Masterthesis

From difficult to easy
-
Developing a language model for translating official into Easy Language

Maximilian Müller

October 2023 – April 2024

Matriculation Id: 3520170
Course of Study: Applied Computer Science

Reviewer:
Professor Dr.-Ing. Zesch
Univ.-Prof. Dr. Beecks

FernUniversität in Hagen
Center of Advanced Technology for Assisted Learning and Predictive Analytics
Research Professorship Computational Linguistics
58097 Hagen

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere weiterhin, dass ich diese Arbeit noch keinem anderen Prüfungsgremium vorgelegt habe.

München, im April 2024

....

Maximilian Müller

Abstract

The field of Natural Language Programming (NLP) and its Large Language Models (LLMs) is rapidly evolving, with new breakthroughs being published almost weekly. However, not everyone can keep up with this rapid evolution, let alone understand it. To create societal benefits with these capabilities, it is necessary to simplify complex information. This is particularly relevant for government communication with citizens, as governments must provide accessible information and services for all members of our digital society, regardless of their language, education or abilities.

Current advances in NLP, especially LLMs, allow us to automatically translate large amounts of text into a simplified version. However, the quality of these text simplifications is not yet sufficient. Therefore, the aim of this study is to identify the most appropriate enhancement method for a Large Language Model to facilitate the intralingual translation of German text into Easy Language (EL). In doing so, we also discuss how data should be processed to become a valuable asset.

This thesis is based on a real-life scenario and its results are meant to be a future implementation in the City of Munich's internal LLM Chatbot "MUCGPT". Consequently, the base data for our experiments is built from public websites from the domain of the City of Munich. The website pairs, in both normal and Easy Language, represent the ideal text/simplification pairs mapped in our parallel corpora. We translate all parallel corpora records using LLMs that have been enhanced with Embedding, Fine Tuning, Prompt Engineering and the combinations of these methods. Afterwards, the translations are evaluated using various metrics that cover the categories of Direct Assessment of Simplicity, Simplicity Gain and Structural Simplicity. The reviewed enhancement methods are ranked based on their suitability for Easy Language translation according to these metrics.

Based on the evaluation metrics used in our thesis, it can be concluded that all enhancement methods are suitable to some extent for intralingual translation of German text into Easy Language. However, Fine Tuning demonstrates the highest level of suitability and effectiveness among the tested methods.

The research addresses unresolved issues from previous papers and confirms the feasibility of LLM-based Easy Language translation. However, human review of these translations remains necessary. Additionally, this thesis presents future research opportunities, such as the development of a standardised, automated framework for evaluating Easy Language.

Our code and results have been published on GitHub/D2E. We hope that our work will assist future research.

Contents

Chapter 1

Introduction

It is a crucial obligation for German governments to provide modern digital points of contact that are accessible to everyone (§1 Abs. 1 Onlinezugangsgesetz (OZG)), as 74% of citizens in Germany still use pen and paper to interact with the government (Kühnhenrich and Michalik 2020). Although governmental services, such as Bremen's application for parental allowance *"ELFE - Einfach Leistungen für Eltern"* (Kühnhenrich and Michalik 2020), are already available online, they still face the challenge of being provided without barriers (§3 Abs. 1 OZG, Art. 3 Abs. 3 GG Grundgesetz (GG), § 11 Behindertengleichstellungsgesetz (BGG)). In fact, 61% of citizens have difficulty understanding government information (Kühnhenrich and Michalik 2020). The obligation to use digital media may worsen communication barriers, especially for people with disabilities or non-native speakers (Maaß 2020, pp. 22-26).

To promote inclusion for all citizens in its eGovernment services (LHM a), the City of Munich has taken steps such as publishing an increasing number of public websites in Easy Language (LHM b).

Easy Language was developed as an accessibility standard to make information accessible *"for people with limited language skills (Petersen and Ostendorf 2007; Watanabe et al. 2009; Allen 2009; De Belder and Moens 2010; Siddharthan and Katsos 2010) or language impairments such as dyslexia (Rello et al. 2013), autism (Evans et al. 2014), and aphasia (Carroll et al. 1999)"* (Xu et al. 2016). It adheres to specific rules developed by the "Netzwerk Leichte Sprache" (NLS 2022) for simplifying texts, such as:

- Use simple instead of complicated **words**.
  Good: allow
  Bad: authorise

- Avoid special **characters** or explain them if they are indispensable.
  Good: The sign for paragraph is §
  Bad: §

- Write with simple **sentence** structure.
  Good: Use short sentences.
  Use a new line for each sentence.
  Bad: Use very long sentences, which have relative clauses and span over several topics and lines.

Additionally, in 2023, the City of Munich has launched an internal NLP-based tool called "MUCGPT" to provide a valuable service for its residents in the future. MUCGPT's operational capabilities include

- a chat dialogue,
- summarisation of a given text
- and generating a mind map on a given topic.

To develop MUCGPT into a standalone eGovernment service, it needs to be accessible to all users. Therefore, we propose adding a new feature that translates text into Easy Language. Currently, the chat dialogue fails to meet our standards (Table 1.1).

Prompt
Bitte in Leichte Sprache übersetzen: "Sie erhalten den Personalausweis im Scheckkartenformat."

Expected Translation
Ein Personal-Ausweis ist ein Ausweis-Dokument.
Der Personal-Ausweis sieht aus wie eine Plastik-Karte.

Actual Translation
Sie bekommen eine kleine Karte als Personalausweis.

Table 1.1: A comparison between the desired translation and the current translation by MUCGPT.

Even in this brief example, the untrained LLM fails to identify important characteristics of German text simplification (NLS 2022):

- No explanation of the difficult word "Personalausweis" is given.
- No decomposition of the compound word "Personalausweis" into "Personal-Ausweis" was executed.

Several studies have explored the potential of automatic text simplification (Ebling et al. 2022; Xu et al. 2016; Sulem et al. 2018c), automatic evaluation for text simplification (Alva-Manchego et al. 2021; Ebling et al. 2022; Sun et al. 2021), or even whether text simplification preserves meaning (Agrawal and Carpuat 2023). However, these studies often focus on specific approaches and do not necessarily use LLMs as their foundation. As a result, existing research does not compare which LLM enhancement method is most suitable for improving the quality of automatic text simplification.

The objective of this thesis is to explore the potential of using LLMs to improve government services. Using MUCGPT as an impulse, we will examine the following objectives, in order to harness the possibilities of LLMs to improve the accessibility of the eGovernment of the City of Munich.

Which technique is most suitable for improving an LLM's ability to translate complex text into Easy Language?

- *How does the existing public data need to be prepared for this purpose?*
- *How does the existing language model need to be enhanced with the prepared data?*

We aim to demonstrate the significant impact that the choice of the appropriate LLM enhancement method has on the generation of Easy Language. EL is an essential tool for overcoming one of the barriers that eGovernment faces. To achieve this, we present parallel corpora compiled from the City of Munich's public websites. This data serves as the basis for the chosen enhancement methods. The results of the text simplification were evaluated using automated metrics that cover different aspects of simplicity. Rankings for the enhancement methods were derived based on the metric results.

This paper solely focuses on Easy Language as defined by the NLS (2022). This thesis does not address Easy Language Plus, Plain Language, or any other forms of simplified language, as described by Maaß (2020), or otherwise. Furthermore, our research is limited to the textual component of Easy Language. The NLS (2022) has established rules for font size, font highlighting, paper characteristics and picture specifications, none of which are within the scope of this research.

The study compares the currently most common enhancement methods. The evaluation was limited to automated metrics due to the fact that certified human evaluation exceeded both the thesis time limit and budget.

Chapter 2

Related Work

The thesis focuses on enhancing the ability of an LLM to simplify text for Easy Language. To achieve this goal, existing research on the four main topics of our work is reviewed in this chapter:

- data preparation
- automatic text simplification
- evaluation of simplicity
- LLM enhancement

Our domain-specific data (Section 3.2) is derived from the government realm of the City of Munich. In this regard, we furthermore reviewed existing research on the situation of Natural Language Processing capabilities and accessibility in government services.

2.1 Data Preparation

The use of parallel corpora in LLM text simplification has advanced significantly over the years. Previous studies utilised parallel corpora as a resource for *"a sentence simplification model by tree transformation"*, which learned its parameters from the provided examples of complex and simplified sentences (Zhu et al. 2010). The use of parallel corpora to train and fine tune machine translation models has become more common with the rise of neural networks and deep learning. This allows for the generation of simplified text from complex input, resulting in improved quality and efficiency of automated text simplification (Sulem et al. 2018b).

Parallel corpora have been crucial in assessing and utilising lexical, syntactic and semantic complexities in texts, allowing for more targeted and effective simplification strategies (Zhu et al. 2010). Additionally, researchers have employed parallel corpora to create machine learning models that can automate the evaluation of the text simplifications (Sulem et al. 2018b). These models learn from pairs of complex and simplified sentences and are fine tuned to generate a structure-aware evaluation metric for of simplified texts.

Researchers generally agree that parallel corpora are valuable for training text simplification systems. They offer examples of complex and simplified sentences that can effectively train models. Additionally, they serve as a gold standard for evaluating simplification output (Kajiwara and Komachi 2018). However, there is disagreement among scholars regarding the necessity of simplified corpora. Kajiwara and Komachi (2018) propose a method for achieving text simplification without simplified corpora. Their method involves sequentially simplifying sentences based on word difficulty and sentence readability. In contrast, Martin et al. (2023) emphasize the importance of parallel corpora for reliable automatic text simplification. They highlight the challenges that need to be addressed to boost the potential of parallel corpora.

