht und Bildung. Komplexe Fachsprache kann Menschen mit eingeschränkten Sprachkenntnissen oder kognitiven Beeinträchtigungen den Zugang zu wichtigen Informationen erschweren. Um dem entgegenzuwirken, wurden verschiedene Sprachvereinfachungsansätze entwickelt, die sich in Leichte Sprache (Easy Language) und Einfache Sprache (Plain Language) unterteilen.

Dieses Kapitel gibt einen Überblick über diese beiden Ansätze, ihre gesetzlichen Grundlagen sowie ihre Bedeutung für die Textvereinfachung durch moderne KI-Modelle. Zudem wird der aktuelle Stand der Technik im Bereich der LLMs erläutert und deren Entwicklung bis hin zur Transformer-Architektur nachgezeichnet.

2.1 Überblick über Leichte und Einfache Sprache

Die Leichte Sprache zielt darauf ab, Informationen für Menschen mit eingeschränkten Sprachkompetenzen zugänglich zu machen. Sie folgt festen Regeln, die sowohl sprachliche als auch gestalterische Aspekte berücksichtigen. Zu den zentralen Merkmalen zählen die Verwendung kurzer Sätze, die in der Regel nur eine Aussage enthalten, sowie die Nutzung einfacher und bekannter Wörter. Fachbegriffe und Fremdwörter werden vermieden oder unmittelbar erklärt. Zusammengesetzte Wörter werden durch Bindestriche getrennt, um die Lesbarkeit zu erhöhen. Zusätzlich wird großer Wert auf eine übersichtliche Textgestaltung gelegt, wobei jeder Satz in einer eigenen Zeile steht und unterstützende Bilder oder Symbole verwendet werden, um den Inhalt zu veranschaulichen. Diese Maßnahmen sollen sicherstellen, dass die Texte für die Hauptzielgruppe leicht verständlich sind. Die Leichte Sprache richtet sich primär an Personen mit geistiger Behinderung, Demenz, Aphasie sowie an Personen mit sehr geringen Lesekompetenzen. Auch Nicht-Muttersprachler profitieren von Leichter Sprache [13].

Im Gegensatz dazu ist die Einfache Sprache weniger strikt reglementiert und näher an der Standardsprache angesiedelt. Sie zielt darauf ab, komplexe Informationen für eine breitere Öffentlichkeit zugänglich zu machen. Empfehlungen für die Einfache Sprache beinhalten die Verwendung kurzer, klar strukturierter Sätze mit maximal 15 bis 20 Wörtern, die Vermeidung von Fachjargon oder dessen Erklärung sowie eine klare und logische Gliederung des Textes. Obwohl es kein festes Regelwerk gibt, orientiert sich die Einfache Sprache an Prinzipien, die darauf abzielen, die Verständlichkeit zu erhöhen, ohne den Inhalt übermäßig zu simplifizieren. Die Zielgruppe der Einfachen Sprache ist vielfältig und umfasst neben Menschen mit leichten kognitiven Beeinträchtigungen auch

Personen mit geringen Sprachkenntnissen, wie Migrant*innen und Geflüchtete, sowie Menschen mit geringer Lesekompetenz. Auch Fachtexte, die für Laien verständlich aufbereitet werden sollen, profitieren von der Anwendung der Einfachen Sprache [14].

2.1.1 Gesetzliche Grundlagen und historische Entwicklung

Die Entwicklung vereinfachter Sprachstandards im englischsprachigen Raum begann bereits in den 1940er Jahren, als erste Initiativen zur Förderung verständlicher Verwaltungssprache entstanden. Besonders in den Vereinigten Staaten und Großbritannien wurden früh Konzepte zur Verbesserung der Verständlichkeit administrativer und rechtlicher Dokumente entwickelt [15]. In den USA erhielt die Bewegung für Einfache Sprache mit dem Plain Writing Act of 2010 [6] eine gesetzliche Grundlage. Dieses Gesetz verpflichtet Bundesbehörden, Informationen in einer klaren, präzisen und verständlichen Sprache bereitzustellen [6].

In Großbritannien setzt sich die Plain English Campaign seit den 1980er Jahren für eine klare und verständliche Verwaltungssprache ein. Die Organisation vergibt offizielle Zertifizierungen für leicht verständliche Texte [16]. Auch Kanada und Australien haben ähnliche Programme entwickelt, um eine bessere Verständlichkeit in behördlichen Dokumenten zu gewährleisten [17].

Parallel dazu entwickelte sich die Bewegung für Leichte Sprache, die speziell für Menschen mit kognitiven Einschränkungen oder begrenzten Sprachkenntnissen geschaffen wurde. Eine der frühesten europäischen Initiativen wurde 1998 von Inclusion Europe ins Leben gerufen, um Standards für verständliche Texte für Menschen mit geistigen Behinderungen zu etablieren [18]. Diese Leitlinien wurden in mehreren Ländern übernommen, darunter das Konzept Leichte Sprache in Australien, das auf die Bedürfnisse von Menschen mit Lernschwierigkeiten und Sprachbarrieren zugeschnitten ist [17].

Die gesetzlichen Vorschriften zur sprachlichen Barrierefreiheit haben sich in den letzten Jahrzehnten kontinuierlich erweitert. In den USA setzen sich Organisationen wie das National Center on Disability and Access to Education (NCDAE) für eine klare und verständliche Sprache in Bildung und Verwaltung ein [19]. Diese Entwicklungen tragen dazu bei, dass Verwaltungs- und rechtliche Texte für eine größere Bevölkerungsgruppe zugänglich werden, insbesondere für Menschen mit geringer formaler Bildung, ältere Menschen und Migranten.

2.2 Stand der Technik

Die Vereinfachung von Verwaltungstexten mithilfe von Sprachmodellen ist ein wachsendes Forschungsfeld, das darauf abzielt, komplexe administrative Dokumente verständlicher und zugänglicher zu machen. Sprachmodellen bieten eine innovative Lösung, da sie in der Lage sind,

linguistische Strukturen effizient zu analysieren, Textkomplexität zu reduzieren und dabei die wesentlichen Inhalte zu bewahren.

2.2.1 Fortschritte bei der Textvereinfachung mit LLMs

Moderne LLMs sind in der Lage, komplexe Sätze umzustrukturieren, Fachterminologie zu vereinfachen und eine höhere Lesbarkeit zu gewährleisten. Zu den am häufigsten untersuchten Modellen gehören sowohl Open-Weight-Modelle wie BLOOM [20], T5 [21] und LLaMA [22] als auch Closed-Weight-Modelle wie GPT-4 [23] und dessen API-basierte Anwendungen. Diese Modelle haben gezeigt, dass sie Textvereinfachung effektiv durchführen können.

T5 wurde als ein vielseitiges Text-zu-Text-Modell entwickelt, das erfolgreich in verschiedenen NLP-Aufgaben wie maschineller Übersetzung, Textzusammenfassung und auch Textvereinfachung eingesetzt wurde [21]. Es behandelt Vereinfachung als eine Form der Transformation zwischen komplexer und vereinfachter Sprache und hat sich als leistungsfähig für bildungsbezogene Anwendungen erwiesen.

Ein weiterer bedeutender Fortschritt in diesem Bereich ist die Entwicklung Parameter-Efficient Fine-Tuning (PEFT) [24]. Diese Technik ermöglicht es, große Modelle mit minimalem Rechenaufwand an domänenspezifische Aufgaben anzupassen [25] [26] [27].

2.2.2 Domänenspezifische LLMs zur Textvereinfachung

Die Anpassung vortrainierter LLMs an spezifische Domänen ist ein wichtiger Forschungsbereich. Studien zeigen, dass spezialisierte Modelle besonders effektiv in der Verwaltung, Medizin, Rechtsprechung und anderen komplexen Bereichen eingesetzt werden können.

Eine aktuelle Untersuchung von Musumeci et al. [28] zeigt, wie LLMs zur Generierung semi-strukturierter Verwaltungstexte eingesetzt werden können. Ihr Ansatz kombiniert Prompt Engineering mit Multi-Agenten-Systemen, wodurch eine verbesserte Strukturierung und Vereinfachung von behördlichen Dokumenten erreicht wird. Durch die automatische Analyse und Umstrukturierung semi-strukturierter Dokumente konnten bedeutende Fortschritte in der verständlichen Aufbereitung von Inhalten erzielt werden [28].

Mandravickaitė et al. [29] untersuchten die Vereinfachung litauischer Verwaltungstexte mit den Modellen mT5 und mBART. Die Ergebnisse zeigten, dass mBART konsistente Verbesserungen in den Metriken BLEU, SARI und ROUGE erzielte. Das Modell konnte juristische Präzision bewahren und gleichzeitig komplexe Satzstrukturen vereinfachen. Diese Arbeit unterstreicht die Bedeutung der Feinabstimmung vortrainierter Modelle auf bestimmte Sprachen und Fachbereiche [29].

Im deutschsprachigen Raum hebt sich die Studie von Klöser et al. [30] hervor. Sie entwickelten ein semi-synthetisches Korpus, um die Problematik der Datenknappheit in der Textvereinfachung zu adressieren. Durch das Training von LLMs mit bis zu 13 Milliarden Parametern konnte die Verständlichkeit von Verwaltungstexten erheblich verbessert werden. Die Autoren weisen darauf hin, dass sowohl automatische als auch manuelle Bewertungen die Wirksamkeit von spezialisierten LLMs bestätigen [30].

Martínez et al. [31] untersuchten den Einsatz von LLaMA 2 für die Verwaltungstextvereinfachung, insbesondere für Personen mit kognitiven Beeinträchtigungen. Durch gezieltes Fine-Tuning auf ein speziell erstelltes Korpus wurden Verwaltungsdokumente gemäß den "Easy to Read"-Richtlinien umformuliert. Expertenbewertungen bestätigten, dass diese Methode eine bessere Verständlichkeit und Zugänglichkeit ermöglicht [31].

2.2.3 Prompt-Engineering und automatische Optimierung

Ein wesentlicher Faktor für die Qualität der Vereinfachung ist das Prompt-Engineering, das gezielt darauf abzielt, LLMs mit effektiven Eingabeanweisungen zu optimieren. Brown et al. [32] heben hervor, dass In-Context Learning (ICL) und Few-Shot Learning es ermöglichen, LLMs durch die Bereitstellung weniger Beispiele gezielt auf eine bestimmte Umformulierung auszurichten [32]. Diese Techniken sind besonders relevant für die Verwaltungstextvereinfachung, da sie es ermöglichen, die Modelle direkt auf die gewünschten sprachlichen Anpassungen zu konditionieren, ohne dass ein aufwendiges Fine-Tuning erforderlich ist.

Ein weiterer vielversprechender Ansatz ist das Chain-of-Thought (CoT) Prompting, das von Wei et al. [33] entwickelt wurde. Diese Technik ermutigt LLMs dazu, komplexe Aufgaben durch das Generieren von Zwischenschritten zu lösen, anstatt direkt eine Endantwort zu geben. [33] zeigten, dass CoT-Prompting die Leistung von Modellen in logischen, arithmetischen und textbasierten Schlussfolgerungen signifikant verbessern kann, indem es eine strukturierte Denkweise simuliert. Dies macht die Methode besonders relevant für die schrittweise Vereinfachung komplexer Verwaltungstexte, indem die Umformulierung in logische Teilprozesse zerlegt wird.

Ein weiterer wichtiger Ansatz ist die automatische Prompt-Optimierung. Amatriain [34] hebt hervor, wie fortgeschrittene Methoden wie CoT Prompting und Reflection genutzt werden können, um die Genauigkeit und Transparenz der Modelle zu verbessern. Diese Techniken strukturieren die Denkprozesse der Modelle explizit, sodass logischere und konsistentere Ergebnisse erzielt werden können. Zusätzlich zeigt [34], dass der Einsatz von Automatic Prompt Engineering (APE) die Effizienz und Effektivität der Vereinfachung erheblich steigern kann.

2.2.4 Herausforderungen und offene Fragen

trotz signifikanter Fortschritte in der LLM-gestützten Vereinfachung von Verwaltungstexten bestehen weiterhin mehrere Herausforderungen. Eine zentrale Schwierigkeit liegt in der Balance zwischen Verständlichkeit und Präzision, insbesondere in rechtlichen oder administrativen Kontexten. Während die Vereinfachung von Texten die Zugänglichkeit erhöht, besteht die Gefahr, dass essenzielle juristische Feinheiten verloren gehen, was zu Missverständnissen führen kann. Zudem bleibt das Problem der "Halluzinationen" bestehen, bei den Modellen plausible, aber faktisch falsche Informationen generieren. Dieses Phänomen kann insbesondere in Verwaltungstexten zu erheblichen Fehlinformationen führen. Gekhman et al. [35] untersuchten dieses Problem und stellten fest, dass das Einführen neuer Fakten während des Fine-Tunings die Tendenz der Modelle zu Halluzinationen erhöhen kann.

Ein weiterer offener Forschungsbereich ist die kulturelle und sprachliche Anpassung von LLMs. Gooding [36] diskutierte die ethischen Implikationen der Textvereinfachung und betonte, dass Modelle kulturelle Unterschiede nicht immer ausreichend berücksichtigen, was zu unangemessenen oder ungenauen Vereinfachungen führen kann. Dies unterstreicht die Notwendigkeit, LLMs so zu gestalten, dass sie kulturelle Nuancen und sprachliche Besonderheiten adäquat erfassen.

Zusammenfassend zeigt die aktuelle Forschung, dass LLMs ein enormes Potenzial für die Vereinfachung von Verwaltungstexten besitzen. Dennoch bleiben Herausforderungen wie die Sicherstellung der Faktentreue, die kulturelle Adaption und die Balance zwischen Vereinfachung und Präzision zentrale Forschungsfragen für die Zukunft.

2.3 Vorgänger der Transformer-Architektur

Bevor die Transformer-Architektur [10] in der NLP-Forschung Einzug hielt, basierten Ansätze zur Sprachverarbeitung auf regelbasierten Systemen, statistischen Modellen und frühen neuronalen Netzwerken. Diese Methoden hatten jedoch erhebliche Einschränkungen bei der Handhabung von Kontext, langen Sequenzen und der Modellierung semantischer Beziehungen. Insbesondere scheiterten viele dieser Ansätze daran, die Bedeutung der Wortreihenfolge und den Kontext eines gesamten Satzes oder Dokuments effektiv zu erfassen. Die Entwicklung der Transformer-Architektur war ein revolutionärer Schritt, der viele dieser Herausforderungen überwand und die Grundlage für moderne LLMs schuf.

2.3.1 Regelbasierte und frühe maschinelle Lernverfahren

Frühe Ansätze in der NLP-Forschung stützten sich stark auf regelbasierte Systeme und statische Modelle. Regelbasierte Systeme, wie sie in der maschinellen Übersetzung oder Syntaxanalyse eingesetzt

wurden, nutzten explizit definierte grammatikalische Regeln zur Textverarbeitung. Diese Systeme waren jedoch stark limitiert, da sie Schwierigkeiten hatten, mit der Komplexität natürlicher Sprache und variierenden Kontexten umzugehen [37].

Eine statische Erweiterung dieser Ansätze war das Bag-of-Words (BoW)-Modell. Dieses Modell ignorierte die Reihenfolge der Wörter und repräsentierte Texte lediglich durch die Häufigkeit der darin enthaltenen Begriffe [37]. BoW wurde häufig in Textklassifikations- und Informationsabrufsystemen eingesetzt. Obwohl es effizient war, führte der Verlust der Kontextinformationen dazu, dass semantische Zusammenhänge nur unzureichend erfasst wurden.