Current research on parallel corpora in text simplification highlights several limitations. One of these limitations is the scarcity of parallel corpora across different languages, domains and text types, which restricts the usefulness of models trained on these datasets (Martin et al. 2023). Secondly, the challenge of creating large parallel corpora due to the heavy reliance on human annotators, which is often an expensive and time-consuming process, remains unresolved (Kajiwara and Komachi 2018). Additionally, evaluating simplified texts still poses a significant challenge. While parallel corpora can serve as a reference, determining the 'best' simplification is subjective due to varying individual needs and preferences. Research is needed to investigate how to modify text simplification systems trained on parallel corpora for specific user groups, such as non-native speakers or individuals with cognitive disabilities. Most existing systems concentrate on simplifying for a general audience.

2.2 Automatic Text Simplification

Automatic text simplification research has shifted from simple language models (Kauchak 2013) to a focus on Large Language Models (Zhou et al. 2023). For example Deilen et al. (2023) who use GPT (Generative Pre-Training, Radford et al. 2018) for automatic text simplification or Ormaechea et al. (2023) who base their approach to enhance sentence complexity assessment on BERT (Birectional Encoder Representations from Transformers, Devlin et al. 2019).

Table 2.1 demonstrates that simplifying texts can *"be implemented by three major types of operations: splitting, deletion and paraphrasing (Feng 2008)"* (Xu et al. 2016). During translation, it is crucial to preserve the original context (Devaraj et al. 2022; Agrawal and Carpuat 2023; Bang et al. 2023). Texts should not be split or rearranged in a way that alters the original meaning. However, preserving context can be problematic, as demonstrated by Djeffal and Horst (2021). This issue can be addressed by using the Barzilay and Elhadad (2003) algorithm for alignment, which ensures coherent meaning in the translated text, as demonstrated by Klaper et al. (2013).

Several papers share the finding that current systems still make critical errors and have not yet achieved perfection. Maaß (2020) for example argues, that in order to be accessible, simplified text *"has to be retrievable, perceptible, comprehensible, linkable, acceptable and action-enabling."* Critical errors may stem from the fact that the *"relation between these different requirements are dilemmatic and actions taken to improve one of these requirements may incur damage for another."* Furthermore, Agrawal and Carpuat

Normal Language
Die Stadt München arbeitet beständig an der Verbesserung ihres Webauftritts und konnte aufgrund der Fülle des Materials und der Komplexität der Seite noch nicht alle Inhalte und Services digital barrierefrei gestalten.
Easy Language
Die Internet-Seite soll noch weniger Barrieren haben. Die Stadt München arbeitet immer weiter daran.

Table 2.1: A real-life example of text simplification. The original sentence was split into two and the order of the sentences was rearranged. Paraphrased counterparts are highlighted with green and blue underscores. Removed content is marked with a red underscore.

(2023) suggest that *"prompted LLMs (as ChatGPT)"* are not as accurate as *"[s]upervised systems that leverage pre-trained knowledge"*. Those supervised systems approach the level of *"human written texts regarding reading comprehension accuracy"*. However, they still fail to preserve meaning in at least 14% of cases.

2.3 Evaluation of Simplicity

From Flesch (1948) to modern metrics like SARI (Xu et al. 2016), all share a mathematical approach for the evaluation of text simplicity. However, new challenges, such as hallucinations in LLM-based text simplifications, have led to the emergence of new metric demands, such as *"Evaluating Factuality in Text Simplification"* (Devaraj et al. 2022). Their research highlights the *"need for research into ensuring the factual accuracy of automated simplification models"*, as they found that *"inserting statements unsupported by the corresponding original text, or [...] omitting key information"* occured in *"standard simplification datasets and state-of-the art model outputs"*.

Before delving into more complex analysis, such as factual accuracy, it is important to start with a foundational understanding of text simplicity. Text simplicity can be classified into three types:

- **Direct Assessment (DA)** evaluates translations in isolation because *"there is the potential for ratings of different translations to influence each other."* The system uses *"a continuous rating scale"* where rankings can be *"combined into mean and median scores"*. Thus, it facilitates *"powerful statistical analys[i]s of systems"*, as *"the law of large numbers suggests that as increasing numbers of individual scores are collected for each system, the mean score will increasingly approach the true score. By simply increasing the number of assessments [...] accuracy is improved."* (Graham et al. 2017)

- **Simplicity Gain (SG)** is assessed *"by counting how many successful lexical or syntactic paraphrases occurred in the simplification. [...] [U]sing simplicity gain avoids over-punishment of errors, which are already penalised for poor meaning* *retention and grammaticality, and thus reduces the bias towards very conservative models."* (Xu et al. 2016)

- **Structural Simplicity (SS)** refers to techniques that simplify the text, such as splitting complex sentences, using simpler synonyms for complex word, or removing unnecessary or redundant information to streamline the content. (Sulem et al. 2018c; Sun et al. 2021; NLS 2022)

While BLEU (Bilingual Evaluation Understudy, Papineni et al. 2002) metrics are widely used in Natural Language Processing tasks for evaluating the quality of machine-produced translations, they may not entirely ensure the factual accuracy of automatically simplified texts. BLEU significantly focuses on comparing n-grams between the machine-produced text and a human reference, providing a score that encapsulates mainly the fluency and coherence. Factual accuracy, on the other hand, requires a deep understanding of the content semantics, which BLEU may not fully capture. Thus, while BLEU can provide insights into the legibility and linguistic quality of the simplified text, additional metrics or methods might be required to confirm factual consistency (Sulem et al. 2018a). Tan et al. (2015) share these findings, as they state that *"studies on the schism between BLEU and manual evaluation highlighted the poor correlation between [machine translation] systems with low BLEU scores and high manual evaluation scores."* Agrawal and Carpuat (2023) also suggest that *"SARI is a better metric than meaning-preservation metrics such as [...] BLEU to rank systems by adequacy"*.

Therefore we do not use BLEU but a variety of metrics, including SARI, to evaluate text simplicity as well as meaning preservation of our translations. The selected metrics and their functionality are further discussed in Section 3.4.

When choosing metrics for this research, we needed to setup a mixture of metrics with different focuses, as there are many metrics to evaluate text simplicity, but Easy Language *"poses a special challenge compared to other text simplification tasks. It is specified by rules and regulations and, in the future, also by a DIN standard [DI23]. In this respect, all texts would have to be checked for precisely these sets of rules in order to determine their level of Leichte Sprache. Currently, there are no scientific publications on this topic."* (Schomacker 2023). Furthermore, Agrawal and Carpuat (2023) highlight, that *"current evaluation protocols assess system outputs for simplicity and meaning preservation without regard for the document context in which output sentences occur and for how people understand them."* Both of which are key factors for our research that need to be mitigated by the chosen combination of metrics.

2.4 LLM Enhancement Methods

The development of enhancement methods is closely linked to the evolution of LLMs themselves. GPT-3, for example, has *"175 billion parameters, 10x more than any previous nonsparse language model"*. As a result, it has become *"competitive with prior state-of-the-art Fine Tuning approaches"* (Brown et al. 2020). In the following we will discuss the three current go-to enhancement methods. We hope that they will help us improve our results.

2.4.1 Embedding

Over time, Embedding has evolved from the introduction of efficient algorithms with Word2Vec (Mikolov et al. 2013b) to the implementation of generative pre-training for language models (Radford et al. 2018). Word2Vec represented a paradigm shift in how machines could comprehend word semantics and relationships by representing words as dense vectors in a continuous vector space. GPT-3 (Brown et al. 2020) builds upon this by dynamically computing word representations based on the context in which the word appears. Its embeddings play a crucial role in improving language understanding.

Word embeddings are effective in text analysis due to their unique properties and inherent structure. They can translate high-dimensional complex data, such as text, into a more manageable, lower-dimensional, dense vector space. This is known as representation learning, and it is the key to their success. The vectors can capture and represent semantic and syntactic relationships between words, preserving subtle linguistic patterns. These embeddings can hold substantial meaning, making them an essential tool in various natural language processing and machine learning applications. (Turney and Pantel 2010; Mikolov et al. 2013a).

Research indicates that contextualised embeddings are more effective than traditional static embeddings in capturing nuanced semantic information Devlin et al. (2019). Static word embeddings represent words as fixed, context-independent vectors in a high-dimensional space. In this paradigm, a word has the same representation regardless of its context, meaning that it can not capture polysemy or the different meanings a word can have based on where and how it is used. Studies such as Kumar et al. (2020), on the other hand, have investigated preprocessing techniques tailored for text simplification tasks and domain-specific data augmentation. They propose that *"pre-trained models such as BERT have provided significant gains across different NLP tasks."* Radford et al. (2018) as well demonstrate that language models pre-trained with such embeddings can be fine-tuned to perform specific language tasks more effectively.

Mikolov et al. (2013b), Devlin et al. (2019) and Brown et al. (2020) share the finding, that contextualised embeddings produced by models such as BERT and GPT are highly effective and useful in various NLP tasks. Peters et al. (2018) and Ruder et al. (2019) discuss the different approaches of generalisation and specificity in embedding models and explore techniques for adapting pre-trained embeddings to specific tasks or domains.