Probabilistische Modelle wie Hidden Markov Models (HMMs) und Conditional Random Fields (CRFs) boten eine statische Herangehensweise, die besser geeignet war, Unsicherheiten in der Sprache zu modellieren. HMMs fanden Anwendungen in der Spracherkennung und maschinellen Übersetzung, während CRFs vor allem bei Aufgaben wie Named Entity Recognition (NER) und Sequenzkennzeichnung erfolgreich eingesetzt wurden [37]. Trotz ihrer Erfolge hatten diese Modelle Schwierigkeiten, lange Abhängigkeiten und komplexe semantische Beziehungen zu modellieren.

2.3.2 Neuronale Ansätze

Die Einführung neuronaler Netzwerke führte zu signifikanten Fortschritten in der Sprachverarbeitung. Besonders Feedforward- und rekurrente Architekturen ermöglichten eine effizientere Verarbeitung und Generalisierung von Sprachmustern.

2.3.2.1 Feedforward Neural Networks

Die Einführung von Feedforward Neural Networks markierte einen Wendepunkt in der NLP-Forschung. Diese Netzwerke basieren auf einer Architektur, die eine Eingabe durch mehrere Schichten von Neuronen verarbeitet, um eine Ausgabe zu generieren. Sie waren besonders effektiv in Aufgaben wie Textklassifikation und Sentimentanalyse. Allerdings konnten sie die Reihenfolge und den Kontext von Wörtern nicht berücksichtigen, was [38] dazu veranlasste, ein probabilistisches Sprachmodell zu entwickeln, das Wortrepräsentationen in einem kontinuierlichen Raum erlernt. Dies machte deutlich, dass spezialisierte Modelle für sequentielle Daten erforderlich waren.

2.3.2.2 Recurrent Neural Networks (RNNs)

Die Einführung von Recurrent Neural Networks (RNNs) brachte eine entscheidende Verbesserung in der Verarbeitung von Sprachsequenzen. Durch Rückkopplungsschleifen konnten frühere Eingaben

gespeichert und bei späteren Berechnungen berücksichtigt werden. Diese Eigenschaft machte sie besonders wertvoll für maschinelle Übersetzung, Sprachmodellierung und Textgenerierung.

Ein zentrales Problem dieser Netzwerke war jedoch das sogenannte Vanishing Gradient Problem, welches dazu führte, dass Informationen über lange Sequenzen hinweg nur schwer erfasst wurden.

2.3.2.3 LSTM und GRUs

Um die Schwächen von RNNs zu beheben, wurden Long Short-Term Memory (LSTM)-Netzwerke [39] entwickelt. Diese Netzwerke führten Gedächtniszellen ein, die entscheiden können, welche Informationen über längere Zeiträume hinweg gespeichert oder vergessen werden sollen. LSTMs wurden für viele NLP-Aufgaben wie maschinelle Übersetzung und Sprachmodellierung eingesetzt.

Später wurden Gated Recurrent Units (GRUs) entwickelt, die eine kompaktere Struktur als LSTMs aufweisen und ähnliche Ergebnisse liefern konnten. Trotz dieser Fortschritte blieb die sequenzielle Verarbeitung ein Engpass, da diese Modelle weiterhin auf eine schrittweise Verarbeitung von Token angewiesen waren. Die Unterschiede zwischen RNNs, LSTMs und GRUs sind in Abbildung 2 dargestellt.

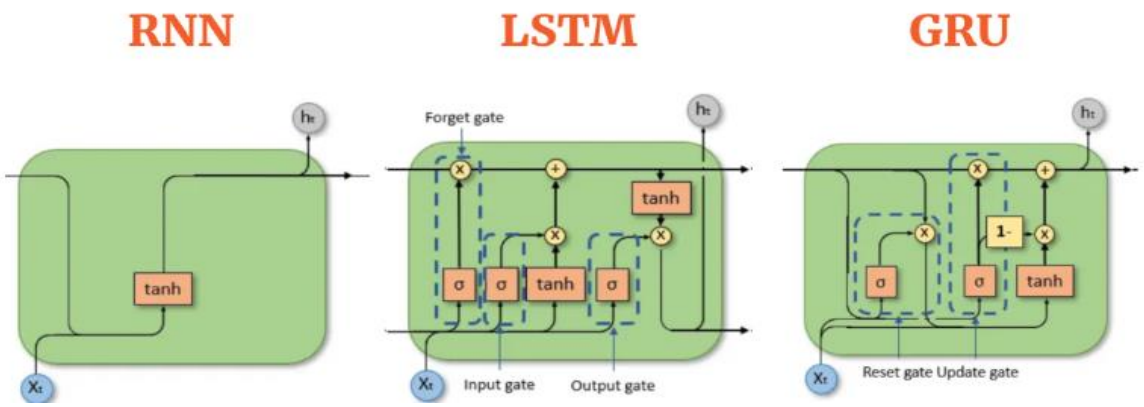

Abbildung 2: Vergleich verschiedener Sequenzmodelle [40]

2.3.3 Attention-Mechanismus

Die Einführung von Attention-Mechanismen war ein entscheidender Schritt zur Verbesserung der Verarbeitung natürlicher Sprache. Der bahnbrechende Ansatz wurde erstmals im Kontext von Seq2Seq-Modellen mit Attention von [41] vorgestellt. Diese Technik ermöglichte es Modellen, sich dynamisch auf relevante Teile einer Eingabesequenz zu konzentrieren, anstatt alle Informationen gleich zu gewichten. Attention adressierte damit die Limitierungen früherer Modelle wie LSTMs, die Schwierigkeiten hatten, langreichweitige Abhängigkeiten effizient zu modellieren.

Attention-Mechanismen wurden zunächst in der maschinellen Übersetzung eingesetzt, wo sie dazu beitrugen, Quell- und Zielwörter effektiver zu verknüpfen. Sie boten eine klare Verbesserung der Genauigkeit und eröffneten neue Möglichkeiten für die Verarbeitung komplexer Sequenzen. Die Bedeutung dieses Fortschritts zeigt sich in der weiteren Entwicklung von Sprachmodellen, wie in Abbildung 3 dargestellt. Hier wird der Übergang von statistischen Sprachmodellen zu neuronalen Netzwerken sowie der Einfluss von Attention auf die Entwicklung moderner Transformer-Modelle veranschaulicht.

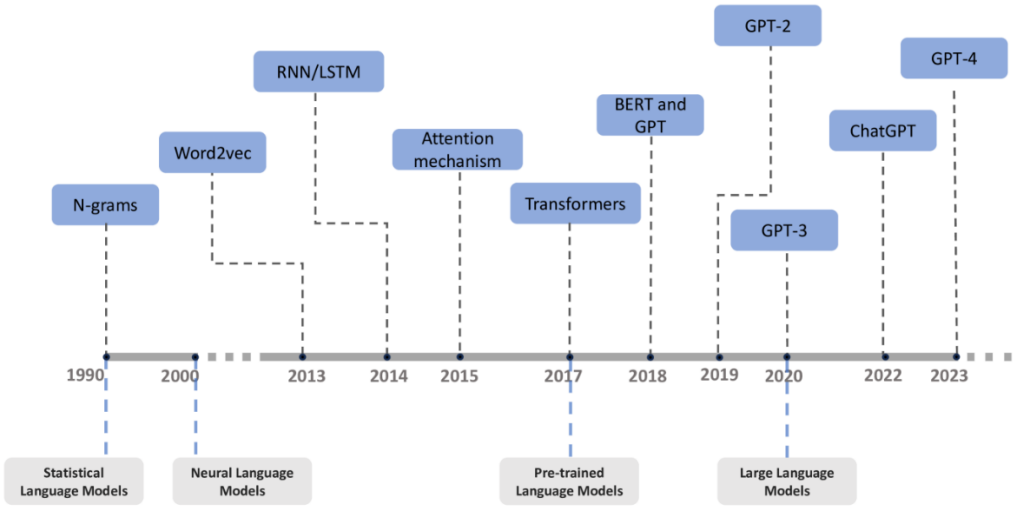

Abbildung 3: Geschichte und Entwicklung von Sprachmodellen [42]

2.3.4 Self-Attention als Grundlage der Transformer-Architektur

Ein zentraler Durchbruch war die Einführung von Self-Attention in der Arbeit "Attention is All You Need" von [10]. Self-Attention ermöglicht es Modellen, Beziehungen zwischen allen Wörtern in einer Sequenz gleichzeitig zu berücksichtigen, unabhängig von deren Entfernung. Diese Fähigkeit zur parallelen Verarbeitung war eine Revolution im Vergleich zu LSTMs und GRUs, die Wörter nur sequentiell verarbeiteten.

Self-Attention berechnet Gewichtungen für jedes Wort basierend auf seiner Relevanz für andere Wörter in der Sequenz. Dadurch können Modelle besser kontextuelle Informationen erfassen und auch lange Abhängigkeiten effizient modellieren. Die Kombination von Self-Attention mit Feedforward-Schichten und Layer-Normalisierung führte zur Transformer-Architektur, die Skalierbarkeit, Geschwindigkeit und Genauigkeit in NLP-Aufgaben erheblich verbesserte.

2.4 Übergang zu Transformer-Modellen

Mit der Einführung des Transformer-Modells [10] wurde ein paradigmatischer Wandel in der Verarbeitung von Sequenzdaten vollzogen. Im Gegensatz zu früheren Modellen, die auf rekurrenten oder konvolutionalen Architekturen basierten, verzichtet der Transformer vollständig auf diese Strukturen und nutzt stattdessen eine Self-Attention-Mechanik. Diese ermöglicht die parallele Modellierung globaler Abhängigkeiten innerhalb einer Eingabesequenz und verbessert dadurch sowohl die Verarbeitungsgeschwindigkeit als auch die Skalierbarkeit und Effizienz, insbesondere bei langen Sequenzen.

2.4.1 Transformer-Architektur

Die Transformer-Architektur, wie in Abbildung 4 dargestellt, folgt einem Encoder-Decoder-Design. Der Encoder verarbeitet die Eingabesequenz über mehrere Schichten hinweg und generiert eine kontinuierliche Repräsentation, die anschließend vom Decoder genutzt wird, um die Zielsequenz zu erzeugen. Beide Komponenten bestehen aus einer Sequenz identischer Schichten, die jeweils eine Multi-Head-Attention-Schicht und ein Position-wise Feed-Forward-Netzwerk als wesentliche Bausteine enthalten. Um Stabilität und eine effizientere Konvergenz während des Trainings zu gewährleisten, werden Residual-Verbindungen und Layer-Normalisierung integriert.

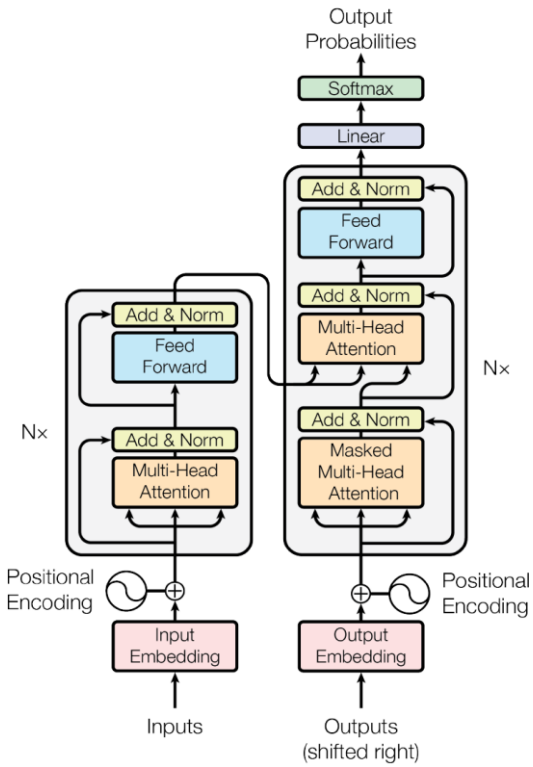

Abbildung 4: Die Transformer-Modellarchitektur [10]

2.4.2 Encoder-Decoder-Struktur

Der Encoder transformiert die Eingabesequenz zunächst in Vektorrepräsentationen (Embeddings), die durch Positional Encodings ergänzt werden, um die Reihenfolge der Tokens explizit zu berücksichtigen [10]. Anschließend wird diese angereicherte Eingabe in die erste Encoderschicht überführt, die auf dem Multi-Head-Attention-Mechanismus basiert [10].

Der Decoder hingegen verfügt über eine zusätzliche Masked Multi-Head-Attention-Schicht, die sicherstellt, dass während der Generierung einer Sequenz zukünftige Tokens nicht berücksichtigt werden. Diese Maskierung gewährleistet die Kausalität der Modellierung und verhindert Datenlecks innerhalb des Modells [10].

2.4.3 Multi-Head Attention und Scaled Dot-Product Attention

Ein zentraler Mechanismus des Transformers ist die Multi-Head Attention, die auf dem Prinzip der Scaled Dot-Product Attention basiert. Dieser Mechanismus ermöglicht es dem Modell, verschiedene Teile einer Sequenz parallel zu betrachten, wodurch komplexe Abhängigkeiten und Beziehungen zwischen Tokens effizient erfasst werden können [10].

Die Berechnung der Scaled Dot-Product Attention folgt der Gleichung (1):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

Hier stehen Q (Queries), K (Keys) und V (Values) für die Eingabematrizen, die aus den eingebetteten Token-Vektoren einer Sequenz gebildet werden. Die Skalierung durch $\sqrt{d_k}$, wobei d_k die Dimension der Keys ist, verhindert numerische Instabilitäten bei großen Werten. Der Softmax-Operator normalisiert die Gewichtungen der Werte V, sodass sich die Aufmerksamkeit selektiv auf relevante Tokens konzentrieren kann [10].

Die Architektur dieses Mechanismus wird in Abbildung 5 dargestellt. Sie zeigt den mehrstufigen Prozess, beginnend mit der Matrixmultiplikation von Q und K, gefolgt von der Skalierung und einer optionalen Maskierung, um kausale Abhängigkeiten sicherzustellen. Schließlich wird der Softmax-Operator angewendet und die gewichteten Werte mit V multipliziert, um die endgültige Ausgabe zu generieren.

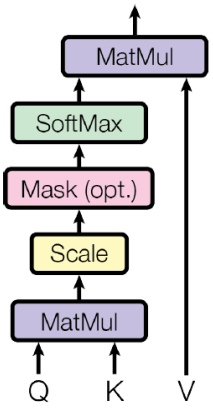

Abbildung 5: Scaled Dot-Product Attention [10]

Multi-Head Attention erweitert das Konzept der Scaled Dot-Product Attention, indem mehrere Attention-Köpfe parallel arbeiten. Dadurch kann das Modell unterschiedliche Aspekte der Eingabesequenz simultan erfassen, was zu einer reichhaltigeren Repräsentation des Kontexts führt. Die mathematische Darstellung des Multi-Head Attention Mechanismus ist in Gleichung (2) definiert:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \quad (2)$$

Hierbei wird jeder Kopf durch individuelle Gewichtungsmatrizen W^Q, W^K, W^V transformiert, bevor die Ergebnisse zusammengeführt werden.

Die schematische Darstellung in Abbildung 6 zeigt, wie mehrere parallele Scaled Dot-Product Attention-Berechnungen durchgeführt werden. Jede Attention-Einheit verarbeitet die Eingabe unabhängig, sodass unterschiedliche Aspekte der Sequenz erfasst werden können. Anschließend werden die Ergebnisse der einzelnen Attention-Köpfe zusammengeführt (Concat-Schritt) und mittels einer linearen Transformation weiterverarbeitet. Diese Visualisierung ist entscheidend für das Verständnis, wie der Transformer verschiedene semantische Relationen innerhalb der Sequenz effizient modelliert.