Many word embeddings are trained on generic text corpora, which may not capture domain-specific terminology and semantics. Some papers examine the advantages of pre-training on domain-specific data for text simplification, including Yeung (2019) and Schomacker et al. (2023). However, Schomacker et al. (2023) have also identified research gaps, such as:

- *"Identification and investigation of existing texts, which are tailored to the needs of the target group and improve the readability of texts both in monolingual and* *parallel datasets."*

- *"Extension of parallel datasets by adding topics, domains and sub-domains, that are relevant for the everyday life of the target group."*
- *"The transferability of the model to domains and sub-domains (e.g., legal sub-domains) for which it has not been trained."*

We hope that our research helps to narrow this gap.

2.4.2 Fine Tuning

Fine Tuning initially involved adjusting pre-trained models on smaller, specific datasets to enhance performance on particular tasks. This approach is effective because it utilises the pre-trained model's general knowledge while adapting to the specific task. This reduces the need for extensive training data and computational resources, resulting in a faster training process. Additionally, it offers the benefits of generalization and adaptability, making it a highly efficient technique. Howard and Ruder (2018) introduced this concept and demonstrated its effectiveness through their work on the Universal Language Model Fine Tuning (ULMFiT) method. As the field advanced, research by Devlin et al. (2019), with the introduction of BERT, showed how deep bidirectional training could further improve the Fine Tuning process, allowing models to better understand context and subtleties in language. Later studies, such as those by Liu et al. (2019) on RoBERTa, optimised these Fine Tuning strategies by exploring different training methodologies, hyperparameters, and larger datasets. The current trend in language models is towards efficiency and adaptability. LLMs such as GPT-3 (Brown et al. 2020) demonstrate how few-shot learning and Prompt Engineering can reduce the need for extensive Fine Tuning. This indicates a shift towards more flexible and generalised approaches to model adaptation.

Garbacea and Mei (2022) suggest that *"when directly adapting a Web-scale pre-trained language model to low-resource text simplification tasks, fine-tuning based methods present a competitive advantage over metalearning approaches."* Moslem et al. (2023) *"emphasise the significance of fine-tuning efficient LLMs [...] to yield high-quality zero-shot translations [...]."* This Fine Tuning can be achieved by few-shot learning, which can enhance LLMs to *"achieve strong performance on many NLP datasets, inclusing translation [...]."* (Brown et al. 2020)

Research agrees, that fine tuning pre-trained models on task-specific datasets can significantly improve their performance for specialised tasks (Howard and Ruder 2018; Devlin et al. 2019). It is widely accepted that preparing and curating datasets tailored to the specific needs of the task at hand is necessary (Dodge et al. 2020). The importance of high-quality and relevant data is consistently emphasised in studies as crucial for successful model adaptation. Although there is agreement on the importance of Fine Tuning, researchers differ on the best methodologies and depth of Fine Tuning required. Some argue for extensive Fine Tuning on large, diverse datasets to capture nuances (Devlin et al. 2019), while others suggest that focused, task-specific Fine Tuning may be more efficient and effective (Howard and Ruder 2018). There is a debate on the most effective approach to Fine Tuning language models. The introduction of models such as GPT-3 (Brown et al. 2020) has sparked this debate, with some questioning on the reliance on Fine Tuning and proposing innovative approaches such as Prompt Engineering.

Current research gaps exist in the specific application of LLMs for translating complex text into Easy Language, particularly with minimal resources. Although Fine Tuning and Prompt Engineering have been well-explored for general tasks, their effectiveness and optimisation strategies for Easy Language translation remain an ongoing topic. Furthermore, there is a shortage of dedicated datasets for Easy Language, which are crucial for effective training and Fine Tuning of LLMs.

2.4.3 Prompt Engineering

Prompt Engineering is a technique used to direct language models towards generating more accurate and efficient outputs. It involves carefully crafting prompts by adding extra information or formulating them in specific ways to better define the task for the model. The effectiveness of Prompt Engineering lies in its ability to reduce ambiguity and provide clear directions to the model, thereby improving the quality of the output. Language models can be highly sensitive to input phrasing, which is why carefully engineered prompts are essential to achieving specific results. Prompt Engineering allows us to fully utilise the potential of a language model and guide it towards meeting our requirements. It has evolved significantly alongside the development of LLMs.

The publication of GPT-3 by Brown et al. (2020) was a significant milestone. The model demonstrated its capacity to perform a broad range of tasks with minimal adjustments, solely through innovative prompt design. This has led to a surge in research focusing on systematic approaches to Prompt Engineering, with the aim of optimising prompt efficiency and effectiveness (Liu et al. 2023). Current trends involve exploring automatic prompt generation and the impact of prompt format and structure on model performance. This highlights a shift from heuristic approaches to more empirical, automated methodologies in crafting effective prompts (Reynolds and McDonell 2021; Wu et al. 2022).

Studies such as Brown et al. (2020) demonstrate the potential of fine-tuned LLMs, such as GPT-3, in understanding and generating text in specific formats. Liu et al. (2023) emphasise the importance of systematically exploring prompting methods. They indicate that the structuring and formatting of prompts significantly impact the performance of LLMs in specialised tasks, including translation. Shin et al. (2020) conducted research on auto-generating prompts, which provides a way to prepare data and improve language models for efficient processing and output of Easy Language translations.

Researchers generally agree on the effectiveness of Prompt Engineering as a powerful tool for eliciting specific responses from LLMs. It has demonstrated its versatility across a range of tasks, including translation (Brown et al. 2020; Liu et al. 2023). It is widely agreed that the design of the prompt has a significant impact on the performance of

LLMs. This underscores the importance of prompt structure and content (Liu et al. 2019). However, there is disagreement regarding the best methods for prompt creation. Some advocate for manual crafting to precisely control model output (Brown et al. 2020), while others see potential in automating prompt generation to efficiently leverage model capabilities across tasks (Shin et al. 2020). There is debate surrounding the scalability of Prompt Engineering, particularly in its application to complex tasks such as translating into Easy Language. Some researchers question whether current methodologies can sufficiently maintain the nuances and simplicity at the same time in translated text. Additionally, there are differing views on whether Prompt Engineering alone can achieve high levels of accuracy and fidelity in task-specific applications without additional Fine Tuning.

There is a lack of systematic studies on the efficacy of Prompt Engineering for specific linguistic simplification tasks (Liu et al. 2023). Despite recent advances, there is still a notable absence of guidelines on crafting prompts that consistently produce Easy Language outputs. Therefore, more focused research in this area is needed. Furthermore, automating the process of generating prompts while ensuring high-quality and contextually appropriate translations in Easy Language has not been fully addressed yet. Existing studies, such as Shin et al. (2020), primarily focus on automation in broader contexts. There is a lack of research on whether manual or automated prompts are more effective for language simplification tasks. This leaves questions about the best practices for Prompt Engineering in this specialised application unanswered. Additionally, the potential of Prompt Engineering in combination with other enhancement methods, such as Fine Tuning with task-specific datasets for Easy Language, remains an underexplored synergy in current literature.

2.5 NLP in Government Communication

In recent years, research has focused on the utilisation of artificial intelligence (AI) in our research fields. The topic has become so important that the European Comission has publishes a paper discussing the *"[a]pplication areas"* for NLP, how to structure an NLP project and which *"[r]elevant examples of NLP usages [exist] in the public sector in Europe"* (Barthélemy et al. 2022). In addition to these opportunities, the European Council has recognised the potential dependence on companies that provide LLMs, as well as the limitations of LLMs due to their *"purely mathematical approach to reasoning.* It is crucial that *"key principles of public administration such as accountability, transparency, impartiality, or reliability"* are thoroughly considered in the integration of NLP into public administration (EU 2023). Digital interaction with the government has also become a very important issue for citizens. So far, that they even work as Open Source communities on projects such as PolicyWeb (2024), which *"aims to develop a large-scale app that serves as an interface for users to engage with the government and influence policy."*

Djeffal and Horst (2021) discuss the potential of AI for inter- and intra-lingual translation. They suggest that automatic translation to Easy Language can serve social and inclusive purposes. Additionally, they address the question of how these techniques can support public administrators without compromising their competences. Deilen et al. (2023) specialise in translating *"citizen-oriented administrative texts into German Easy Language"*. They use *"ChatGPT to translate texts from websites of German public authorities"*. The analysis of *"correctness, readability, and syntactic complexity"* reveals that the texts were simplified but do not meet Easy Language standards. Furthermore, they found, that *"the content of the [translated] texts was not always correct."* Maaß (2020) suggests that Easy Language texts *"also have a symbolic function: They are a token of the inclusion-friendliness of the [..] public authorities [...] that publicly display such texts, for example on their websites. The symbolic function should, however, be executed in a way that facilitates acceptability by the majority society."* Otherwise they risk *"resistance and rejection with regard to communication impairments and making it visible through publicly displayed accessible communication products."*

The papers we reviewed in this study examined the use of NLP in government communication from different angles. They all agree that NLP can be a valuable tool for governments if used effectively. Additionally, all papers discussing LLM text simplification acknowledge that texts are simplified but do not meet Easy Language standards. For instance, Kopp et al. (2023) attribute these inadequate translations to the *"lack of training data"*. Although *"algorithm-based alignment suggestions"* can assist the alignment of translations, they still require a significant amount of manual work.