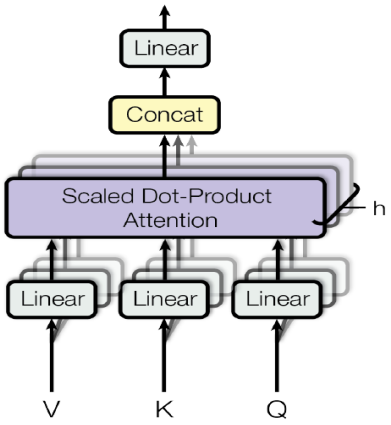

Abbildung 6: Multi-Head Attention [10]

2.4.4 Position-wise Feed-Forward-Netzwerk

Jede Schicht des Encoders und Decoders enthält ein Feed-Forward-Netzwerk (FFN), das unabhängig auf jedes Token angewendet wird. Dieses Netzwerk dient der nichtlinearen Transformation der Eingabedaten, um die Repräsentationsfähigkeit des Modells zu verbessern [10]. Mathematisch wird das FFN durch die folgende Gleichung (3) definiert:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (3)$$

Dabei handelt es sich um eine zweischichtige Transformation mit einer Zwischendimension von $d_{ff} = 2048$. Die erste lineare Transformation (W_1, b_1) projiziert den Eingabevektor in einen höherdimensionalen Raum, gefolgt von einer ReLU-Aktivierung $max(0, x_1)$, die Nichtlinearität einführt. Anschließend erfolgt eine zweite lineare Transformation (W_2, b_2), die die Daten wieder in die ursprüngliche Dimension zurückprojiziert.

2.4.5 Positional Encodings

Da der Transformer keine rekurrenten Strukturen verwendet, wird die Reihenfolge der Tokens durch Positional Encodings dargestellt. Diese Formeln werden mit den Eingabe-Embeddings kombiniert und nutzen sinus- und cosinusbasierten Funktionen, wie in Gleichung (4) dargestellt:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right), PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (4)$$

Hierbei repräsentieren pos die Position des Tokens und i die Dimension. Diese Funktionen gewährleisten, dass das Modell die Reihenfolge der Sequenz erfassen kann.

Nachdem die Architektur des Transformers ausführlich erläutert wurde, bietet die folgende Tabelle 1 einen strukturierten Vergleich zwischen RNNs, LSTMs, GRUs und der Transformer-Architektur. Dabei werden zentrale Eigenschaften, Vorteile und Einschränkungen der einzelnen Modelle gegenübergestellt, um ihre jeweiligen Stärken und Schwächen zu verdeutlichen.

Tabelle 1: Vergleich verschiedener Sequenzmodelle [40]

Beschreibung	RNN	LSTM	GRU	Transformer
Überblick	Iterative Verarbeitung von Sequenzen, wobei frühere Outputs als Inputs für die nächsten Schritte dienen.	Erweiterung von RNNs zur besseren Erfassung von Langzeitabhängigkeiten.	Vereinfachte Version von LSTMs mit optimiertem Gating-Mechanismus.	Nutzt Selbstaufmerksamkeit anstelle von Rekurrenz für parallele Verarbeitung.
Hauptmerkmale	- Speichert frühere Informationen über rekurrente Verbindungen.	- Nutzt Gates (Eingangs-, Vergessens-, Ausgangsgate) zur Regulierung des Informationsflusses	- Verwendet Reset- und Update-Gates zur effizienten Steuerung der Informationsweitergabe.	- Selbstaufmerksamkeitsmechanismus gewichtet Eingaben dynamisch. - Besteht aus Encoder-Decoder-Struktur mit paralleler Datenverarbeitung.
Vorteile	- Einfache Struktur. - Geeignet für Aufgaben mit kurzen Sequenzen.	- Erfassen und Speichern von Langzeitabhängigkeiten. - Mildert das Vanishing-Gradient-Problem.	- Weniger Parameter als LSTMs, oft schnelleres Training. - Gute Balance zwischen Effizienz und Genauigkeit.	- Lernfähigkeit für Langzeitabhängigkeiten ohne Rekurrenz. - Hohe Parallelisierbarkeit beschleunigt das Training.
Nachteile	- Vanishing-Gradient-Problem, begrenzte Langzeitspeicherung.	- Höhere Rechenanforderungen als RNNs. - Längere Trainingszeiten.	- Kann in manchen Szenarien schlechter als LSTMs sein.	- Hohe Speicheranforderungen und Rechenlast. - Benötigt große Datenmengen für effizientes Training.

2.5 Einführung in LLM

LLMs sind leistungsfähige neuronale Netzwerke, die auf umfangreichen Textkorpora trainiert werden, um menschenähnliche Sprachverarbeitung zu ermöglichen. Die Einführung der Transformer-Architektur revolutionierte diesen Bereich, da sie durch Selbstaufmerksamkeit eine effiziente Modellierung von Langzeitabhängigkeiten erlaubt und frühere rekurrente Ansätze ersetzte [10].

Bekannte Vertreter sind GPT-3 [32], BERT [43] und T5 [21], die wesentliche Fortschritte in maschineller Übersetzung, Textklassifikation und Fragebeantwortung erzielt haben. Neben der Fähigkeit zur Textgenerierung zeichnen sie sich durch ihre hohe Präzision in NLP-Aufgaben aus.

Ein zentraler Aspekt der Leistungsfähigkeit von LLMs ist ihre Skalierbarkeit. Größere Modelle, die mit umfangreicheren Datensätzen trainiert wurden, erzielen eine höhere Genauigkeit. Darüber hinaus kann durch gezieltes Fine-Tuning auf domänenspezifische Daten die Modellleistung für spezifische Anwendungen weiter optimiert werden.

3 Datensatz, Modellauswahl und Modelloptimierung

Die Effektivität der automatisierten Vereinfachung englischer Verwaltungstexte hängt maßgeblich von der Qualität des Datensatzes, der Auswahl geeigneter Sprachmodelle und deren Optimierung ab. Dieses Kapitel beschreibt zunächst die Erstellung eines spezialisierten Textkorpus als Grundlage für das Modelltraining. Anschließend werden die verwendeten vortrainierten Modelle und deren technische Eigenschaften erläutert. Im weiteren Verlauf wird die Modellanpassung und Optimierung detailliert betrachtet, wobei der Fokus auf Hyperparametern, Trainingsmethoden und technischen Herausforderungen liegt. Abschließend werden die Ergebnisse der Feinabstimmung vorgestellt.

3.1 Entwicklung des spezialisierten Textkorpus

Die Grundlage dieser Arbeit bildet ein speziell erstellter Datensatz, der Verwaltungstexte aus gesellschaftlich relevanten Bereichen umfasst. Im Fokus stehen Dokumente der Bundesagentur für Arbeit sowie der Ausländerbehörden, da diese essenziellen Informationen für Bürgerinnen und Bürger bereitstellen und somit eine zentrale Rolle für die gesellschaftliche Teilhabe und das Verständnis administrativer Prozesse spielen.

3.1.1 Datenerhebung und Aufbereitung

Die Erstellung des Datensatzes erfolgte in einem mehrstufigen, systematischen Verfahren, das sowohl die sprachliche Verständlichkeit als auch die inhaltliche Präzision sicherstellen sollte. Zunächst wurden Verwaltungstexte in deutscher und englischer Sprache aus offiziellen Quellen, darunter Formulare der Bundesagentur für Arbeit und der Ausländerbehörden, extrahiert. Aufgrund der besseren Verarbeitung englischsprachiger Inhalte durch moderne vortrainierten Sprachmodelle wurden die deutschen Verwaltungstexte anschließend ins administrative Englisch übersetzt. Dabei wurde besonderes Augenmerk auf eine konsistente Terminologie sowie die Erhaltung des rechtlichen und formellen Charakters der Originaltexte gelegt, um die sprachliche Präzision und inhaltliche Integrität zu gewährleisten.

Dieser iterative Ansatz gewährleistet, dass die finalen Texte sowohl sprachlich verständlich als auch rechtlich präzise bleiben. Die Kombination aus automatisierter Vereinfachung mit ChatGPT-4o und manueller Nachbearbeitung ermöglicht eine gezielte Anpassung an die Prinzipien der Einfachen Sprache, während gleichzeitig der formelle und rechtliche Charakter der Verwaltungstexte erhalten bleibt. Durch diese methodische Vorgehensweise entsteht ein qualitativ hochwertiger Datensatz, der sich sowohl für die Vereinfachung administrativer Kommunikation als auch für das Fine-Tuning von LLMs eignet. Zudem schafft die gezielte Verbindung von Übersetzungs- und Vereinfachungsstrategien

eine solide Grundlage für zukünftige Anpassungen an weitere Sprachen, insbesondere Deutsch, um die Anwendbarkeit der entwickelten Modelle langfristig zu erweitern.

3.1.2 Skalierte Datensatzstrukturen

Der generierte Datensatz wurde systematisch in drei Größen skaliert. ein Minimaldatensatz mit 100 Einträgen, ein mittlerer Datensatz mit 500 Einträgen und ein vollständiger Datensatz mit 1000 Einträgen. Diese gestufte Skalierung dient dazu, die Auswirkungen der Datensatzgröße auf die Qualität der automatisierten Textvereinfachung zu analysieren und zu bewerten.

Der Datensatz ist in einer zweispaltigen Struktur organisiert. Die erste Spalte enthält die englische Verwaltungstexte, die den komplexen, formellen Stil der Verwaltungssprache nachbilden. Die zweite Spalte enthält die entsprechende vereinfachte englische Version, die speziell unter Berücksichtigung der Prinzipien der Einfachen Sprache erstellt wurde. Diese Struktur unterstützt die vortrainierten Sprachmodelle dabei, die Transformation von komplexen zu vereinfachten Texten effizient zu erlernen.

Durch diese Herangehensweise werden die Stärken vortrainierter Sprachmodelle optimal genutzt. Gleichzeitig wird der besondere Stil und die Komplexität der Verwaltungssprache in die Datensatzgestaltung einbezogen. Der erstellte Datensatz bietet somit eine solide Grundlage für die Entwicklung effektiver und anwendungsorientierter Strategien zur Textvereinfachung im Verwaltungskontext.

3.2 Verwendete vortrainierte Modelle

In dieser Arbeit werden drei state-of-the-art vortrainierte Modelle – LLama 3, Phi-3 mini und Mistral 7B – untersucht und hinsichtlich ihrer Eignung zur Vereinfachung von Verwaltungstexten evaluiert. Im Folgenden werden diese Modelle im Detail vorgestellt.

3.2.1 LLaMA 3 8B Instruct: Architektur und Leistungsfähigkeit

LLaMA 3 8B Instruct ist ein leistungsfähiges Open-Source-Sprachmodell, das von Meta AI [44] im Jahr 2024 veröffentlicht wurde. Es basiert auf einer decoderbasierten Transformer-Architektur und wurde gezielt für Aufgaben optimiert, die eine präzise Verarbeitung natürlicher Sprache erfordern. Das Modell ist speziell für Instruction-Following-Szenarien konzipiert, wodurch es besonders gut darin ist, komplexe Eingaben in verständliche und präzise Antworten zu überführen [44].Im Vergleich zu vorherigen Generationen bietet LLaMA 3 verbesserte Rechenleistung, Skalierbarkeit und Effizienz, was es ideal für die Vereinfachung von Verwaltungstexten macht.

Das Modell wurde auf einer umfassenden und qualitativ hochwertigen Datenbasis von über 15 Billionen Tokens trainiert. Die Trainingsdaten stammen aus CommonCrawl, wissenschaftlichen Publikationen, Lehrbüchern und technischen Dokumentationen. Die Datenvielfalt ermöglicht LLaMA 3 8B, komplexe Verwaltungstexte präzise zu analysieren und verständlich umzuwandeln, wodurch Behörden die Zugänglichkeit für Bürger mit geringer Sprachkompetenz oder kognitiven Einschränkungen verbessern.

Abbildung 7 veranschaulicht die Wettbewerbsfähigkeit von LLaMA 3 8B im direkten Vergleich mit anderen Modellen. In den Massive Multitask Language Understanding (MMLU) Benchmarks (5-shot) erreichte das Modell eine Punktzahl von 68,4 und übertraf damit leicht konkurrierende Modelle wie Mistral 7B und Gemma 7B.

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

Abbildung 7: Leistung des vortrainierten Modells [44]

3.2.2 Phi-3 Mini: Architektur und Leistungsfähigkeit

Die Phi-Modellfamilie, entwickelt von Microsoft Research [45], repräsentiert eine bedeutende Weiterentwicklung im Bereich ressourcenschonender und leistungsfähiger Sprachmodelle. Von Phi-1 über Phi-2 bis hin zu Phi-3 Mini zeigt sich eine konsequente Optimierung hin zu effizienteren und flexibleren KI-Modellen. Phi-3 Mini, mit nur 3,8 Milliarden Parametern, stellt eine kompakte, aber leistungsstarke Alternative zu größeren Sprachmodellen dar und eignet sich besonders für Anwendungen mit begrenzten Rechenressourcen [45].

Eine der herausragenden Eigenschaften von Phi-3 Mini ist seine hohe Effizienz relativ zu seiner Modellgröße. Abbildung 8 zeigt die Positionierung von Phi-3 Mini im Vergleich zu anderen Small Language Models (SLMs) und LLMs auf der MMLU-Benchmark. Hier wird deutlich, dass Phi-3 Mini

trotz seiner geringen Parameteranzahl eine bemerkenswerte Qualität erreicht und sich gegenüber anderen Modellen mit vergleichbarer Größe behauptet.

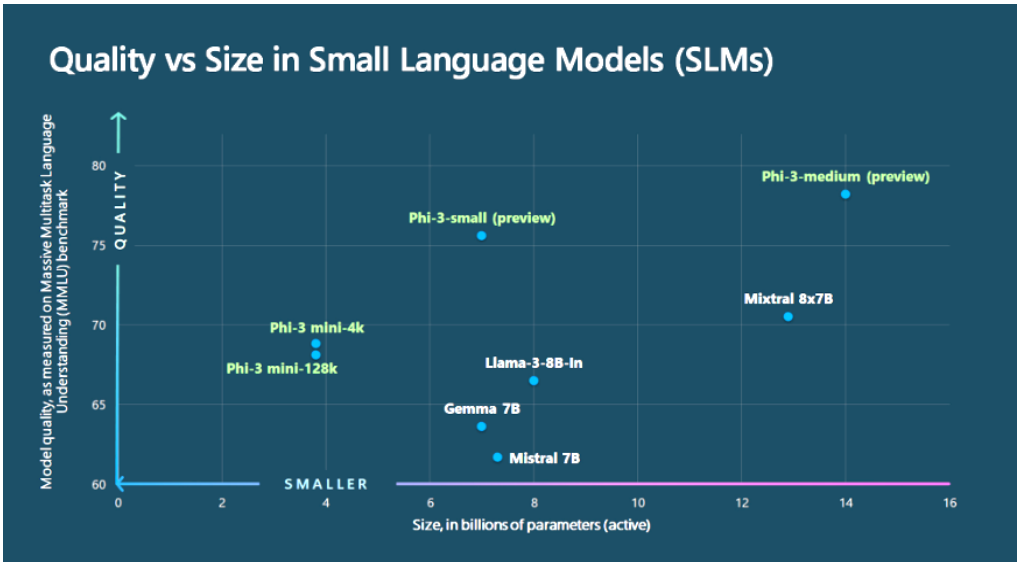

Abbildung 8: Vergleich der Modellqualität und Größe bei SLMs [45]

Das Modell wurde auf der Grundlage von "textbook-quality data" trainiert. Diese Daten umfassen kuratierte Inhalte aus Lehrbüchern, wissenschaftlichen Publikationen und Programmierbeispielen, wobei spezialisierte Filtertechniken angewendet wurden, um die Qualität der Trainingsdaten sicherzustellen. Microsoft Research setzte bewusst auf qualitativ hochwertige Datensätze, um eine solide Grundlage für die Textgenerierung zu schaffen und gleichzeitig die Effizienz der Sprachverarbeitung zu maximieren. Durch diese datengetriebene Optimierung erzielt Phi-3 Mini eine bemerkenswerte Textgenerierungsleistung, die sich durch Präzision und Konsistenz auszeichnet [45].