Kühnhenrich and Michalik (2020) demonstrate that governmental services should not only enhance their digital accessibility but also present information in a way that is easily understandable for citizens. This requirement is inherent in the definition of E-Governance itself. As Ghosh (2009) states: *"E-governance is the public sector's use of information and communication technologies (ICT) with the aim of improving information and service delivery, encouraging citizen participation in the decision-making process and making government more accountable, transparent, and effective."* Easy Language (NLS 2022) is a method of presenting information more coherently through text simplification, which can be seen as a form of intralingual translation. Gille et al. (2023) note that there is currently a lack of appropriate parallel corpora for training LLMs to achieve automated text simplification. While their research ideas overlap with ours, they take a more theoretical approach and do not provide a detailed analysis of the various enhancement methods for LLMs, let alone a comparison.

2.6 Related Work Conclusion

Several papers have explored text simplification in the context of the German (Easy) Language, particularly in government settings. Klaper et al. (2013) as well as Toborek et al. (2023) concentrate on developing a sentence-aligned parallel corpus and the algorithm required for this task. Ebling et al. (2022) share the same focus but enhance their research with a neural automatic translation approach for different levels of simplified German and performance improvements. Schomacker (2023) discusses involving the target group in creating and evaluating an Easy Language dataset. Additionally, the author focuses on the Easy Language rules and the requirements they derive for automatic text simplification. Anschütz et al. (2023) demonstrate that Fine Tuning an LLM on a corpus of German Easy Language adapts the LLM *"to the style characteristics of Easy Language and output[s] more accessible texts."* The *"results indicate that pre-training on unaligned data can reduce the required parallel data while improving the performance on downstream tasks."*

Although these papers focus on automated text simplification in the context of LLMs, none of them compare different LLM enhancement methods and their respective impact on the quality of text simplification for Easy Language. Therefore, they do not provide a definitive answer regarding the best method for enhancing LLMs in the context of text simplification.

Chapter 3

Data Collection

To gain a better understanding of real-life applications, we met with the web editor responsible for all easy language translations in the City of Munich's domain, as well as the administrative team behind MUCGPT. Currently, each translation has to be written individually. Automated text simplification could significantly improve speed and translation diversity, as an LLM could propose multiple translation suggestions. However, the current LLM prototype, MUCGPT, lacks the necessary capabilities to deliver translations up to our standard. Therefore, it requires improvement. This leads us to the central question of this thesis:

Which technique is most suitable for improving an LLM's ability to translate complex text into Easy Language?

The proposition regarding the best enhancement method to address the research topic must be deduced from collecting and analysing data. In the following, we cover the first research question *"How does the existing public data need to be prepared for this purpose?"*. We will outline the selected data and metrics, the processing steps involved and how they contribute to providing qualitative research data.

3.1 Language Model

The language model serves as the foundation for all enhancement methods. The thesis follows a pragmatic philosophy based on our real-life use case. Initially, several LLMs were used to produce preliminary translations without any enhancements, enabling a review of their core translation quality (Chapter 4). Although none of the results met Easy Language standards, we selected GPT-3.5-Turbo-1106 as the foundational LLM for several reasons:

- Our company's prototype, MUCGPT, is based on this language model.
- It provided superior support for enhancing possibilities, including production ready Fine Tuning and Embedding APIs.
- It produced the most coherent results among the LLMs that were tested.
- At the time of research, this version represented the latest freely available version with the highest level of public interest. We hope this provides an easy entry point for possible future research building upon this thesis.

GPT-3.5-Turbo-1106 has a context window of 16,385 tokens and was trained on data up to September 2021 (OpenAI). For the sake of brevity, the GPT-3.5-Turbo-1106 model will be referred to as GPT throughout the thesis.

3.2 Parallel Corpora

Our parallel corpora are derived from normal/Easy Language website pairs from the City of Munich. The URLs and additional examples of our parallel corpora can be found in Appendix A and B to offer a more comprehensive understanding. All texts adhere to the NLS (2022) standards and have been written, revised and approved by certified translators and auditors (LHM b). However, these websites were not composed in a standardised format and cannot be processed automatically. As a result, the base texts required manual transformation to serve as translation input, evaluation reference data and data for the enhancement methods. In base text transformation, a distinction is made between page pairs (Section 3.2.1) and text pairs (Section 3.2.2). The quality of our parallel corpora is limited because it was not curated by a certified editor.

In the subsequent sections, we will provide examples from the page *"Abmeldung eines Hundes"* and its Easy Language counterpart *"Hunde-Steuer abmelden"* to illustrate our parallel corpora in more detail.

3.2.1 Pages

The public websites were downloaded using the command line. Subsequently, all code was removed from the pages, leaving only the raw text. Each website is represented as an individual document. These documents were then converted into multi-level vectors, via the GPT embedding API, to serve as input data for Embedding (Section 4.3). Pairs of normal and Easy Language websites are also represented as pairs in the page- and vector-file pairs. In Listing 1, the vector dimensions were reduced from 400 to 6 to improve overall text spacing of the thesis and comprehensibility of the example. The example depicts the page *"Hunde-Steuer abmelden"* in vector format. This visualisation aims to enhance the reader's understanding of vectorisation, which is necessary basic knowledge for delving into Embedding in Section 4.3.

3.2.2 Pairs

170 pairs of normal/Easy Language translations were compiled manually from those pages. The text pairs range from single words to paragraphs, aiming to depict as diverse a picture as possible. The length of each pair and the pairing of easy and normal text is determined by the content of each page that best matches its counterpart. The provided information serves as the basis for Fine Tuning (Section 4.4), translation (Section 3.3) and translation evaluation references (Section 3.4). Fine Tuning required for transforming the pairs into JSONL format. It is important to note that the pair has to be presented in the user and assistant context, representing the original text and its expected simplified translation for the Fine Tuning API.

```json
{
   "object": "list",
   "data": [
      {
         "object": "embedding",
         "index": 0,
         "embedding": [
            -0.005888141,
            0.0074991477,
            -0.018609783,
            0.028877739,
            -0.019020502,
            -0.021810552
         ]
      }
   ],
   "model": "text-embedding-ada-002-v2",
   "usage": {
      "prompt_tokens": 9,
      "total_tokens": 9
   }
}
```

Listing 1: A simplified example of page vectorisation in JSON format.

Normal Language Wenn Ihr Hund gestorben ist, Sie ihn abgeben oder Sie aus München wegziehen, müssen Sie Ihren Hund innerhalb von zwei Wochen von der Hundesteuer abmelden.
Easy Language Ihr Hund ist gestorben? Sie haben Ihren Hund abgegeben? Sie ziehen mit Ihrem Hund aus München weg? Dann müssen Sie Ihren Hund abmelden.
JSONL Pair `{"role": "user", "content": "Wenn Ihr Hund gestorben ist, Sie ihn abgeben oder Sie aus München wegziehen, müssen Sie Ihren Hund innerhalb von zwei Wochen von der Hundesteuer abmelden."}, {"role": "assistant", "content": "Ihr Hund ist gestorben?\nSie haben Ihren Hund abgegeben?\nSie ziehen mit Ihrem Hund aus München weg?\nDann müssen Sie Ihren Hund abmelden."}`

Table 3.1: An example pair of our parallel corpora. The two counterparts from the public websites are represented in JSONL format. This JSONL text will be used as data for our experiments.

3.3 Translations

As a first step, we compiled a comprehensive table file containing the base text, the human simplification reference, GPT baseline translations and all enhancement translations. This translation library serves as the single point of truth for all metric calculations (Section 3.4). Tables 3.2 - 3.4 present an example of one of the 170 rows, providing insight into the actual implementation of this thesis and the differences in the results obtained by different methods. To improve formatting, the translation examples had to be divided into three parts. Chapter 5 provides a detailed analysis of the results.

Base Text	Die Gleichstellungsstelle wirkt auf die Gleichstellung von Frauen und Männern in Beruf, Familie und Gesellschaft hin.
Human Translation	Die Gleichstellungs-Stelle ist für Frauen in München da. Und auch für Mädchen.
GPT	Die Gleichstellungsstelle arbeitet daran, dass Frauen und Männer die gleichen Rechte und Chancen in ihrem Beruf, ihrer Familie und der Gesellschaft haben.

Table 3.2: An example row of our table containing the base text, the human and the GPT baseline translations.

EM	Die Gleichstellungsstelle hilft dabei, dass Frauen und Männer die gleichen Chancen haben. Das gilt für den Beruf, die Familie und die Gesellschaft.
FT	Gleichstellungs-Stelle ist ein Amt. Ein Amt gehört zu einer Stadt oder einem Land. Die Gleichstellungs-Stelle kümmert sich um Frauen. Und die Gleichstellungs-Stelle kümmert sich um Mädchen. Die Gleichstellungs-Stelle hilft ihnen.
PE	Die Gleichstellungsstelle arbeitet daran, dass Frauen und Männer in Beruf, Familie und Gesellschaft die gleichen Chancen haben.

Table 3.3: An example row of our table containing the standalone methods' translations.