Die Benchmark-Tests bestätigen die herausragende Leistungsfähigkeit von Phi-3 Mini in verschiedenen sprachlichen Disziplinen. Abbildung 9 stellt die Ergebnisse von Phi-3 Mini im Vergleich zu anderen Modellen, darunter Llama-3 8B, Mistral 7B und Gemma 7B, dar. Die MMLU-Bewertung zeigt, dass Phi-3 Mini trotz seiner kleineren Größe beeindruckende Werte in Kategorien wie Sprachverständnis (HellaSwag, ARC-Challenge) erreicht.

Category	Benchmark	Phi-3				Gemma-7b	Mistral-7b	Mixtral-8x7b	Llama-3-8B-In	GPT3.5-Turbo-1106	Claude-3 Sonnet
		Phi-3-Mini-4K-In	Phi-3-Mini-128K-In	Phi-3-Small (Preview)	Phi-3-Medium (Preview)						
Popular Aggregate Benchmarks	AGI Eval (0-shot)	37.5	36.9	45	48.4	42.1	35.1	45.2	42	48.4	48.4
	MMLU (5-shot)	68.8	68.1	75.6	78.2	63.6	61.7	70.5	66.5	71.4	73.9
	BigBench Hard (0-shot)	71.7	71.5	74.9	81.3	59.6	57.3	69.7	51.5	68.3	--
Language Understanding	ANLI (7-shot)	52.8	52.8	55	58.7	48.7	47.1	55.2	57.3	58.1	68.6
	HellaSwag (5-shot)	76.7	74.5	78.7	83	49.8	58.5	70.4	71.1	78.8	79.2
Reasoning	ARC Challenge (10-shot)	84.9	84	90.7	91	78.3	78.6	87.3	82.8	87.4	91.6
	ARC Easy (10-shot)	94.6	95.2	97.1	97.8	91.4	90.6	95.6	93.4	96.3	97.7
	BoolQ (0-shot)	77.6	78.7	82.9	86.6	66	72.2	76.6	80.9	79.1	87.1
	CommonsenseQA (10-shot)	80.2	78	80.3	82.6	76.2	72.6	78.1	79	79.6	82.6
	MedQA (2-shot)	53.8	55.3	58.2	69.4	49.6	50	62.2	60.5	63.4	67.9
	OpenBookQA (10-shot)	83.2	80.6	88.4	87.2	78.6	79.8	85.8	82.6	86	90.8
	PIQA (5-shot)	84.2	83.6	87.8	87.7	78.1	77.7	86	75.7	86.6	87.8
	Social IQA (5-shot)	76.6	76.1	79	80.2	65.5	74.6	75.9	73.9	68.3	80.2
	TruthfulQA (MC2) (10-shot)	65	63.2	68.7	75.7	52.1	53	60.1	63.2	67.7	77.8
	WinoGrande (5-shot)	70.8	72.5	82.5	81.4	55.6	54.2	62	65	68.8	81.4
Factual Knowledge	TriviaQA (5-shot)	64	57.1	59.1	75.6	72.3	75.2	82.2	67.7	85.8	65.7
Math	GSM8K Chain of Thought (0-shot)	82.5	83.6	88.9	90.3	59.8	46.4	64.7	77.4	78.1	79.1
Code generation	HumanEval (0-shot)	59.1	57.9	59.1	55.5	34.1	28	37.8	60.4	62.2	65.9
	MBPP (3-shot)	53.8	62.5	71.4	74.5	51.5	50.8	60.2	67.7	77.8	79.4

Abbildung 9: Benchmark-Ergebnisse von Phi-3 Mini [46]

Phi-3 Mini optimiert die Vereinfachung von Verwaltungstexten, indem es komplexe Inhalte präzise und verständlich umformuliert. Seine kompakte Architektur erleichtert die Integration in bestehende IT-Infrastrukturen und ermöglicht effiziente Textverarbeitung, insbesondere in ressourcenbeschränkten behördlichen Anwendungen. Durch seine Kombination aus Effizienz, Flexibilität und qualitativ hochwertigen Trainingsdaten trägt Phi-3 Mini zur barrierefreien Kommunikation und verbesserten Informationszugänglichkeit bei.

3.2.3 Mistral 7B Instruct: Architektur und Leistungsfähigkeit

Mistral 7B Instruct ist ein leistungsstarkes Open-Source-Sprachmodell, das von Mistral AI [47] entwickelt und unter dem Apachen 2.0-Lizenz veröffentlicht wurde. Mit 7,3 Milliarden Parametern bietet es eine bemerkenswerte Leistungsfähigkeit, die in vielen Bereichen mit wesentlich größeren Modellen wie LLaMA 2 13B und LLaMA 1 34B konkurrieren kann [47]. Dank seiner optimierten Architektur, die moderne Techniken wie Grouped-Query Attention (GQA) und Sliding Window Attention (SWA) nutzt, erzielt das Modell eine hohe Verarbeitungseffizienz und ermöglicht die Bearbeitung längerer Sequenzen bei reduziertem Rechenaufwand [47].

Das Modell wurde mit einem breit gefächerten hochwertigen Datensatz trainiert, der Inhalte aus wissenschaftlichen Publikationen, technischen Dokumentationen und offenen Textquellen umfasst. Dabei wurde ein besonderer Fokus auf die Optimierung der Sprachverarbeitung gelegt, um die Modellleistung für administrative und juristische Anwendungen zu maximieren. Durch Fine-Tuning für Instruct-Aufgaben wurde Mistral 7B speziell darauf ausgerichtet, Präzision und Klarheit in der Textgenerierung zu gewährleisten [47]. Diese Fähigkeit ist besonders relevant für die Verwaltungstextvereinfachung, da es darauf ankommt, juristische Genauigkeit zu bewahren, während der Text in eine verständliche und zugängliche Form umgewandelt wird.

In Benchmark-Tests zeigt sich die überlegene Leistung von Mistral 7B Instruct im Vergleich zu ähnlich großen Modellen. Besonders in logischen und reasoning-basierten Aufgaben erreicht das Modell eine Effizienz, die mit einem LLaMA 2-Modell mit dreifacher Parametergröße vergleichbar ist. Dies bedeutet, dass Mistral 7B mit deutlich geringeren Rechenressourcen Ergebnisse auf dem Niveau größerer Modelle liefert. Eine Analyse der MMLU-, Knowledge-, Reasoning- und Comprehension-Benchmarks zeigt, dass Mistral 7B in vielen Kategorien mit LLaMA 2 13B und LLaMA 1 34B konkurrieren kann (siehe Abbildung 10).

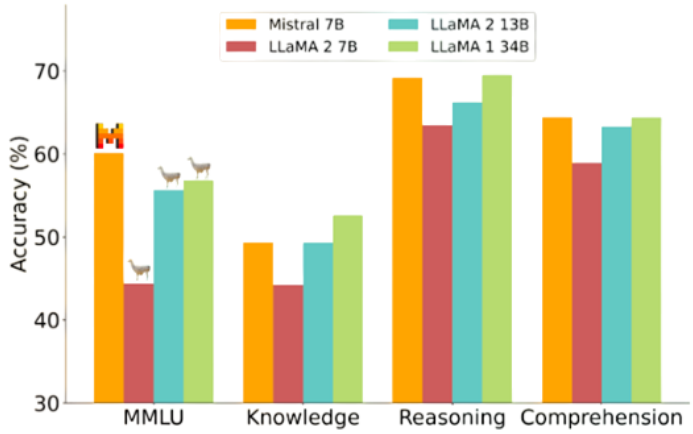

Abbildung 10: Mistral 7B im Benchmark-Vergleich mit LLaMA-Modellen [46]

Darüber hinaus zeigt eine detaillierte Bewertung der MT Bench-Ergebnisse, dass Mistral 7B Instruct alle anderen 7B-Modelle übertrifft und sich mit Modellen der 13B-Klasse messen kann. Besonders hervorzuheben ist die höhere Punktzahl im Vergleich zu LLaMA-2-7B-chat sowie Vicuna-7B-16k, was seine Stärke in der präzisen und verständlichen Sprachverarbeitung unterstreicht (siehe Abbildung 11). Diese herausragende Leistungsfähigkeit unterstreicht die Fähigkeit des Modells, komplexe sprachliche Aufgaben zu bewältigen, einschließlich der Textvereinfachung, bei der Kohärenz, Verständlichkeit und Präzision entscheidend sind.

Model	MT Bench
WizardLM-13b-v1.2	7.2
Vicuna-13B-16k	6.92
Mistral 7B Instruct	**6.84 ± 0.065**
WizardLM-13B-v1.1	6.76
Llama-2-13b-chat	6.65
Llama-2-7b-chat	6.27
Vicuna-7B-16k	6.22
Alpaca-13B	4.53

Abbildung 11: Mistral 7B Instruct – MT Bench-Ergebnisse [47]

Mistral 7B ist flexibel anpassbar und ermöglicht gezieltes Fine-Tuning für die Verwaltungstextvereinfachung, wodurch Fachjargon reduziert und komplexe Satzstrukturen vereinfacht werden.. Durch seine kompakte Modellgröße und effiziente Architektur eignet sich Mistral 7B besonders für Einsätze in Verwaltungen, Bildungseinrichtungen und öffentlichen Institutionen, wo es problemlos in bestehende IT-Systeme integriert werden kann [47].

3.3 Modellanpassung und Optimierung

Im Rahmen der systematischen Fine-Tuning-Experimente werden drei ausgewählte Sprachmodelle durch die gezielte Variation mehrerer Hyperparameter optimiert. Die methodische Evaluierung umfasst die schrittweise Anpassung der Größe des Trainingsdatensatzes sowie die Untersuchung verschiedener maximale Trainingsschritte (max-steps). Darüber hinaus werden unterschiedliche Lernraten getestet, während gleichzeitig verschiedene Prompt-Strukturen evaluiert werden. Diese strukturierte Herangehensweise ermöglicht die Identifikation der optimalen Parameterkombination für jedes der drei Modelle und gewährleistet somit die bestmögliche Performanz bei der Vereinfachung von Verwaltungstexten.

Aufgrund begrenzter GPU-Ressourcen und Speicherkapazitäten wurde auf eine automatisierte Hyperparameter-Optimierung mittels Grid Search oder Bayesian Optimization verzichtet. Über die untersuchten Parameter hinaus existieren jedoch weitere bedeutsame Optimierungsmöglichkeiten für die Textvereinfachungsaufgabe, die im Rahmen dieser Arbeit nicht näher betrachtet werden. So könnte die Implementierung verschiedener Loss-Funktionen, beispielsweise durch die Kombination von Cross-Entropy-Loss mit zusätzlichen Metriken zur Messung lexikalischer Komplexität, die Qualität der Vereinfachung weiter verbessern.

Darüber hinaus könnten Early-Stopping-Strategien auf Basis von Simplizitätsmetriken sowie die Integration von Readability-Scores in den Trainingsprozess die Modellleistung weiter steigern. Auch Ansätze wie die Implementierung von Warmup-Phasen oder das schrittweise Entfrieren von

Modellschichten (gradual unfreezing) bieten weiteres Optimierungspotenzial. Eine detaillierte Untersuchung dieser Parameter würde jedoch den definierten Rahmen dieser Bachelorarbeit überschreiten.

3.3.1 Technische Umsetzung und Trainingsumgebung

Der zugrunde liegende Code ist für alle drei vortrainierten Sprachmodelle einheitlich strukturiert, wobei Unterschiede nur bei der Auswahl des Modells und den genannten Parametern bestehen.

Der gesamte Code wurde in der Google-Colab-Umgebung ausgeführt, die sich aufgrund ihrer GPU-Unterstützung und der einfachen Integration externer Ressourcen wie Google Drive ideal für das Training großer Sprachmodelle eignet. Der Quellcode wurde zudem in einem Gitlab-Repository gespeichert, um die Nachvollziehbarkeit und Wiederverwendbarkeit der Implementierung zu gewährleisten.

3.3.2 Code-Struktur und Einsatz von Unsloth

Die Implementierung basiert auf der spezialisierten Bibliothek Unsloth, die für die effiziente Nutzung und Anpassung großer Sprachmodelle entwickelt wurde. Diese Bibliothek zeichnet sich durch eine benutzerfreundliche Schnittstelle aus, die das Laden, Trainieren und Feinjustieren von Modellen vereinfacht. Insbesondere für komplexe Modelle mit Milliarden Parametern ist Unsloth aufgrund seiner Ressourcenoptimierung hervorragend geeignet.

Ein wesentliches Merkmal von Unsloth ist die Unterstützung des 4-Bit-Formats für die Modellverarbeitung, das den Speicherbedarf drastisch reduziert. Diese Funktion ist von entscheidender Bedeutung, um große Modelle auf GPUs mit begrenzten Ressourcen auszuführen, ohne dass die Modellleistung beeinträchtigt wird. Darüber hinaus integriert Unsloth moderne Techniken wie PEFT, einschließlich Low-Rank Adaptation (LoRA) [48]. Diese Techniken optimieren Modelle für spezifische Aufgaben wie die Verwaltungstextvereinfachung, minimieren den Rechenaufwand und machen Unsloth zu einem essenziellen Werkzeug für die automatisierte Textvereinfachung.

3.3.3 Bibliotheken und Datensatzverarbeitung

Die Implementierung nutzt eine Vielzahl bewährter Python-Bibliotheken, um die Datenverarbeitung und Modellanpassung zu unterstützen. Die Bibliothek *Transformers* wird für die Bereitstellung und Feinabstimmung vortrainierter Sprachmodelle verwendet. *Datasets* dient zur effizienten Verwaltung und Organisation der Trainingsdaten, während *Pandas* für die Datenanalyse und Vorverarbeitung eingesetzt wird. Der Datensatz wird direkt aus Google Drive geladen, was die Handhabung großer

Datenmengen erleichtert und eine zentrale Datenquelle bietet. Nach dem Importieren in ein pandas-DataFrame wird der Datensatz in ein Hugging-Face-kompatibles Format umgewandelt, um eine nahtlose Integration in das Trainingsframework zu gewährleisten.

Während der Datenvorbereitung werden die Verwaltungstexte systematisch standardisiert und für das Training optimiert. Jeder Eintrag im Datensatz wird in ein Eingabe-Ausgabe-Paar umgewandelt, bei dem die Eingabe den Originaltext und die Ausgabe die vereinfachte Version des Textes darstellt. Die strukturierte Datenvorbereitung und die Verwendung modernster Bibliotheken gewährleisten eine präzise und effiziente Umsetzung der Aufgabenstellung.

3.3.4 Hyperparameter und Prompts

Die Optimierung der Modellleistung wurde durch eine systematische Untersuchung zentraler Hyperparameter sowie durch gezielte Variation der Eingabeformulierung durchgeführt. Die Lernrate wurde in den Werten 3e-5 und 5e-5 getestet, um die besten Bedingungen für das Training zu identifizieren. Gleichzeitig wurden die maximalen Trainingsschritte (max steps) in den Intervallen 25, 50 und 100 angepasst, um den Einfluss unterschiedlicher Trainingslängen auf die Modellleistung zu bewerten. Der Umfang des Datensatzes wurde ebenfalls variiert, mit Größen von 100, 500 und 1000 Beispielen, um zu analysieren, wie der Datenumfang die Ergebnisse beeinflusst.