EM-FT	Die Gleichstellungs-Stelle hilft Frauen. Und die Gleichstellungs-Stelle hilft auch Männern. Die Gleichstellungs-Stelle hilft bei vielen Dingen. Zum Beispiel: * Wenn Frauen oder Männer Hilfe brauchen. * Wenn Frauen oder Männer eine Beratung brauchen. * Wenn Frauen oder Männer eine Anlauf-Stelle brauchen.
EM-PE	Die Gleichstellungs-Stelle hilft Frauen und Männern. Damit Frauen und Männer die gleichen Rechte haben. Und die gleichen Chancen. In der Familie. Im Beruf. In der Gesellschaft.
FT-PE	Die Gleichstellungs-Stelle arbeitet für die Gleich-Stellung von Frauen und Männern. Das bedeutet: Frauen und Männer sollen die gleichen Rechte haben. Und die gleichen Chancen. In der Arbeit. In der Familie. Und in der Gesellschaft.
EM-FT-PE	Die Gleichstellungs-Stelle arbeitet für die Gleich-Stellung von Frauen und Männern. Das bedeutet: Frauen und Männer sollen die gleichen Rechte haben. Und die gleichen Chancen. In der Arbeit. In der Familie. Und in der Gesellschaft.

Table 3.4: An example row of our table containing the combined methods' translations.

3.4 Metrics

To answer the second research question, *"How does the existing language model need to be enhanced with the prepared data?"*, we need to provide evaluation data of various translations from the selected enhancements (Section 2.4). Hence, we have selected the metrics outlined in this chapter for this purpose.

Translation into Easy Language follows a dedicated set of rules and so does its evaluation (NLS 2022, pp. 61 - 67). Therefore, certified human evaluation against these rules would be the ideal metric for our comparison. However, after soliciting four different quotations, we realised that they exceed both our time limit and budget. Thus, this thesis is limited to the automatic evaluation of text simplification.

As stated in Section 2.3, the assessment of text simplification covers multiple categories. Consequently, to make realistic assessments, it is necessary to apply different metrics that cover these categories (Table 3.5).

All 170 examples from the muenchen.de pages (Appendix A) were translated using each of the enhancement methods. Afterwards, these translations were evaluated using the metrics introduced in this chapter.

It is important to note that readability indices provide estimations and are not absolute measures of difficulty. They serve as useful tools for assessing text complexity and tailoring written content to specific audiences. Therefore, we selected a combination of eight metrics spanning the three categories introduced to obtain a comprehensive evaluation. It is also important to note that certain metrics fall under multiple categories as they cover aspects of more than one simplification category (Section 2.3).

	Direct Assessment	**Simplicity Gain**	**Structural Simplicity**
BERTScore	X	X	
FRE			X
LIX			X
METEOR	X	X	
SARI	X	X	
SNLR			X
TER		X	
TTR			X

Table 3.5: The selected metrics and their corresponding categories.

All metrics calculate results mathematically. However, BERTScore, SARI and TER differ from the other metrics as they also consider reference translations. FRE, LIX, SNLR and TTR focus solely on the translation.

FRE and LIX share many similarities. Both metrics focus on text complexity and readability, evaluating factors such as word frequency, word length and sentence length. However, they differ in that FRE is the only one to consider word frequency and was specifically developed for the German language. Additionally, they favoured different enhancements. Therefore, both metrics are examined in this paper despite their similarities.

FRE, LIX and TER take sentence splitting into account. METEOR may indirectly consider sentence splitting if there are differences in sentence boundaries between the translation and the reference. SNLR focuses on the ratio of sentences to newlines in a neutral manner.

3.4.1 BERTScore

BERTScore is the only metric that uses contextual embeddings to determine token similarity instead of relying on exact matches. It *"computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings."* Zhang et al. (2019). *"The metric provides three approaches: BERTScoreRecall, BERTScorePrecision, and BERTScoreF1. The first match each token in the reference sentence to its most similar in the system output, while the second matches the opposite (output to reference). The third combines the two like a typical equally weighted F1 score."* (Beauchemin et al. 2023). BERTScore provides insights into semantic similarity and the quality of the translation. We have incorporated this metric to enrich the diversity of perspectives in our evaluation. BERTScore is calculated as follows:

$$\begin{aligned}[H]R_{BERT} &= \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^{\intercal} \hat{x}_j \\ P_{BERT} &= \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^{\intercal} \hat{x}_j \\ F_{BERT} &= 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}\end{aligned} \tag{3.1}$$

where the *"complete score matches each token in x to a token in $\hat{x}$ to compute recall, and each token in $\hat{x}$ to a token in x to compute precision. [...] We combine precision and recall to compute an F1 measure."* (Zhang et al. 2019).

The scores range from 0 to 1, with 1 indicating a perfect match between the two texts. A **higher score** suggests **greater similarity** between two texts based on their contextualised embeddings generated by a BERT model.

3.4.2 FRE

The "Flesch Lesbarkeitsindex (Flesch Reading Ease)", based on Flesch (1948), is a metric used to evaluate the readability of German text. It assesses the structural simplicity of a text based on sentence length and the number of syllables per word. This metric is particularly useful for evaluating the readability of simplified texts. FRE is calculated as follows:

$$FRE = 180 - (ASL \times 1.5) - (ASW \times 58.5) \tag{3.2}$$

where:

$$\begin{aligned}ASL &= \text{average sentence length in words} \\ ASW &= \text{average number of syllables per word}\end{aligned}$$

FRE scores range from 0 to 100, with **higher scores** indicating **easier readability**:

Score	FRE Readability
90 - 100	very easy
80 - 89	easy
70 - 79	fairly easy
60 - 69	standard
50 - 59	fairly difficult
30 - 49	difficult
00 - 29	very difficult

Table 3.6: FRE scores and their text complexity classification.

3.4.3 LIX

The "Lesbarkeitsindex (LIX)" (Björnsson 1968) is a readability metric that evaluates the structural simplicity and readability of German text. It considers the average sentence length and the percentage of long words. It is calculated as follows:

$$LIX = \frac{W}{S} + \frac{CW \times 100}{W} \tag{3.3}$$

where:

$$\begin{aligned} W &= \text{the number of words} \\ S &= \text{the number of sentences} \\ CW &= \text{the number of long words (more than six letters)} \end{aligned}$$

Contrary to FRE, a **lower** LIX **value** indicates a **lower difficulty** level.

Score	LIX Readability
$LIX < 25$	very easy (for children)
$25 \leq$ LIX < 35	easy
$35 \leq$ LIX < 45	standard
$45 \leq$ LIX < 55	difficult
$55 \leq LIX$	very difficult

Table 3.7: LIX scores and their text complexity classification.

3.4.4 METEOR

The "**M**etric for **e**valuation of **t**ranslation with **e**xplicit **or**dering (METEOR)" offers holistic evaluations of text quality, taking into account various linguistic features, alignment with the reference text and content overlap. It provides insights into the adequacy, fluency and informativeness of the system-generated text in comparison to the reference text. *"METEOR [...] is based on a generalized concept of unigram matching between the machine produced translation and human-produced reference translations. [...] METEOR computes a score for this matching using a combination of unigram-precision,* *unigram-recall, and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the machine translation are in relation to the reference."* (Banerjee and Lavie 2005).

It is calculated as follows:

$$\begin{aligned}\text{METEOR} &= \frac{P \times R}{(1-\alpha) \times R + \alpha \times P} \\ \text{Precision} &= \frac{\text{exact matches} + \beta \times \text{fragmentation penalty}}{\text{number of words in the system-generated translation}} \\ \text{Recall} &= \frac{\text{exact matches} + \beta \times \text{fragmentation penalty}}{\text{number of words in the reference translation}}\end{aligned} \tag{3.4}$$

where:

$$\begin{aligned}P &= \textit{precision} \\ R &= \textit{recall} \\ \alpha &= \text{tunable parameter (usually set to 0.5)} \\ \beta &= \text{tunable parameter (usually set to 0.5)}\end{aligned}$$

METEOR scores range from 0 to 1, with 1 representing a perfect match between the system-generated and the reference translation. Therefore, a **higher** METEOR **score** suggests that the system-generated translation is closer in terms of precision and recall to the reference translation, indicating **higher** translation **quality**.

3.4.5 SARI

SARI (Xu et al. 2016) *"compares **s**ystem output **a**gainst **r**eferences and against the **i**nput sentence. It explicitly measures the goodness of words that are added, deleted and kept by the systems. [...] Together, in SARI, we use arithmetic average of n-gram precisions $P_{operation}$ and recalls $R_{operation}$".*

It is specifically designed to assess text simplification and incorporates additional considerations specific to it, such as simplicity and appropriateness, which are not addressed by metrics such as BLEU. SARI *"achieve[s] a much better correlation with humans in simplicity judgment[, than existing metrics (i.e. FK, BLEU, iBLEU)], while still capturing the notion of grammaticality and meaning preservation."* Furthermore, Sulem et al. (2018c) have shown *"that BLEU is not suitable for the evaluation of sentence splitting, the major structural simplification operation.* Moreover, SARI outperforms all BERTScore variants in terms of Simplicity Gain (Alva-Manchego et al. 2021). We therefore prefer SARI to the available BLEU metrics.

SARI is calculated as follows:

$$\begin{aligned} \text{SARI} &= d_1\text{Fadd} + d_2\text{Fkeep} + d_3\text{Pdel} \\ &\text{where } d_1 = d_2 = d_3 = \frac{1}{3} \text{ and} \\ P_{\text{operation}} &= \frac{1}{k}\sum_{n=1}^{k} p_{\text{operation}}(n) \\ R_{\text{operation}} &= \frac{1}{k}\sum_{n=1}^{k} r_{\text{operation}}(n) \\ F_{\text{operation}} &= \frac{2 \times P_{\text{operation}} \times R_{\text{operation}}}{P_{\text{operation}} + R_{\text{operation}}} \\ &\text{operation} \in \{\text{del, keep, add}\} \end{aligned} \tag{3.5}$$

where k is the highest n-gram order and set to 4 in our experiments." (Xu et al. 2016)

SARI provides a direct measure of the improvement in the simplicity of the text, ranging from 0 to 1, where 1 indicates the highest similarity to the references. A **higher** SARI **score** means that the simplified text is **more similar** to the references, which in our case indicates greater simplicity.