Zusätzlich wurden drei unterschiedliche Prompts verwendet, um die Auswirkungen der Eingabeformulierung auf die Modellleistung zu untersuchen:

a) **Einfacher Prompt:**
   Simplify text: {input_text}
b) **Präziser, verlustfreier Prompt:**
   Simplify the following text without any information lost: {input_text}
c) **Ausführlicher Experten-Prompt:**
   You are an expert in text simplification, specializing in transforming complex administrative language into Easy Language in English. Your goal is to rewrite the text so it is easy to understand for everyone, including people with limited language skills, cognitive impairments, or those who are not native speakers. Follow these guidelines for the best results: - Use simple and clear words: Choose everyday words that most people know. Avoid technical terms, jargon, or difficult expressions. - Keep sentences short and direct: Use one idea per sentence to avoid confusion. Break complex thoughts into multiple sentences or step-by-step instructions. - Explain concepts clearly: If a term or idea is necessary but difficult, provide a simple explanation or example. Assume the reader has no prior knowledge of the topic. - Avoid figurative language: Do not use idioms, metaphors, or abstract expressions that might be

confusing. Use straightforward language instead. - Structure the information clearly: Use lists or step-by-step formatting when appropriate to make the text easy to read and scan. - Be respectful and encouraging: Maintain a neutral and supportive tone. Make the text helpful without being condescending. Avoid creating a sense of urgency or pressure. - Simplify grammar and punctuation: Use basic punctuation like commas and periods. Avoid complex punctuation like semicolons. Write numbers as words for small numbers and as figures for larger ones. - Focus on active voice: Use direct, active sentences rather than passive voice. - Repeat key points if necessary: Repetition can help reinforce important information. The goal is to make the text as clear, concise, and easy to understand as possible, while keeping all essential information. Simplify the following text: {input_text}.

Die Kombination aus variierenden Hyperparametern und differenzierten Prompt-Formulierungen wurde gezielt eingesetzt, um die optimalen Bedingungen für eine effektive Textvereinfachung zu identifizieren. Die Experimente zeigten, dass sowohl Modellparameter als auch Eingabeformulierung die Qualität der Vereinfachung maßgeblich beeinflussen.

3.3.5 Einsatz von PEFT-Techniken mit LoRA

Das Fine-Tuning der Modelle basiert auf PEFT. Innerhalb des PEFT-Frameworks wird LoRA verwendet, eine Methode, die es ermöglicht, lediglich ausgewählte Parametergruppen wie die Projektionsmatrizen zu trainieren, während die übrigen Modellparameter eingefroren bleiben. Diese gezielte Anpassung reduziert die Rechenkosten erheblich und bewahrt gleichzeitig die hohe Leistung der vortrainierten Modelle.

LoRA ist besonders vorteilhaft, wenn nur begrenzte Trainingsdaten zur Verfügung stehen, wie es bei der Erstellung von spezialisierten Datensätzen für die Textvereinfachung oft der Fall ist. Die Implementierung im vorliegenden Projekt nutzt eine Dropout-Rate von 0 und einen Alpha-Wert von 32, was eine effiziente Balance zwischen der Modellanpassungsfähigkeit und der Stabilität der Trainingsergebnisse gewährleistet. Diese Parameter wurden sorgfältig gewählt, um sicherzustellen, dass das Modell trotz der Ressourcenschonung domänenspezifische Anforderungen präzise erfüllen kann.

Die Integration von LoRA innerhalb der Unsloth-Bibliothek und ihrer Unterstützung für speichereffizientes Laden im 4-Bit-Format ist ein wesentlicher Bestandteil des Codes. Dies erlaubt es, das Training selbst auf Geräten mit begrenztem GPU-Speicher durchzuführen, während die leistungsstarken Eigenschaften der Modelle erhalten bleiben.

3.3.6 Trainingssetup und Modellbewertung

Das Trainingssetup wurde sorgfältig konzipiert, um optimale Bedingungen für die Feinabstimmung vortrainierter Sprachmodelle sicherzustellen. Im Mittelpunkt standen die systematische Variation zentraler Hyperparameter wie Lernrate, Trainingsschritte, Datensatzgröße und die Formulierung der Prompts. Ziel war es, den Einfluss dieser Faktoren auf die Modellleistung umfassend zu analysieren. Für die Implementierung kam die leistungsfähige *transformers*-Bibliothek in Kombination mit der *SFTTrainer*-Klasse zum Einsatz, die speziell für die effiziente Anpassung großer Sprachmodelle entwickelt wurde. Um eine vergleichbare Bewertung der Experimente sicherzustellen, blieben Parameter wie Batch-Größe, Warmup-Schritte und Optimierungsalgorithmus über alle Tests hinweg konstant.

Das Training basierte auf einem spezialisierten Datensatz mit vereinfachten Verwaltungstexten, wobei jede Modelliteration mit drei unterschiedlichen Prompt-Varianten durchgeführt wurde. Diese dienten dazu, die Reaktion des Modells auf unterschiedliche Eingabequalitäten zu evaluieren und die bestmögliche Strategie für die Textvereinfachung zu identifizieren.

Während des Trainings wurde der Training Loss kontinuierlich überwacht, um die Konvergenz des Modells und dessen Fähigkeit zur präzisen Vorhersage der Zielausgabe zu bewerten. Dieser Wert lieferte entscheidende Hinweise darauf, wie gut das Modell die angestrebte Vereinfachung umsetzen konnte.

3.3.7 Technische Herausforderungen und Lösungen

Während der Implementierung und Feinabstimmung der Sprachmodelle traten mehrere technische Herausforderungen auf, die gezielt adressiert wurden, um eine effiziente Modellanpassung sicherzustellen.

Eine zentrale Schwierigkeit bestand in der begrenzten GPU-Verfügbarkeit in Google Colab, die zu Speicherengpässen führte und die maximale Modell- und Batchgröße einschränkte. Zur Lösung dieses Problems wurde die 4-Bit-Quantisierung mithilfe der Unsloth-Bibliothek implementiert, wodurch der Speicherverbrauch erheblich reduziert, aber die Modellleistung weitgehend erhalten blieb. Dies ermöglichte ein effizientes Fine-Tuning trotz eingeschränkter Hardware-Ressourcen.

Ein weiteres Problem war die Modellanpassung bei begrenzter Datenmenge, da ein vollständiges Training aus Rechen- und Datensicht nicht praktikabel war. Um dies zu lösen, wurde LoRA (Low-Rank Adaptation) eingesetzt, das eine selektive Anpassung spezifischer Gewichtsmatrizen innerhalb der Selbstaufmerksamkeitsmodule erlaubt, anstatt das gesamte Modell zu aktualisieren. Dies optimierte

nicht nur die Recheneffizienz, sondern bewahrte auch bereits erlernte Sprachstrukturen, wodurch der Trainingsprozess stabiler wurde.

Eine weitere Herausforderung war die begrenzte Sitzungsdauer in Google Colab, die zu Verbindungsabbrüchen während längerer Trainingszyklen führte. Um dies zu umgehen, wurde ein Google Colab Pro-Abonnement genutzt, das eine verlängerte Sitzungsdauer und priorisierten GPU-Zugriff ermöglichte. Die Wahl fiel auf die NVIDIA T4-GPU, da sie ein optimales Verhältnis zwischen Leistung und Speicherverbrauch bietet.

Die Verwaltung großer Textdatenmengen stellte ebenfalls eine Herausforderung dar. Die Integration von Google Drive ermöglichte eine effiziente Speicherung und direkte Verfügbarkeit der Trainingsdaten, wodurch der Workflow erheblich verbessert wurde.

Durch die Kombination dieser technischen und methodischen Lösungen konnte das Training trotz begrenzter Ressourcen erfolgreich durchgeführt und die Effizienz sowie Skalierbarkeit der Implementierung optimiert werden.

3.3.8 Designentscheidungen

Die Entscheidung, sich auf die Vereinfachung englischer Verwaltungstexte zu konzentrieren, basiert auf der Tatsache, dass die verwendeten vortrainierten Modelle überwiegend auf umfangreichen englischsprachigen Korpora trainiert wurden. Dadurch können sie in englischer Sprache eine höhere Präzision und Leistung erzielen.

Die Wahl der Bibliotheken und Frameworks orientierte sich an der Effizienz und Skalierbarkeit der Implementierung. Die Unsloth-Bibliothek wurde aufgrund ihrer nativen Unterstützung für LoRA und die 4-Bit-Quantisierung priorisiert. Diese Kombination minimierte den Speicherverbrauch, während das Fine-Tuning trotz begrenzter Hardware-Ressourcen ermöglicht wurde.

Bei der Modellauswahl spielten mehrere Faktoren eine Rolle. LLaMA 3 7B, Phi-3 mini und Mistral 7B wurden aufgrund ihrer bekannten Leistungsfähigkeit in vergleichbaren Szenarien gewählt. Die Architektur dieser Modelle bietet eine optimale Balance zwischen Größe und Anpassungsfähigkeit, wodurch sie sowohl für den Einsatz in ressourcenbeschränkten Umgebungen als auch für spezialisierte Fine-Tuning-Aufgaben geeignet sind.

Ein zentraler Bestandteil des Designs war die Entwicklung und Strukturierung der Prompts, die die Eingaben für die Modelle steuern. Diese wurden gezielt auf die Zielgruppe und die Anforderungen der Textvereinfachung abgestimmt. Neben einem einfachen Prompt, der lediglich eine grundlegende Vereinfachung forderte, wurde ein präziser, verlustfreier Prompt entwickelt, der sicherstellte, dass keine Informationen verloren gingen. Zusätzlich wurde ein ausführlicher Experten-Prompt konzipiert, der

detaillierte Anweisungen und Leitlinien enthielt, um die Modelle gezielt auf die Übersetzung komplexer administrativer Sprache in leicht verständliches Plain English auszurichten. Diese differenzierte Herangehensweise ermöglichte nicht nur eine gezielte Analyse der Auswirkungen verschiedener Eingabequalitäten, sondern trug auch maßgeblich zur Optimierung der Modellleistung bei.

3.4 Ergebnisse der Feinabstimmung

In diesem Abschnitt werden die Ergebnisse der Experimente für jedes Modell detailliert dargestellt. Für jedes der untersuchten Modelle wurde eine optimale Kombination von Hyperparametern identifiziert, die die beste Leistung bei der Vereinfachung von Verwaltungstexten erzielte. Die Analyse umfasst die Auswirkungen von Lernrate, Trainingsschritten, Datensatzgröße und der Gestaltung von Prompts, um die jeweiligen Stärken und Schwächen der Modelle hervorzuheben und ihre Leistungsfähigkeit zu maximieren.

3.4.1 LLaMA 3 8B Instruct

Die Feinabstimmung von LLaMA 3 8B Instruct identifizierte eine optimale Kombination von Hyperparametern, die eine besonders gute Leistung bei der Vereinfachung komplexer Verwaltungstexte ermöglichte. Die Experimente zeigten, dass eine mittlere Datensatzgröße von 500 Instanzen die beste Balance zwischen Trainingsstabilität und Generalisierungsfähigkeit bot. Mit 100 Trainingsschritten und einer Lernrate von 5×10^{-5} wurde ein Final Loss von 0,4231 erreicht, während eine alternative Lernrate von 3×10^{-5} höhere Loss-Werte von 0,8288 erzielte, jedoch weiterhin stabile Ergebnisse lieferte.

Interessanterweise führte eine Erhöhung des Trainingsdatensatzes auf 1000 Instanzen nicht zwangsläufig zu besseren Ergebnissen. In einigen Fällen lagen die Loss-Werte zwischen 1,0331 und 1,6159, was darauf hindeutet, dass größere Datensätze ab einem bestimmten Punkt keine signifikanten Leistungssteigerungen mehr bringen und möglicherweise zu einer Erhöhung der Modellkomplexität ohne zusätzlichen Nutzen führen. Gleichzeitig zeigten kleinere Datensätze mit nur 100 Instanzen eine unzureichende Generalisierung, da die Loss-Werte selbst nach 100 Trainingsschritten über 16,7485 blieben. Dies unterstreicht die Bedeutung einer ausreichend großen, aber nicht überdimensionierten Datenbasis für eine effektive Modellanpassung.

Die Anzahl der Trainingsschritte hatte einen direkten Einfluss auf die Modellleistung. Während 25 oder 50 Schritte akzeptable Ergebnisse lieferten, zeigte sich, dass 100 Trainingsschritte insbesondere bei mittleren und großen Datensätzen eine deutliche Verbesserung bewirkten. Bei 1000 Instanzen flachte der Trainingseffekt ab, während bei 500 Instanzen mit 100 Trainingsschritten der niedrigste Loss-Wert

(0,4231) erreicht wurde, was auf eine optimale Balance zwischen Datenmenge und Trainingsdauer hindeutet.

Auch die Lernrate spielte eine entscheidende Rolle. Eine Lernrate von 5×10^{-5} führte durchweg zu den besten Ergebnissen, insbesondere bei 500 Instanzen und 100 Trainingsschritten. Eine konservativere Lernrate von 3×10^{-5} war zwar stabil, erzielte jedoch höhere Loss-Werte. Größere Datensätze (1000 Instanzen) profitierten tendenziell von einer leicht höheren Lernrate, während kleinere Datensätze mit einer moderateren Optimierung stabilere Resultate lieferten.

Ein zentraler Aspekt der Experimente war der Einsatz verschiedener Prompts, die das Modell zur Vereinfachung der Verwaltungssprache anleiteten. Die drei getesteten Prompt-Varianten unterschieden sich in Detailgrad, Struktur und sprachlicher Ausrichtung. Die besten Ergebnisse wurden mit klaren, präzisen Anweisungen erzielt. Besonders wirkungsvoll waren Prompts, die einfache Worte, kurze Sätze und die Vermeidung von Fachjargon betonten. Ein weiterer wichtiger Aspekt war der Umgang mit komplexen Begriffen. Während einige Prompts das Modell dazu veranlassten, schwierige Begriffe zu eliminieren, führten gezielte Anweisungen dazu, dass diese verständlich erklärt wurden. Dies verbesserte die Zugänglichkeit, ohne dass wesentliche Inhalte verloren gingen.

3.4.2 Phi-3 mini

Die Feinabstimmung von Phi-3 Mini zeigte, dass eine Lernrate von 5×10^{-5}, 50 Trainingsschritte und eine Datensatzgröße von 1000 Instanzen die besten Ergebnisse lieferten. Diese Konfiguration erreichte einen Final Loss von 0,9385, was darauf hindeutet, dass das Modell eine präzise und stabile Textvereinfachung durchführen konnte.

Die Experimente bestätigten die bedeutende Rolle der Datensatzgröße für die Modellleistung. Größere Datensätze mit 1000 Instanzen führten durchweg zu besseren Loss-Werten als kleinere. Dies zeigt, dass Phi-3 Mini stärker auf eine größere Anzahl an Trainingsbeispielen angewiesen ist, um eine robuste Generalisierung zu erreichen. Allerdings zeigte sich, dass eine Erhöhung der Trainingsschritte von 50 auf 100 bei 1000 Instanzen nicht zu besseren Ergebnissen führte. Stattdessen stieg der Loss-Wert auf 1,6441, was darauf hindeutet, dass das Modell bei längerer Trainingsdauer zu stark an die Trainingsdaten angepasst wurde und damit weniger flexibel auf neue Eingaben reagierte.

Bei 500 Instanzen war ein ähnliches Muster zu beobachten. Während 50 Trainingsschritte (Loss: 1,1142) solide Ergebnisse lieferten, führte eine Erhöhung auf 100 Schritte (Loss: 1,4698) nicht zu einer weiteren Verbesserung. Dies unterstreicht, dass eine angemessene Trainingsdauer gefunden werden muss, um Overfitting oder unnötige Rechenzeit zu vermeiden. Kleine Datensätze mit nur 100 Instanzen zeigten hingegen stark erhöhte Loss-Werte von über 12, selbst bei 100 Trainingsschritten. Dies bestätigt,

dass das Modell mit einer so begrenzten Datenbasis nicht effektiv lernen und generalisieren konnte, wodurch die Qualität der Vereinfachungen stark beeinträchtigt wurde.

Die Wahl der optimalen Lernrate variierte je nach Datensatzgröße. Für große Datensätze (1000 Instanzen) erwies sich eine Lernrate von 5×10^{-5} als besonders effektiv, da sie das Modell in der Lage versetzte, sich effizient zu optimieren, ohne dabei zu instabilen Trainingsdynamiken zu führen. Bei mittleren Datensätzen (500 Instanzen) war eine Lernrate von 3×10^{-5} mit 50 Schritten (Loss: 1,2494) eine stabile Alternative. Eine zu hohe Lernrate (5×10^{-5}) führte hier mit 100 Schritten zu suboptimalen Ergebnissen (Loss: 1,4698), was darauf hindeutet, dass kleinere Datensätze empfindlicher auf aggressive Lernraten reagieren und mit moderateren Werten stabilere Ergebnisse liefern.

Neben den Hyperparametern hatte auch die Struktur der Prompts einen erheblichen Einfluss auf die Qualität der erzeugten Vereinfachungen. Prompt c, der detaillierte Anweisungen zur Vermeidung von Fachjargon und zur Nutzung kurzer, prägnanter Sätze enthielt, führte zu den besten Ergebnissen.

3.4.3 Mistral 7B Instruct

Die Feinabstimmung von Mistral 7B Instruct zeigte, dass eine Lernrate von 5×10^{-5}, kombiniert mit 100 Trainingsschritten und einer Datensatzgröße von 500 Instanzen, die beste Modellleistung erzielte. Diese Konfiguration erreichte einen Final Loss von 0,0982, was auf eine stabile und effiziente Modellanpassung hinweist.

Die Experimente bestätigten, dass eine mittlere Datensatzgröße (500 Instanzen) die beste Balance zwischen Datenvielfalt und Trainingsstabilität bot. Während eine Erhöhung auf 1000 Instanzen mit derselben Lernrate und 100 Trainingsschritten einen Final Loss von 0,3615 erreichte, lag dieser dennoch über dem Wert für 500 Instanzen. Dies deutet darauf hin, dass zusätzliche Trainingsdaten nicht zwangsläufig zu besseren Ergebnissen führen, sondern ab einem bestimmten Punkt die Konvergenz verlangsamen oder zu geringfügigem Overfitting führen können.

Kleinere Datensätze (100 Instanzen) erwiesen sich als unzureichend, um eine effektive Generalisierung zu gewährleisten. Selbst bei 100 Trainingsschritten lagen die Loss-Werte über 9,6191, was verdeutlicht, dass das Modell nicht genügend Trainingsbeispiele hatte, um Verwaltungsstrukturen zuverlässig zu vereinfachen. Diese Ergebnisse unterstreichen, dass eine Mindestmenge an Daten erforderlich ist, um stabile und kohärente Vereinfachungen zu erzeugen.

Die Anzahl der Trainingsschritte spielte ebenfalls eine wesentliche Rolle. Während 50 Trainingsschritte mit 500 Instanzen bereits akzeptable Ergebnisse erzielten (Loss: 0,7082), führte eine Erhöhung auf 100 Schritte zu einer deutlichen Verbesserung (Loss: 0,0982). Dies zeigt, dass eine längere Trainingsdauer

das Modell weiter optimierte, ohne Overfitting zu verursachen. Bei 1000 Instanzen flachte der Effekt zusätzlicher Trainingsschritte jedoch ab, sodass die Verbesserung nicht mehr so signifikant war.

Die Wahl der Lernrate beeinflusste die Trainingsstabilität erheblich. Eine Lernrate von 5×10^{-5} erwies sich als optimal, da sie eine schnelle Konvergenz ermöglichte, ohne zu Instabilitäten zu führen. Eine konservativere Lernrate von 3×10^{-5} zeigte eine langsamere, aber stabilere Lernkurve, was sich besonders bei kleineren Datensätzen als vorteilhaft erwies. Allerdings führte diese reduzierte Lernrate bei größeren Datensätzen zu einer ineffizienteren Optimierung, sodass das Modell mehr Trainingsschritte benötigt hätte, um vergleichbare Ergebnisse zu erzielen.

Die Experimente bestätigten, dass auch bei Mistral 7B die Qualität der Vereinfachung stark von der Gestaltung der Prompts abhing. Die besten Ergebnisse wurden mit klar strukturierten und detaillierten Prompts erzielt, insbesondere mit Prompt c, der präzise Anweisungen zur Vereinfachung komplexer Sätze enthielt. Einfachere Prompts, die weniger Vorgaben zur Struktur und Wortwahl machten, führten zu weniger kohärenten und teilweise redundanten Vereinfachungen. Besonders problematisch war dies bei administrativen Fachbegriffen, da ungenaue Prompts oft zu übermäßigen Vereinfachungen führten, wodurch rechtliche Nuancen verloren gingen. Dies verdeutlicht, dass ein strukturiertes Prompt-Design eine Schlüsselrolle spielt, um die Effektivität von Mistral 7B für die Vereinfachung von Verwaltungstexten zu maximieren.

3.4.4 Zusammenfassung der Feinabstimmungsergebnisse

Die folgende Tabelle 2 gibt einen Überblick über die optimalen Hyperparameter-Kombinationen der drei getesteten Modelle. Diese Konfigurationen wurden iterativ durch umfangreiche Experimente ermittelt, um die bestmögliche Leistung in der Vereinfachung administrativer Texte zu erreichen.

Tabelle 2: Optimale Parameterkonfigurationen der Modelle

Modell	Datensatzgröße	Max-Step	Learning Rate	Prompt	Final Loss
Llama-3	500	100	5e-5	c	0,4231
Phi	1000	50	5e-5	c	0,9385
Mistral	500	100	5e-5	c	0,0982

Die experimentellen Ergebnisse zeigen, dass eine mittelgroße Datensatzgröße von 500 Instanzen in Kombination mit 100 Trainingsschritten für die meisten Modelle die beste Balance zwischen Effizienz

und Generalisierungsfähigkeit bot. Während Phi-3 Mini mit 1000 Instanzen eine stabile Leistung erreichte, führte eine weitere Erhöhung der Datenmenge nicht zwangsläufig zu einer proportionalen Verbesserung. Kleinere Datensätze (100 Instanzen) erwiesen sich als unzureichend, da sie zu stark erhöhten Loss-Werten führten und eine fehlende Generalisierungsfähigkeit aufzeigten.

Ein entscheidender Faktor für die Modellleistung war die Anzahl der Trainingsschritte. 100 Trainingsschritte erwiesen sich als optimale Wahl, insbesondere für Mistral 7B, das mit dieser Konfiguration einen Final Loss von 0,0982 erzielte. Phi-3 Mini hingegen profitierte von nur 50 Trainingsschritten, da längere Trainingsläufe keinen zusätzlichen Mehrwert boten und möglicherweise sogar die Loss-Werte leicht verschlechterten. Dies zeigt, dass die optimale Trainingsdauer modellabhängig ist.

Die Wahl der Lernrate spielte ebenfalls eine zentrale Rolle. Eine Lernrate von 5×10^{-5} erwies sich durchgängig als die beste Wahl, insbesondere für größere Datensätze und längere Trainingszeiten. Kleinere Lernraten (3×10^{-5}) führten zwar zu stabileren Trainingsverläufen, erwiesen sich jedoch als weniger effizient, da sie die Konvergenz verlangsamten. Dies war insbesondere bei LLaMA 3 und Mistral 7B zu beobachten, bei denen die höhere Lernrate schnellere und präzisere Ergebnisse lieferte.

Ein weiterer Schlüsselfaktor war die Gestaltung der Prompts. Strukturierte Eingabeaufforderungen mit klaren Vorgaben zur Vermeidung von Fachjargon, Satzlänge und Wortwahl führten bei allen Modellen zu signifikanten Verbesserungen. Besonders Prompt c, der detaillierte Anweisungen zur Verständlichkeit und sprachlichen Klarheit enthielt, zeigte die besten Ergebnisse. Weniger strukturierte Prompts erschwerten es den Modellen, kohärente und verständliche Vereinfachungen zu erzeugen, da detaillierte Vorgaben für Satzbau und Terminologie fehlten.

4 Ergebnisse und Diskussion

In diesem Kapitel werden die Ergebnisse dieser Arbeit vorgestellt und bewertet. Der Fokus liegt auf der Analyse der Leistung der eingesetzten Modelle bei der Vereinfachung von Verwaltungstexten. Dabei wird sowohl eine subjektive als auch eine objektive Bewertung der vereinfachten Texte vorgenommen. Während die objektive Bewertung auf quantitativen Metriken basiert, die die Genauigkeit und Kohärenz der Sprache messen, zielt die subjektive Bewertung darauf ab, die Verständlichkeit und Lesbarkeit aus menschlicher Perspektive zu beurteilen.

4.1 Objektive Bewertung

Die objektive Bewertung der Modellleistung erfolgte durch die Anwendung standardisierter quantitativer Metriken, insbesondere BLEU (Bilingual Evaluation Understudy) [49] und SARI (System Output Against References and Input) [50].

4.1.1 Bewertung mit BLEU

BLEU wurde ursprünglich für die automatische Bewertung von maschinellen Übersetzungen entwickelt. Es misst die Übereinstimmung von n-Grammen (Wortgruppen) zwischen dem generierten Text und Referenztexten. Dabei basiert BLEU auf der Annahme, dass eine höhere Übereinstimmung mit den Referenztexten auf eine bessere Qualität des Outputs hinweist. Die Metrik kombiniert modifizierte Präzision mit einer Kürzungsstrafe, um sicherzustellen, dass die generierten Texte nicht unnötig verkürzt werden [49].

Ein Vorteil von BLEU ist seine Spracheunabhängigkeit, was es vielseitig einsetzbar macht. Es wurde jedoch beobachtet, dass BLEU-Änderungen, die die Lesbarkeit verbessern oder den Text vereinfachen, nicht immer positiv bewertet, da es sich primär auf die Ähnlichkeit mit den Referenzen konzentriert. Dadurch eignet sich BLEU weniger, um spezifische Vereinfachungsaspekte zu bewerten [49].

4.1.2 Bewertung mit SARI

SARI wurde speziell für die Textvereinfachung entwickelt und berücksichtigt drei zentrale Operationen: Hinzufügen, Löschen und Beibehalten von Informationen. Im Gegensatz zu BLEU vergleicht SARI den generierten Text nicht nur mit den Referenztexten, sondern auch mit dem Originaltext. Dadurch ermöglicht es eine detaillierte Bewertung, wie gut ein Text vereinfacht wurde, ohne dabei wichtige Informationen zu verlieren [50]. Die Qualität der Vereinfachung wird durch eine Kombination von Präzision und Vollständigkeit (Recall) für jede dieser drei Operationen gemessen. Ein hoher SARI-Wert

zeigt an, dass das Modell erfolgreich unnötige Komplexität entfernt hat, ohne den Sinn des Textes zu verändern [50].

4.1.3 Die Ergebnisse der objektiven Bewertung

Zur objektiven Bewertung wurden Verwaltungstexte manuell vereinfacht und als Referenz verwendet, um die Leistung der Modelle systematisch zu vergleichen. Dabei wurde für jedes der drei Modelle eine Stichprobe von zehn Sätzen aus Verwaltungstexten ausgewählt und mit den Modellen vereinfacht. Anschließend wurden die generierten Texte anhand der BLEU- und SARI-Metriken bewertet, indem sie mit den menschlich erstellten Referenzversionen verglichen wurden. Der Durchschnitt der BLEU- und SARI-Werte aus diesen zehn Sätzen wurde berechnet und in der folgenden Tabelle 3 dokumentiert, um eine fundierte Analyse der Modellleistung zu gewährleisten.

Tabelle 3: BLEU- und SARI-Bewertung

LLM-Modell	**BLEU**	**SARI**
LLaMA 3 8B	0.35130	63.20302
Phi-3 mini	0.30812	58.64624
Mistral 7B	0.35430	63.78774

Die Ergebnisse zeigen, dass Mistral 7B mit einem SARI-Wert von 63.78 die beste Leistung in der Vereinfachung von Verwaltungstexten erzielte. Dies bedeutet, dass das Modell besonders effektiv darin war, komplexe Satzstrukturen zu vereinfachen und gleichzeitig relevante Informationen zu bewahren. Ein hoher SARI-Wert deutet darauf hin, dass das Modell erfolgreich überflüssige Komplexität entfernt hat, ohne den Inhalt zu verzerren.

LLaMA 3 8B, mit einem SARI-Wert von 63.20, zeigte eine ähnlich hohe Vereinfachungsleistung, könnte aber in einigen Fällen geringfügig mehr Informationen entfernt haben als Mistral 7B. Dennoch bleibt es für die administrative Textvereinfachung gut geeignet.

Phi-3 Mini erreichte mit 58.64 den niedrigsten SARI-Wert, was darauf schließen lässt, dass die vereinfachten Texte nicht in jedem Fall die optimale Balance zwischen Verständlichkeit und Informationswahrung erreichten. Während das Modell in der Lage war, Sätze zu verkürzen und vereinfacht darzustellen, könnte es häufiger als die anderen Modelle wichtige inhaltliche Details ausgelassen haben.

Hinsichtlich der BLEU-Werte zeigen die Ergebnisse, dass Mistral 7B mit 0.354 die höchste strukturelle Übereinstimmung mit den Referenztexten aufweist, gefolgt von LLaMA 3 8B mit 0.351 und Phi-3 Mini

mit 0.308. Ein höherer BLEU-Wert deutet auf eine größere Ähnlichkeit mit den menschlich erstellten Referenztexten hin.

Jedoch bedeutet ein niedriger BLEU-Wert nicht zwangsläufig eine schlechtere Vereinfachung. Die Kürzungsstrafe (brevity penalty) beeinflusst die BLEU-Metrik erheblich, da die getesteten Modelle darauf optimiert wurden, kürzere und einfachere Sätze zu erzeugen. Da BLEU kürzere Texte im Vergleich zu den Referenzen bestraft, könnten effektiv vereinfachte Texte niedrigere BLEU-Werte aufweisen, obwohl sie in der tatsächlichen Verständlichkeit überlegen sind.

4.2 Subjektive Bewertung

Zur subjektiven Bewertung wurden dieselben zehn Sätze herangezogen, die bereits für die objektive Bewertung verwendet wurden. Die von den drei Modellen mit den besten Hyperparameter-Kombinationen generierten Texte wurden von drei Personen mit unterschiedlichen sprachlichen und kognitiven Hintergründen bewertet. Ziel war es, die Modellleistung aus der Perspektive realer Nutzergruppen zu beurteilen und die Verständlichkeit sowie Lesbarkeit der vereinfachten Texte zu analysieren. Die Auswahl der Bewerter erfolgte mit dem Ziel, eine möglichst breite Zielgruppe abzudecken, die von der Vereinfachung administrativer Texte profitieren könnte.

Person A war ein Flüchtling mit begrenzter formaler Bildung (bis zur Grundschule) und Deutschkenntnissen auf dem Niveau A1. Seine Englischkenntnisse entsprachen dem Niveau A2, weshalb er administrative Formulare in englischer Sprache bevorzugte. Dennoch hatte er erhebliche Schwierigkeiten, komplexe Sätze und Fachbegriffe zu verstehen, was die Notwendigkeit einer klaren und strukturierten Vereinfachung betonte.

Person B war ein internationaler Student mit Englischkenntnissen auf dem Niveau B1. Er hatte vor allem Schwierigkeiten, englischsprachige Formulare von Behörden wie der Ausländerbehörde vollständig zu verstehen und auszufüllen. Besonders herausfordernd waren für ihn unklare Formulierungen und komplizierte Begriffe, weshalb er detailliertere Erklärungen und eine präzisere Ausdrucksweise bevorzugte.

Person C lebt nach drei Schlaganfällen mit kognitiven Einschränkungen. Diese Person spricht weder Deutsch noch Englisch fließend, verfügt jedoch über Englischkenntnisse auf dem Niveau A2. Aufgrund der kognitiven Beeinträchtigungen und der begrenzten Sprachkenntnisse hatte sie erhebliche Schwierigkeiten, lange und komplexe Sätze zu erfassen. Für ihn waren Texte mit einfacher Struktur, klaren Anweisungen und der Vermeidung von Fachjargon besonders wichtig.