3.4.6 SNLR

The current metrics may not accurately reflect the quality of translations, as they ranked Embedding and Prompt Engineering equally overall in our tests. However, Embedding did not follow basic NLS (2022) rules in many of its translations, such as including a newline for each sentence. In our view, the most distinguishable difference at first sight between Easy Language and everyday language is the impact of this rule on text structure. Therefore, we propose a new simple metric called the '**S**entence **N**ew-**L**ine **R**atio'.

$$\textbf{SNLR} = \min\left(\frac{N_{\text{newline}} + 1}{N_{\text{sentence}}}, 1\right) \tag{3.6}$$

where

- N_{newline} represents the quantity of newlines present in the text.
  - Adding one newline compensates for missing newline characters at the end of the text, which would otherwise falsely decrease the score.
- N_{sentence} represents the quantity of sentences present in the text.
- The min function is used to enforce an upper limit of 1.
  - This is because additional newlines added for formatting purposes are not considered, as it is impossible to evaluate whether these additional newlines are positive or negative.

If there are no sentences (N_{sentence}), SNLR is considered undefined or zero to prevent division by zero. The scores range from 0 to 1, with 1 representing a perfect score. A **higher** SNLR **score** indicates that the number of newlines is closer to the number of sentences, which suggests **better translations** in terms of Easy Language.

3.4.7 TER

Snover et al. (2006) define that *"**T**ranslation **E**dit **R**ate (TER) measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation."* They have shown *"that TER is adequate for research purposes as it correlates reasonably well with human judgments [...]."* While primarily used to evaluate machine translation, it can also provide insight into the structural changes introduced during text simplification, including sentence splitting and merging. TER helps to assess structural changes and alignment with the reference text. When calculating the score, TER takes into account both the number of changes and the length of the reference text.

We selected TER because it provides a simple and straightforward measure of translation quality, allowing us to obtain interpretable scores quickly. However, by focusing directly on the edits required to align the translation with the reference, it offers a unique approach.

and is calculated as follows:

$$TER = \frac{I + D + S}{W} \tag{3.7}$$

where:

$$\begin{aligned} I &= \text{Insertions: Words in the machine-generated translation} \\ &\quad \text{that are not in the reference translation.} \\ D &= \text{Deletions: Words in the reference translation} \\ &\quad \text{that are not in the machine-generated translation.} \\ S &= \text{Substitutions: Words that are different between the machine-generated} \\ &\quad \text{and reference translations.} \\ W &= \text{Words: Total number of words in the reference.} \end{aligned}$$

A **lower** TER **score** indicates a **better translation**, as it means that fewer edits are required to align the machine-generated translation with the reference translation.

3.4.8 TTR

"**T**ype **T**oken **R**atio (TTR)" (Vajjala and Meurers 2012) represents an even more sober calculation of textual simplicity than TER. TTR measures lexical diversity and vocabulary richness in the text, indicating the variety of vocabulary used in the simplified text. It is valuable for understanding the impact of vocabulary simplification on text comprehension.

In this thesis, we use TTR instead of STTR because standardisation of TTR did not significantly affect the average score.

As mentioned above, a certified full evaluation of NLS (2022) rules is not feasible for this thesis. Therefore, we have used this metric to provide an exemplary assessment of how well translations comply with NLS (2022) rule W4, “Always use the same words for the same things.“

TTR is calculated using the formula:

$$TTR = \frac{TY}{TO} \tag{3.8}$$

where:

$$\begin{aligned} TY &= \text{number of types} \\ TO &= \text{number of tokens} \end{aligned}$$

A higher TTR score indicates greater lexical diversity, suggesting a broader vocabulary in the text. Conversely, a **lower** TTR **score** indicates repetitive language or a more limited vocabulary, which is consistent with our goal of text simplification.

3.5 Metric Scores

All metrics were calculated using Python scripts. The results were consolidated in tabular form. For each metric, all data was collected in a ‘master table’ to facilitate better comparison of results and to enable the creation of helpful visualisations, such as various graph formats.

Figure 3.1 and Figure 3.2 are examples of such visualisations. These box-plots allow us to compare the distribution of scores for each Enhancement Method and draw conclusions based on this comparison. In these plots, the data for each Enhancement Method is divided into four equal parts, where

- the median 50% of the scores, depicted as a box, representing the interquartile range (IQR) between Q1 (lower quartile) and Q3 (upper quartile).
- the median, depicted as a horizontal line inside the box, dividing the data equally into upper and lower halves.
- the lower and upper whiskers, which typically represent minimum and maximum values within the range of 1.5 times the IQR.
- individual data points that fall significantly outside the overall pattern of the data (outliers), which are marked as dots.

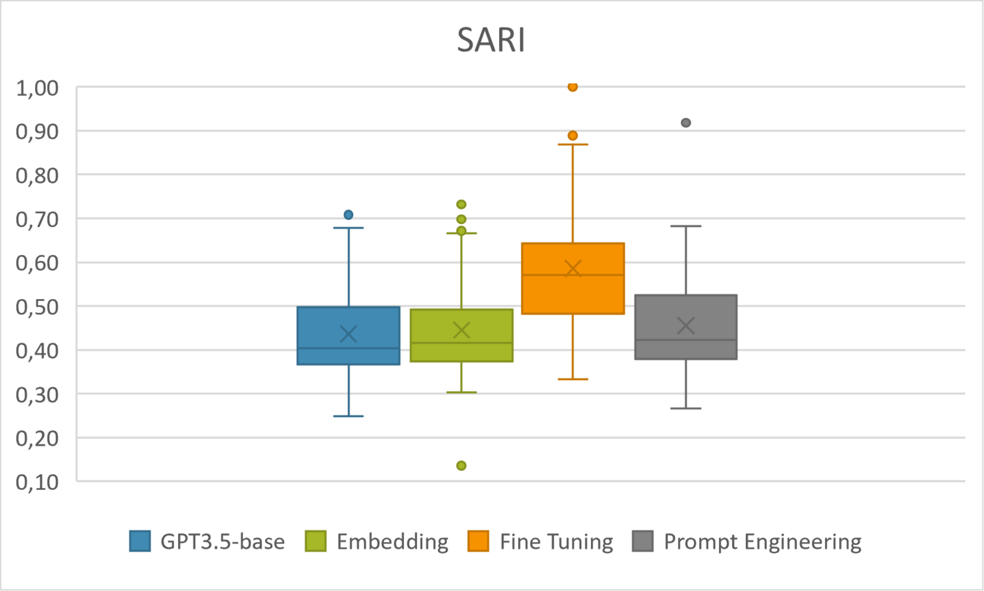


Figure 3.1: The box plots show the SARI scores achieved by the standalone enhancement methods.

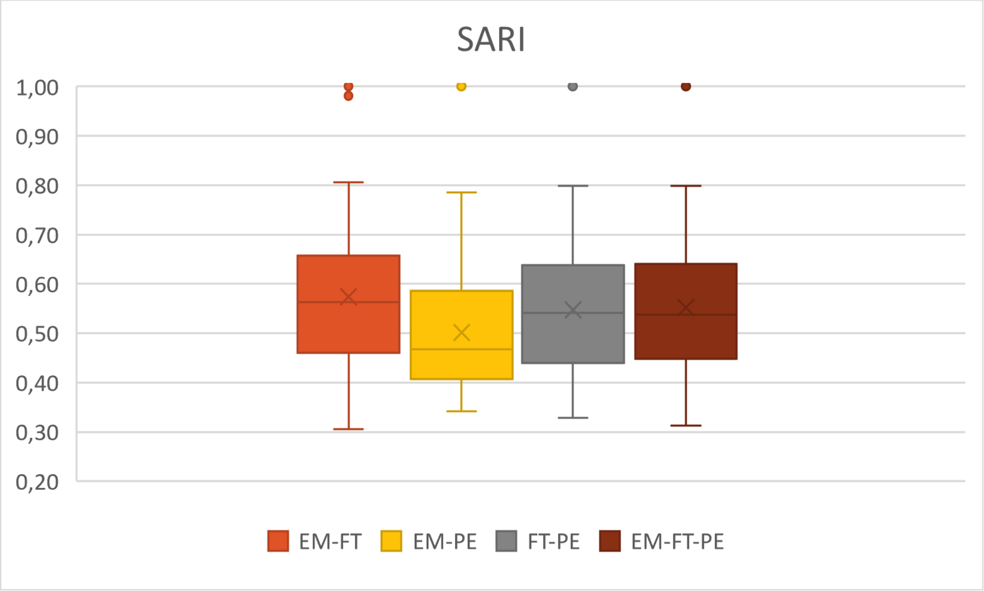


Figure 3.2: The box plots show the SARI scores achieved by the combined enhancement methods.

Chapter 4

Experimental Studies

This chapter addresses the second research question: *"How does the existing language model need to be enhanced with the prepared data?"*. To address this question, we performed experiments using the following enhancement methods and their combinations: Embedding, Fine Tuning, and Prompt Engineering.