Die Bewertung der vereinfachten Texte erfolgte anhand spezifischer Kriterien, die auf den Prinzipien der Einfachen Sprache basieren und um zusätzliche Dimensionen ergänzt wurden:

1. Verständlichkeit: Ein Satz ist verständlich, wenn er für die Zielgruppe ohne zusätzliche Erklärungen sofort erfassbar ist. Dazu gehört die Verwendung einfacher Wörter, der Verzicht auf Fachbegriffe (es sei denn, sie werden direkt erklärt) und die Vermeidung von abstrakten oder komplizierten Formulierungen.

2. Klarheit: Ein Satz ist klar, wenn seine Aussage eindeutig und direkt formuliert ist. Doppeldeutigkeiten, unnötige Einschübe oder lange verschachtelte Satzkonstruktionen sollten vermieden werden. Jede Aussage sollte klar das ausdrücken, was gemeint ist, ohne überflüssige oder schwer verständliche Informationen.

3. Lesbarkeit: Ein Text ist gut lesbar, wenn er kurze, aktive Sätze enthält, unnötige Füllwörter vermeidet und eine natürliche Satzstruktur aufweist. Aktiv formulierte Sätze („The office processes the application “) sind klarer und verständlicher als passive Formulierungen („The application is processed by the office “). Zudem sollte die Satzlänge auf etwa 15 Wörter begrenzt sein, um die Lesbarkeit zu gewährleisten.

4. Präzision: Ein Satz ist präzise, wenn er die wichtigsten Informationen vollständig und korrekt vermittelt. Der Inhalt sollte weder zu vage noch zu detailliert sein, sondern sich auf das Wesentliche konzentrieren. Überflüssige Informationen oder unklare Formulierungen sollten vermieden werden.

5. Kohärenz: Ein Satz ist kohärent, wenn er sich logisch in den Gesamtkontext des Textes einfügt. Die Verbindung zwischen den Sätzen sollte klar sein, sodass der Text flüssig und nachvollziehbar bleibt. Logische Übergänge und strukturierte Informationen erleichtern das Verständnis.

Die Bewerter wurden gebeten, die von den drei Modellen generierten vereinfachten Texte mit den ursprünglichen, komplexen Verwaltungstexten sowie mit den von GPT-4o generierten vereinfachten Versionen zu vergleichen. Dabei beurteilten sie die Verständlichkeit der Texte auf einer Skala von 1 bis 5, wobei 5 „sehr gut vereinfacht“ und 1 „nicht gut vereinfacht“ bedeutete (Abbildung 12). Diese Methodik ermöglichte eine differenzierte Analyse der Modellleistung und lieferte wertvolle Einblicke in die praktische Anwendbarkeit der generierten Vereinfachungen.

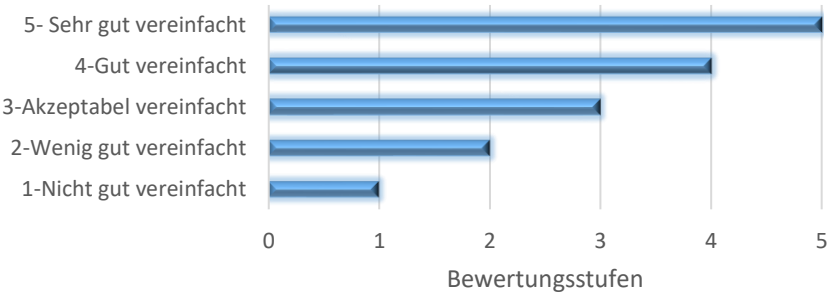

Abbildung 12: Skala zur Bewertung der Textvereinfachung

Durch diese Kriterien konnte die Qualität der generierten Texte nicht nur sprachlich und inhaltlich bewertet, sondern auch ihre Eignung für verschiedene Nutzergruppen untersucht werden. Die Ergebnisse wurden systematisch dokumentiert und analysiert, um die Leistungsfähigkeit der Modelle im Vergleich zu GPT-4 zu evaluieren.

4.2.1 Analyse der Bewertung von Person A

Die Bewertung der Sprachmodelle durch Person A wird im Balkendiagramm (Abbildung 13) visualisiert und hebt die Unterschiede zwischen den Modellen hervor.

Die Verständlichkeit der generierten Texte wurde durchweg positiv bewertet, wobei Mistral 7B mit 94 % die höchste Einstufung erhielt, gefolgt von Phi-3 mini mit 92 % und GPT-4o mit 90 %. LLaMA 3 8B erzielte mit 86 % den niedrigsten Wert, was dennoch auf eine insgesamt verständliche Sprachproduktion hindeutet. Hinsichtlich der Klarheit erzielte Phi-3 mini mit 96 % die beste Bewertung, während GPT-4o mit 94 % und Mistral mit 92 % ebenfalls sehr gut abschnitten. LLaMA 3 8B erreichte mit 88 % den niedrigsten, aber dennoch hohen Wert. Auch die Lesbarkeit wurde durchweg als sehr gut eingestuft, wobei Mistral 7B und GPT-4o mit 96 % die höchsten Werte erreichten, gefolgt von Phi-3 mini mit 92 % und LLaMA 3 8B mit 90 %.

Auffällig waren die Unterschiede in der Präzision. Während Mistral 7B, Phi-3 Mini und LLaMA 3 8B hohe Werte erreichten, blieb GPT-4o leicht zurück, was darauf hindeuten könnte, dass die erstgenannten Modelle genauere und spezifischere Antworten lieferten. Ein besonders konsistentes Ergebnis zeigt sich in der Kohärenzbewertung, bei der alle Modelle eine nahezu perfekte Bewertung von 100 % erhielten. Dies verdeutlicht die Fähigkeit der Modelle, zusammenhängende und logisch strukturierte Texte zu generieren.

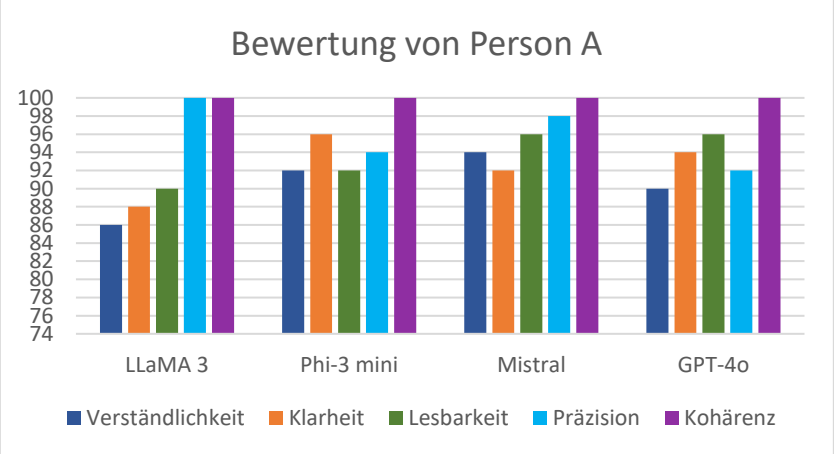

Abbildung 13: Bewertung von Person A

4.2.2 Analyse der Bewertung von Person B

Die Bewertung der Sprachmodelle durch Person B (siehe Abbildung 14) bestätigt die insgesamt hohe Qualität der generierten Texte, weist jedoch einige Unterschiede in der Wahrnehmung einzelner Kriterien im Vergleich zu Person A auf. Während beide Personen eine hohe Verständlichkeit, Klarheit, Lesbarkeit, Präzision und Kohärenz feststellen, zeigen sich in den genauen Bewertungen der einzelnen Modelle leichte Abweichungen, die auf individuelle Präferenzen oder unterschiedliche Interpretationen der Kriterien hindeuten.

LLaMA 3 8B und GPT-4o wurden mit 98 % als die verständlichsten Modelle bewertet, dicht gefolgt von Phi-3 mini mit 96 %. Mistral 7B schnitt mit 90 % etwas schwächer ab, bleibt jedoch weiterhin auf einem hohen Niveau. Hinsichtlich der Klarheit wurde GPT-4o mit 94 % am besten eingestuft, während LLaMA 3 8B mit 92 % ebenfalls gut abschnitt. Phi-3 mini erreichte 84 %, und Mistral 7B erhielt mit 80 % die niedrigste Bewertung, was darauf hinweisen könnte, dass seine Formulierungen als weniger eindeutig wahrgenommen wurden.

Die Lesbarkeit wurde bei LLaMA 3 8B und Phi-3 mini mit 98 % am höchsten bewertet, während GPT-4o mit 96 % ebenfalls sehr gut abschnitt. Mistral 7B erhielt mit 86 % die niedrigste Bewertung, was darauf hindeutet, dass seine Texte für Person B weniger flüssig oder strukturiert wirkten. In der Kategorie Präzision schnitt LLaMA 3 8B mit 92 % am besten ab, während Mistral 7B mit 90% und Phi-3 mini und mit jeweils 86 % leicht darunter lagen. GPT-4o erhielt mit 82 % die niedrigste Bewertung, was darauf hindeuten könnte, dass seine Antworten als weniger spezifisch oder exakt wahrgenommen wurden.

Die Kohärenz wurde insgesamt als sehr hoch bewertet. LLaMA 3 8B erreichte mit 100 % die höchste Bewertung, gefolgt von GPT-4o mit 98 %, Phi-3 mini mit 96 % und Mistral 7B mit 94 %. Diese Ergebnisse zeigen, dass alle Modelle in der Lage sind, zusammenhängende und logisch strukturierte Texte zu generieren, auch wenn leichte Unterschiede in der Wahrnehmung bestehen.

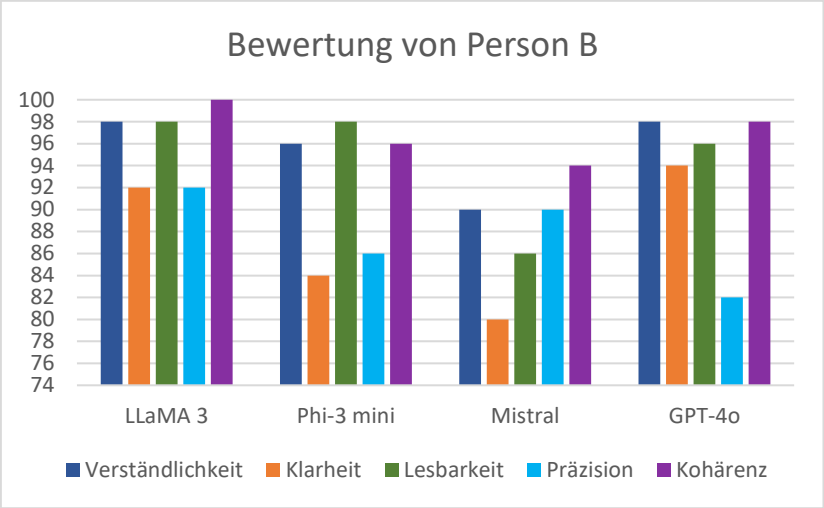

Abbildung 14: Bewertung von Person B

4.2.3 Analyse der Bewertung von Person C

Die Bewertung der Sprachmodelle durch Person C (siehe Abbildung 15) bestätigt die insgesamt hohe Qualität der generierten Texte und zeigt eine noch stärkere Übereinstimmung mit den Einschätzungen der vorherigen Bewertenden. Während sich Person A und B in einigen Aspekten leicht unterschieden, fällt die Bewertung von Person C insgesamt noch positiver aus, insbesondere in den Bereichen Verständlichkeit, Klarheit und Kohärenz. Dies deutet darauf hin, dass die Modelle eine durchweg hohe Qualität aufweisen und sich lediglich in Nuancen unterscheiden.

LLaMA 3 8B und GPT-4o wurden mit einer Verständlichkeit von 100 % bewertet, was darauf hindeutet, dass diese Modelle für Person C besonders leicht verständliche Texte generierten. Phi-3 Mini folgte mit 96 %, während Mistral 7B mit 92 % etwas niedriger eingestuft wurde, aber dennoch einen hohen Verständlichkeitsgrad erreichte. In der Kategorie Klarheit setzten sich LLaMA 3 8B und Phi-3 Mini mit 100 % an die Spitze, während GPT-4o mit 96 % ebenfalls eine sehr hohe Bewertung erzielte. Mistral 7B erhielt mit 92 % die niedrigste Bewertung, was dennoch auf eine durchweg hohe Klarheit hinweist.

Auch die Lesbarkeit wurde durchgehend sehr positiv bewertet. LLaMA 3 8B und GPT-4o erreichten mit 98 % die höchsten Werte, gefolgt von Phi-3 Mini mit 96 %. Mistral 7B erzielte mit 90 % den niedrigsten Wert, blieb jedoch weiterhin auf einem soliden Niveau Ein ähnliches Muster zeigte sich in der Präzision. LLaMA 3 8B erhielt mit 98 % die höchste Bewertung, während Phi-3 Mini und Mistral 7B mit jeweils 94 % sowie GPT-4o mit 92 % leicht darunter lagen. Dies verdeutlicht, dass alle Modelle als präzise empfunden wurden, wobei LLaMA 3 8B erneut leicht vorne lag.

Besonders auffällig ist die durchweg exzellente Bewertung der Kohärenz. LLaMA 3 8B erreichte mit 100 % die höchstmögliche Wertung, während Phi-3 mini, GPT-4o und Mistral mit 98 % nur minimal darunter lagen. Diese Ergebnisse bestätigen, dass alle Modelle in der Lage sind, logisch zusammenhängende und strukturierte Texte zu erzeugen, unabhängig von individuellen Präferenzen.

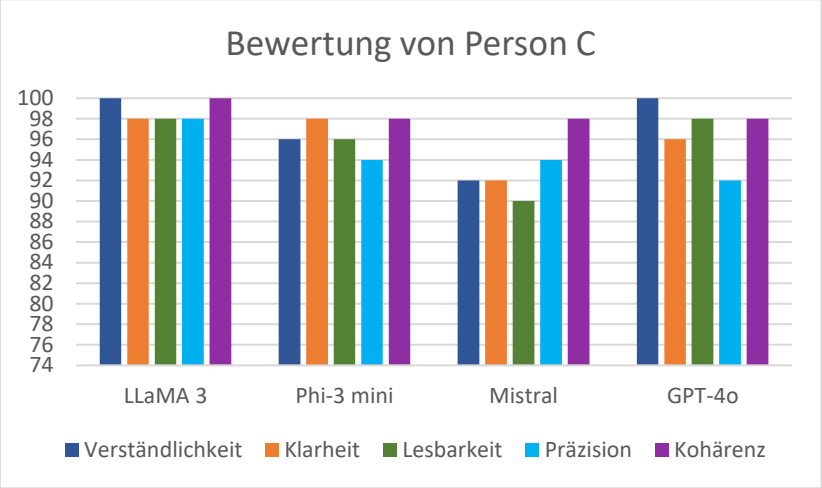

Abbildung 15: Bewertung von Person C

4.3 Gesamtvergleich der Ergebnisse

Die Gesamtbewertung der Sprachmodelle auf Basis sowohl der subjektiven als auch der objektiven Evaluation zeigt ein konsistentes Bild ihrer Leistungsfähigkeit, insbesondere in der Vereinfachung von Verwaltungstexten. Die Ergebnisse beider Ansätze deuten darauf hin, dass Mistral 7B und LLaMA 3 8B besonders leistungsfähig sind und in mehreren Aspekten bessere Ergebnisse als GPT-4o erzielen konnten. Während Mistral 7B im objektiven Vergleich mit einem höchsten SARI-Wert von 63.78 die beste Vereinfachungsleistung zeigte, erzielte LLaMA 3 8B mit 63.20 ein ähnlich hohes Ergebnis, was darauf hindeutet, dass beide Modelle besonders effektiv in der Reduzierung sprachlicher Komplexität sind, ohne wesentliche Inhalte zu verlieren. Phi-3 mini zeigte mit einem SARI-Wert von 58.64 eine etwas geringere Leistung, was darauf schließen lässt, dass es teilweise relevante Informationen stärker vereinfacht oder ausgelassen hat.