Although there are other methods and techniques for enhancing LLMs, such as ensemble methods (Mohammed and Kora 2023), knowledge distillation (Gou et al. 2021) and multi-task learning (Chen et al. 2021), we have chosen the most prominent and well-established ones due to their effectiveness, versatility and widespread adoption in both research and practical applications.

Due to the limited availability of suitable parallel corpora pages and pairs (as discussed in Section 3.2), we are working within the realm of few-shot learning. According to Brown et al. (2020), few-shot learning involves providing a model with *"a few demonstrations of the task at inference time as conditioning"*. These demonstration examples *"typically [have] a context and a desired completion"* (Section 3.2.2). *"[F]ew-shot works by giving K examples of context and completion, and then one final example of context, with the model expected to provide the completion"*. They propose that *"the main advantage of few-shot is a major reduction in the need for task-specific data."*

The translations used in the selected metrics (Section 3.4) were generated by the experiments conducted (Section 3.3).

4.1 Base Translations

The initial stage involved acquainting ourselves with the present state of LLM German Easy Language text simplification and determining whether additional experiments were required. The initial untrained translations (Table 4.1) were generated using the following example prompt, which was derived from the page 'Personal-Ausweis beantragen' (Appendix A):
*"Bitte in Leichte Sprache übersetzen: 'Deutsche Staatsbürger*innen sind verpflichtet, einen gültigen Personalausweis (oder Reisepass) zu besitzen, sobald sie 16 Jahre alt sind.' "*

The translation results were not suitable for productive use as they either distorted or lost the meaning of the original text, or lacked key elements of Easy Language. Table 4.2 presents the real-life counterpart of the example prompt, highlighting the significant disparity between the untrained translations and the original text.

Huggingface/tum-nlp/germant-gpt_easy Der Personalausweis ist in Bayern gültig und kostet 18 Euro. Die Personal-Karten sind ausgestellt
Huggingface/abacusai/Smaug-34B-v0.1 In einfachen Worten: Als deutscher Bürger, wenn du 16 Jahre alt bist, musst du einen gültigen Ausweis haben.
deepinfra.com/mistralai/Mistral-7B-Instruct-v0.1 Wenn man deutscher Staatsbürger*in ist, muss man sich 16 Jahre alt sein, um einen gültigen Personalausweis (oder Reisepass) zu haben.
llama2.ai/llama-2-70B Deutsche Bürger müssen einen gültigen Personalausweis (oder Reisepass) besitzen, wenn sie 16 Jahre alt werden.
GPT-3.5-Turbo In Deutschland müssen Leute, die 16 Jahre alt sind und die Staatsbürgerschaft haben, einen gültigen Personalausweis oder Reisepass besitzen.

Table 4.1: Untrained translations from different Large Language Models.

Although ChatGPT achieved the best baseline results among the LLMs tested, it lacked certain key features, such as a new line for each sentence (NLS 2022, G5) or the separation of compound words, for example, 'Personal-Ausweis' (NLS 2022, W5). This confirms the need to investigate the following enhancement methods in order to evaluate and compare their potential improvements to the translations.

In the upcoming chapters, Figure 4.2, Figure 4.3 and Figure 4.5 schematically demonstrate their respective processes. They are not intended to be technically accurate, but to enhance the reader's understanding of the process. Although they use the same colour palette, each figure represents a separate process without access to the other's data and functions.

Human Translation
Sie sind Deutscher? Wenn Sie 16 Jahre alt werden: Dann müssen Sie einen Personal-Ausweis haben. Das schwere Wort dafür ist: Personal-Ausweis-Pflicht. Es gibt eine Ausnahme: Wenn Sie schon einen Reise-Pass haben. Dann müssen Sie keinen Personal-Ausweis haben.

Table 4.2: The human translation corresponding to the untrained translations from Table 4.1.

4.2 Easy Language Rules

People with learning-disabilities, dementia or non native speakers can have difficulties understanding difficult language, such as foreign words, technical terms or long sentences (NLS 2022). The 'Netzwerk Leichte Sprache' was established in 2006. The German association emerged from a movement called 'people first'. This movement, along with the NLS, fights for the integration of people with disabilities mentioned before (NLS 2024). They designed a set of rules to guide the creation of German texts that are easy to understand. This type of language is called Easy Language. The rules consist of six sets of guidelines. There are

- twelve rules for word usage,
- nine rules on how to use numbers and symbols,
- four rules on sentence structure,
- four rules on text style,
- seventeen rules on everything from font over typeset, to paper type and contrast and picture design
- and also one rule and some helpful suggestions on proofreading.

In our thesis, we apply the first five sets of rules. We do not include rules about paper weight or image design, as they are not applicable to our LLM. Our aim is to enhance the LLM's overall understanding of the desired output.

The 46 rules needed to be processed to make their information available in a suitable data format (JSONL) for enhancing LLMs and evaluating their results. Subsequently, the lines were merged into a unified document (Appendix C), which was then included in the translation prompt (Table 4.6). Table 4.3 shows one of these rules and its JSONL pendant as an example.

NLS rules format
Benutzen Sie einfache Wörter. Schlecht: genehmigen Gut: erlauben
JSONL format
`{"role": "user", "content": "Benutzen Sie einfache Wörter.\nBeispiel\nSchlecht: genehmigen\nGut: erlauben"}`

Table 4.3: An Easy Language rule represented in JSONL format as used in our experiments.

4.3 Embedding

"Embedding-based retrieval (EBR) differs from plain text search in that, it transforms documents into vectors and maps them to a vector space. This allows similar documents

to have a closer distance in the vector space, while dissimilar documents have a greater distance. This enables similarity to be determined by calculating the distance between documents and search content, thus completing the retrieval task.
To be prepared for search content with relativeness, the vector representations of documents are generated and stored in a vector database. Then, query the database with some general words, for instance: 'appearance, identification'. This query should perform a k-nearest neighbour (kNN) search and return multiple most likely results with their distances to the query words. However these results may contain unrelated contents, because there is no such standard distance to filter them." (Peng et al. 2023)

Figure 4.1 illustrates a simplified vector space to aid the reader's comprehension. In reality, vector spaces can have hundreds of thousands of dimensions, allowing for almost infinite categories of content. This example demonstrates how vectors of similar text-pages (Listing 1), containing the categories 'administrative act', 'year' and 'event', are arranged closely to each other in the vector space. Thus, the similarity search (U5, Figure 4.2) can identify the required information to be retrieved.

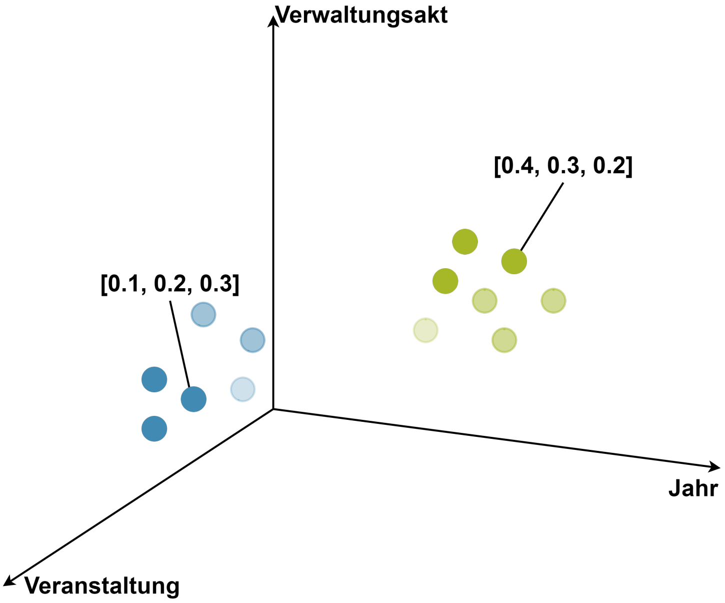


Figure 4.1: A simplified vector space with three dimensions. Similar texts generate vectors that are positioned closer together. This is also indicated by giving them the same colour. Vectors 'farther away' from the reader are displayed with a lighter opacity.

The processes "**E**mbedding 1-3" and "**U**ser 1-6", shown in Figure 4.2, have to be orchestrated by a central instance, such as for example a python script.

The diagram illustrates parallel corpora page pairs (Section 3.2.1), in normal (D) and Easy (D EL) Language, being sent to the Embedding API (E1) as documents. These document pairs represent prompts and their completions. The API then transforms these documents into high-dimensional vectors (E2), which are saved to a vector DB

(E3).

The user submits a request for Easy Language translation to the LLM (U1). The request is then transformed into a vector (U2), which is used to retrieve the most similar content in terms of embeddings (U3). The LLM uses the retrieved similar contents, combined with the LLM's knowledge acquired during its pre-training phase, to generate a translation response (U4). This ensures that the responses generated maintain grammatical structure and semantic coherence, even when translating complex text into Easy Language. Finally, the system returns the response to the user (U5). Due to the significance of shaping the response based on retrieved content, this enhancement is also known as '**R**etrieval **A**ugmented **G**eneration (RAG)' (Lewis et al. 2020).