Zusätzlich wurde GPT-4o als Referenzmodell in den Vergleichen herangezogen, um die Ergebnisse der drei untersuchten Modelle besser einordnen zu können. Die subjektiven Bewertungen bestätigen weitgehend die Tendenzen der objektiven Ergebnisse, zeigen jedoch Unterschiede in der individuellen Wahrnehmung der Modelle.

In Bezug auf die Verständlichkeit schnitten LLaMA 3 8B und Mistral 7B durchweg gut ab. Besonders Mistral 7B wurde von einer Testperson als das verständlichste Modell wahrgenommen, während LLaMA 3 8B vor allem durch seine hohe Klarheit, Lesbarkeit und Kohärenz überzeugte. Phi-3 mini zeigte ebenfalls eine starke Verständlichkeit, wurde jedoch in den Kategorien Klarheit und Präzision inkonsistenter bewertet.

Obwohl Phi-3 mini als kleineres Modell im Vergleich zu Mistral 7B und LLaMA 3 8B eine geringere Anzahl an Parametern aufweist, konnte es dennoch akzeptable Ergebnisse liefern. Dies zeigt, dass auch kompaktere Modelle in der Lage sind, Verwaltungstexte zu vereinfachen. Allerdings weist Phi-3 mini in einigen Bereichen noch Optimierungspotenzial auf und würde durch zusätzliches Fine-Tuning vermutlich deutlich bessere Ergebnisse erzielen.

Auffällig ist, dass GPT-4o in den subjektiven Einschätzungen teilweise hinter den anderen Modellen zurückblieb, insbesondere in der Präzision. Dies deutet darauf hin, dass Mistral 7B und LLaMA 3 8B durch gezieltes Fine-Tuning besonders gut an spezifische Anforderungen angepasst werden können und dadurch eine effektivere Verwaltungstextvereinfachung ermöglichen als GPT-4o.

Zusammenfassend zeigen die Ergebnisse, dass die Kombination aus objektiven und subjektiven Bewertungen ein umfassendes Bild der Modellleistung liefert. Während die objektiven Metriken wie SARI und BLEU eine quantifizierbare Leistung bei der Vereinfachung und strukturellen Übereinstimmung der Texte bieten, reflektieren die subjektiven Bewertungen die tatsächliche Lesbarkeit und Verständlichkeit aus Nutzersicht. Die Konsistenz in den Ergebnissen beider Ansätze zeigt, dass Mistral 7B und LLaMA 3 8B besonders gut für die Textvereinfachung geeignet sind, während Phi-3 mini in bestimmten Bereichen noch Optimierungspotenzial hat. Die Ergebnisse verdeutlichen zudem, dass durch gezieltes Fine-Tuning diese Modelle bessere Ergebnisse als GPT-4o erzielen können.

5 Zusammenfassung und Überblick

Diese Arbeit belegt das Potenzial LLM-gestützter Textvereinfachung zur Optimierung administrativer Kommunikation. Die Analyse verschiedener Modelle zeigt, dass gezieltes Fine-Tuning die Verständlichkeit komplexer Verwaltungstexte signifikant verbessern kann, während die inhaltliche Präzision erhalten bleibt. Dies stellt einen bedeutenden Fortschritt dar, um sprachliche Barrieren abzubauen und behördliche Informationen einem breiteren Publikum zugänglich zu machen.

In den folgenden Abschnitten werden ein kurzes Fazit sowie zukünftige Forschungsansätze zur Weiterentwicklung der KI-gestützten Textvereinfachung vorgestellt.

5.1 Fazit

Die Ergebnisse dieser Arbeit zeigen, dass automatisierte Textvereinfachung mit LLMs eine vielversprechende Lösung zur Verbesserung der Verständlichkeit von Verwaltungstexten darstellt. Die untersuchten Modelle LLaMA 3 8B, Phi-3 Mini und Mistral 7B konnten durch gezieltes Fine-Tuning effektiv auf behördliche Kommunikationsanforderungen abgestimmt werden, sodass sie sprachliche Komplexität reduzieren, ohne wesentliche Inhalte zu verfälschen.

Ein zentrales Ergebnis dieser Untersuchung ist die überlegene Präzision der drei Modelle im Vergleich zu GPT-4o. In der subjektiven Bewertung wurden sie durchweg als genauer und kohärenter eingestuft, insbesondere in der Fähigkeit, komplexe Verwaltungssprache in eine verständlichere Form zu übertragen. Gleichzeitig untermauerten die objektiven Metriken ihre hohe Leistungsfähigkeit und bestätigten, dass auch kleinere, spezialisierte Modelle mit modernsten LLMs konkurrieren können.

Bemerkenswert ist, dass diese ressourcenschonenderen Modelle trotz ihrer geringeren Parameteranzahl nicht nur vergleichbare, sondern in bestimmten Aspekten sogar überlegene Ergebnisse erzielten. Dies unterstreicht das Potenzial dieser Modelle, die mit einem gut strukturierten Datensatz für spezifische Anwendungen sehr effektiv feinabgestimmt werden können. Die Qualität der Trainingsdaten erweist sich dabei als zentraler Faktor für die Leistungsfähigkeit der Modelle und ermöglicht eine gezielte Anpassung an domänenspezifische Anforderungen.

Darüber hinaus zeigt die Arbeit die gesellschaftliche Relevanz KI-gestützter Sprachvereinfachung im öffentlichen Sektor. Die Fähigkeit, behördliche Texte barrierefreier zu gestalten, trägt nicht nur zur Inklusion sprachlich oder kognitiv eingeschränkter Personen bei, sondern fördert auch die allgemeine Transparenz und Verständlichkeit administrativer Prozesse.

5.2 Zukünftige Forschungsansätze

Die Weiterentwicklung LLM-gestützter Textvereinfachung erfordert insbesondere die Optimierung und Erweiterung von Trainingsdatensätzen, da die Modellleistung maßgeblich von der Qualität und Domänenspezifität der Daten abhängt. Ein von Experten kuratierter Korpus könnte sicherstellen, dass vereinfachte Verwaltungstexte nicht nur sprachlich zugänglich, sondern auch inhaltlich präzise und rechtlich korrekt bleiben. Zudem sollten mehrsprachige Modelle erforscht werden, um Verwaltungskommunikation in mehrsprachigen Gesellschaften zu erleichtern.

Ein weiteres zukunftsweisendes Forschungsfeld ist die kontrollierte Textvereinfachung, die es ermöglichen würde, den Grad der Vereinfachung gezielt an unterschiedliche Zielgruppen anzupassen, ohne inhaltliche Genauigkeit zu verlieren. Zudem könnten multimodale Modelle, die Bilder, Symbole oder interaktive Elemente einbinden, die Barrierefreiheit weiter erhöhen. Schließlich erfordert die langfristige Integration von LLMs in Verwaltungssysteme eine interdisziplinäre Zusammenarbeit, um ethische, rechtliche und datenschutzrechtliche Herausforderungen zu bewältigen.

6 Literaturverzeichnis

[1] K. Woodsend und M. Lapata, Learning to Simplifying Sentences with Quasi-Synchronous Grammar and Integer Programming, Association for Computational Linguistics (ACL), 2011.

[2] A. Siddharthan, „A Survey of Research on Text Simplification," *ITL - International Journal of Applied Linguistics,* Bd. 165, 2014.

[3] M. Shardlow, „A Survey of Automated Text Simplification," *International Journal of Advanced Computer Science and Applications (IJACSA),* Bd. 5, pp. 58-65, 2014.

[4] J. McElvenny, „The Application of C.K. Ogden's Semiotics in Basic English," *Language Problems and Language Planning,* Bd. 39, p. 263–285, 2015.

[5] VOA, „The Founding of Special English Program," VOA, 26 10 2009. [Online]. Available: https://www.voanews.com/a/a-13-a-2002-03-05-22-the-66277882/540311.html. [Zugriff am 04 07 2024].

[6] U. Government, „Plain Writing Act of 2010," 2010.

[7] Bundesministerium der Justiz und für Verbraucherschutz, „Gesetz zur Gleichstellung von Menschen mit Behinderungen (Behindertengleichstellungsgesetz - BGG)," 2002. [Online]. Available: https://www.gesetze-im-internet.de/bgg/BJNR146800002.html. [Zugriff am 01 07 2024].

[8] J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin und J. Tait, „Simplifying Text for Language-Impaired Readers," in *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999),* 1999.

[9] G. H. Paetzold und L. Specia, „Lexical Simplification with Neural Ranking," in *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), Volume 2, Short Papers,* 2017.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser und I. Polosukhin, „Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017),* 2017.

[11] Development, Organisation for Economic Co-operation and, „Accessible and Inclusive Public Communication," 2022. [Online]. Available:

https://www.oecd.org/en/publications/2022/09/accessible-and-inclusive-public-communication_60c0dc6e.html. [Zugriff am 14 8 2024].

[12] IAO, Fraunhofer-Institut für Arbeitswirtschaft und Organisation, „KI-Tools für die öffentliche Verwaltung," 24 10 2024. [Online]. Available: https://www.iao.fraunhofer.de/de/veranstaltungen/2024/ki-tools-fuer-die-oeffentliche-verwaltung-okt.html. [Zugriff am 20 02 2025].

[13] C. Maaß, „Leichte Sprache: Das Regelbuch," Berlin, 2015.

[14] Bildung, Bundeszentrale für politische, „Aus Politik und Zeitgeschichte," 19 02 2014. [Online]. Available: https://www.bpb.de/shop/zeitschriften/apuz/179341/leichte-und-einfache-sprache-versuch-einer-definition/?utm_source=chatgpt.com. [Zugriff am 04 06 2024].

[15] M. Cutts, Oxford Guide to Plain English, Oxford University Press, 2009.

[16] Campaign, Plain English, „Plain English Guidelines," [Online]. Available: https://www.plainenglish.co.uk/free-guides.html. [Zugriff am 04 06 2024].

[17] A. Government, „Style Manual: Writing and Editing for Government," [Online]. Available: https://www.stylemanual.gov.au/. [Zugriff am 04 06 2024].

[18] I. Europe, „Information for All: European Standards for Making Information Easy to Read and Understand," Belgium, 2010.

[19] (NCDAE), National Center on Disability and Access to Education, „Best Practices for Institution-Wide Web Accessibility," [Online]. Available: https://ncdae.org/goals/accreditation/bestpractices.php. [Zugriff am 20 02 2025].

[20] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow und e. al., „BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," 2023.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou und W. L. u. P. J. Liu, „Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, pp. 1-67, 2020.

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave und G. Lample, „LLaMA: Open and Efficient Foundation Language Models," 2023.

[23] OpenAI, „GPT-4 Technical Report," 2023.

[24] H. Team, „PEFT,“ Huggingface, [Online]. Available: https://huggingface.co/docs/peft/index. [Zugriff am 09 06 2024].

[25] Z. Han, C. Gao, J. Liu, J. (. Zhang und S. Q. Zhang, „Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey,“ 2024.

[26] L. Wang, S. Chen, L. Jiang, S. Pan, R. Cai, S. Yang und F. Yang, „Parameter-Efficient Fine-Tuning in Large Models: A Survey of Methodologies,“ 2024.

[27] C. C. S. Balne, S. Bhaduri, T. Roy, V. Jain und A. Chadha, „Parameter Efficient Fine Tuning: A Comprehensive Analysis Across Applications,“ 2024.

[28] F. Musumeci, M. Brienza, V. Suriani, D. Nardi und D. D. Bloisi, „LLM Based Multi-Agent Generation of Semi-structured Documents from Semantic Templates in the Public Administration Domain,“ 2024.

[29] J. Mandravickaitė, E. Rimkienė, D. K. Kapkan, D. Kalinauskaitė und T. Krilavičius, „Automatic Simplification of Lithuanian Administrative Texts,“ *Algorithms,* 20 11 2024.

[30] S. Klöser, M. Beele, J.-N. Schagen und B. Kraft, „German Text Simplification: Finetuning Large Language Models with Semi-Synthetic Data,“ 2024.

[31] P. Martínez, L. Moreno und A. Ramos, „Exploring Large Language Models to Generate Easy to Read Content,“ 2024.

[32] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell und e. al., „Language Models are Few-Shot Learners,“ in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le und D. Zhou, „Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,“ 2022.

[34] X. Amatriain, „Prompt Design and Engineering: Introduction and Advanced Methods,“ 2024.

[35] Z. Gekhman, G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart und J. Herzig, „Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?,“ in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Miami, Florida, USA, 2024.

[36] S. Gooding, „On the Ethical Considerations of Text Simplification,“ in *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, Dublin, Irland, 2022.

[37] D. Jurafsky und J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, Online-Manuskript, 2025.

[38] Y. Bengio, R. Ducharme, P. Vincent und C. Jauvin, „A Neural Probabilistic Language Model," *Journal of Machine Learning Research,* p. 1137–1155, 2003.

[39] S. Hochreiter und J. Schmidhuber, „Long Short-Term Memory," *Neural Computation,* p. 1735–1780, 1997.

[40] AIML.com Machine Learning Resources, „Compare the different Sequence model (RNN, LSTM, GRU, and Transformers)," AIML.com Machine Learning Resources, 05 05 2024. [Online]. Available: https://aiml.com/compare-the-different-sequence-models-rnn-lstm-gru-and-transformers/. [Zugriff am 26 02 2025].

[41] D. Bahdanau, K. Cho und Y. Bengio, „Neural Machine Translation by Jointly Learning to Align and Translate," in *3rd International Conference on Learning Representations (ICLR)*, 2015.

[42] Z. Chu, S. Ni, Z. Wang, X. Feng, C. Li, X. Hu, R. Xu, M. Yang und W. Zhang, „History, Development, and Principles of Large Language Models—An Introductory Survey," *AI and Ethics,* 2024.

[43] J. Devlin, M.-W. Chang, K. Lee und K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

[44] M. AI, „Meta LLaMA 3: The Next Generation of Open-Source Language Models," Meta AI, 18 04 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/. [Zugriff am 25 05 2024].

[45] Microsoft, „Tiny but mighty: The Phi-3 small language models with big potential," Microsoft, 23 04 2024. [Online]. Available: https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/. [Zugriff am 04 06 2024].

[46] Microsoft, „Phi-3: Neue Maßstäbe für die Möglichkeiten kleiner Sprachmodelle," 25 04 2024. [Online]. Available: https://news.microsoft.com/de-de/phi-3-neue-massstaebe-fuer-die-moeglichkeiten-kleiner-sprachmodelle/. [Zugriff am 04 06 2024].

[47] M. A. Team, „Mistral 7B," Mistral AI, 27 09 2023. [Online]. Available: https://mistral.ai/news/announcing-mistral-7b. [Zugriff am 04 06 2024].

[48] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang und W. Chen, „LoRA: Low-Rank Adaptation of Large Language Models,“ 2021.

[49] K. Papineni, S. Roukos, T. Ward und W.-J. Zhu, „BLEU: a Method for Automatic Evaluation of Machine Translation,“ in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[50] W. Xu, C. Napoles, E. Pavlick, Q. Chen und C. Callison-Burch, „Optimizing Statistical Machine Translation for Text Simplification,“ in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

Anhang

Der vollständige Code der Implementierung sowie die vollständigen Bewertungen der Testpersonen sind auf der beiliegenden CD gespeichert. Zusätzlich ist der Code unter folgendem Repository verfügbar: https://git.haw-hamburg.de/infwht611/llm-basierte-vereinfachung-von-verwaltungstexten. Die wichtigsten Ergebnisse sind in Kapitel 3 und 4 zusammengefasst.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit mit dem Titel:

LLM-basierte Vereinfachung von Verwaltungstexten

selbständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

Datum	Unterschrift