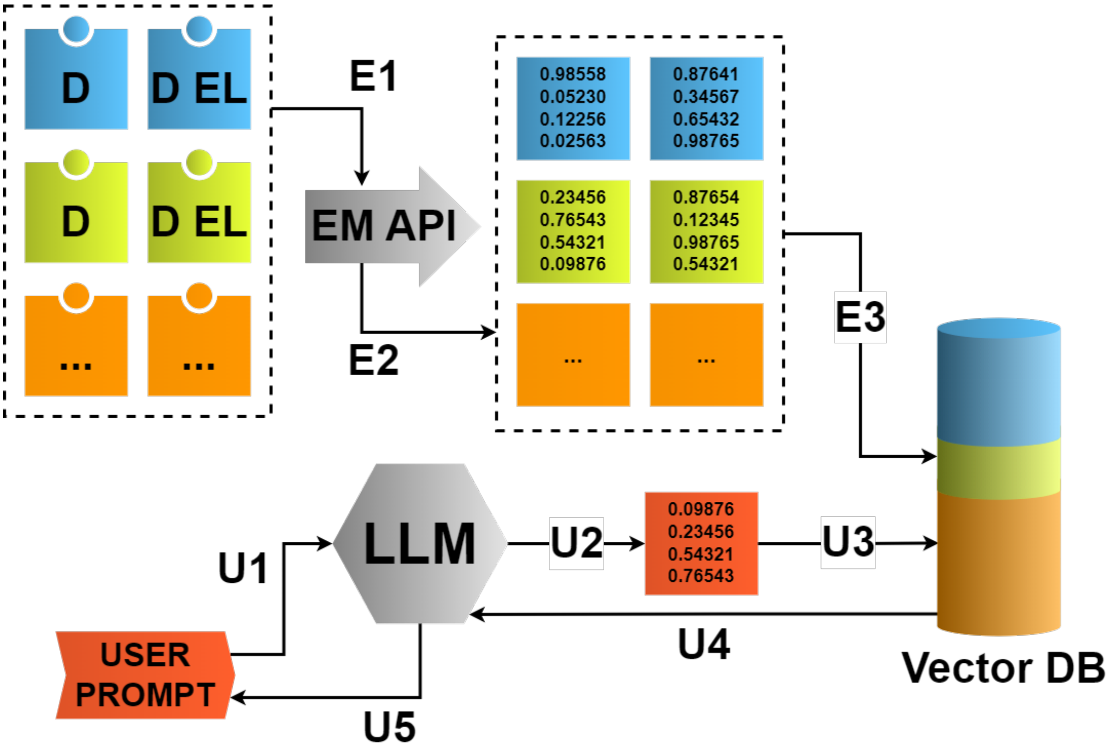


Figure 4.2: The Embedding process simplified. E*-steps display the embedding process, while the U*-steps display the user process. Corresponding normal/ Easy Language pairs share the same colour. Vectors are shown with reduced dimensions to improve visualisation.

We examined the embedding of pages as well as the embedding of pairs. As there was no significant difference in the results, we selected page-based vectors as the basis for Embedding, because page-based Embedding maintains the contextual integrity of the content by treating the entire page's information as a unified entity. This approach can capture subtleties and thematic cues that may be lost when breaking down the content into isolated text pairs, potentially allowing for a more nuanced understanding and processing of the language simplification task.

In simple terms, Embedding can be thought of as providing a comprehensive library of domain-specific documents to your language model, enabling it to search for content similar to the request to generate appropriate responses.

4.4 Fine Tuning

“When applying LLMs in practical applications, such as education, law, and medicine, fine-tuning LLMs with domain-specific data can be essential. Fine-tuning can enrich LLMs with domain knowledge, enhance their specific ability, improve the fairness and reliability of the outputs, and prevent certain damage caused by hallucination (Ji et al. 2023). However, fine-tuning LLMs entails a high demand for computational resources and a substantial amount of domain data that may not be shareable due to privacy concerns.” (Kuang et al. 2023)

Domain-adaptive Fine Tuning was applied, which involves fine tuning the model on data from the target domain, even if it differs from the domain it was pre-trained on. This approach helps the model adapt to the specific characteristics of the target domain, ultimately improving its performance on tasks within that domain.

To train an LLM on domain-specific data, the data must first be prepared in a format suitable for the LLM. The example pair (Table 4.4) of our parallel corpora (Section 3.2.2) was therefore restructured to JSONL format.

JSONL User/Assistant Pair

```
{"role": "user", "content": "Wenn Ihr Hund gestorben ist, Sie ihn abgeben
oder Sie aus München wegziehen, müssen Sie Ihren Hund innerhalb von zwei
Wochen von der Hundesteuer abmelden."}, {"role": "assistant", "content":
"Ihr Hund ist gestorben?\nSie haben Ihren Hund abgegeben?\nSie ziehen mit
Ihrem Hund aus München weg?\nDann müssen Sie Ihren Hund abmelden."}
```

Table 4.4: An example Fine Tuning parallel corpora pair in JSONL format.

Figure 4.3 displays these parallel corpora pairs being sent to the LLM’s Fine Tuning API. Fine Tuning occurs over a variable number of iterations known as ‘epochs’. One epoch refers to one complete pass through the entire training dataset. In this case, we fine tune over three epochs (E1, E2, E3). During each epoch, the LLM aquires new data that it was previously unable to comprehend. Once the Fine Tuning process is complete, the LLM’s knowledge is enriched with domain-specific data. This data can then be utilised by the LLM to process subsequent user requests.

When fine tuning a model, during the epochs, training loss occurs (Figure 4.4). Training loss measures how well a machine learning model performs during the training phase by quantifying the difference between the predicted and actual values in the training dataset. The goal is to reduce training loss as training progresses, indicating that the model is becoming more accurate in its predictions on the training data.

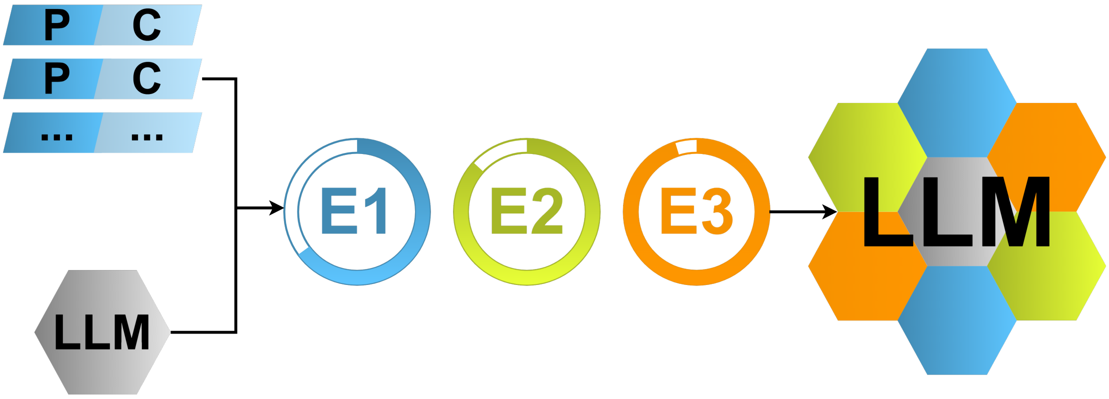
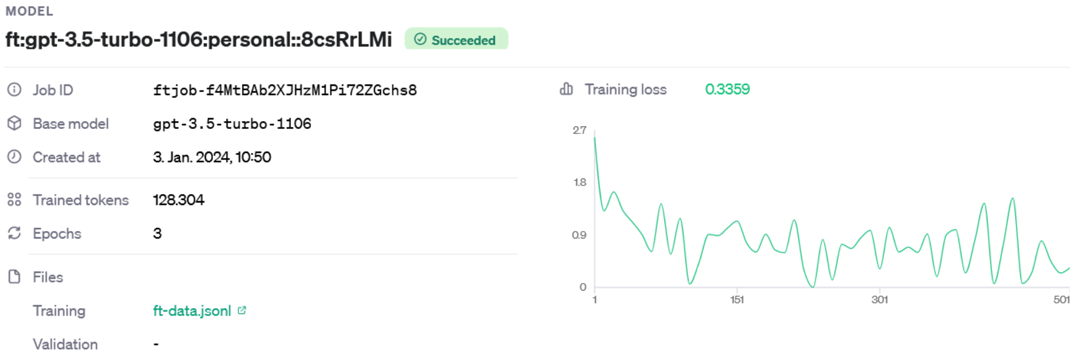


Figure 4.3: The Fine Tuning process simplified. Training loss in epochs is represented as partially filled circles. The LLM acquires the domain-specific knowledge in varying proportions corresponding to the training loss of each epoch.

In simple terms, Fine Tuning can be thought of as providing your domain-specific 'vocabulary' to your language model, teaching it how to speak in the style of your domain.



MODEL

ft:gpt-3.5-turbo-1106:personal::8csRrLMi Succeeded

Job ID	`ftjob-f4MtBAb2XJHzM1Pi72ZGchs8`
Base model	`gpt-3.5-turbo-1106`
Created at	3. Jan. 2024, 10:50
Trained tokens	128.304
Epochs	3
Files	
Training	ft-data.jsonl
Validation	-

Training loss 0.3359

Figure 4.4: Our actual Fine Tuning job output containing important information such as the base model, number of trained tokens, number of epochs, the training file and training loss.

4.5 Prompt Engineering

"Prompt Engineering is an increasingly important skill set needed to converse effectively with large language models (LLMs), such as ChatGPT. Prompts are instructions given to an LLM to enforce rules, automate processes, and ensure specific qualities (and quantities) of generated output. Prompts are also a form of programming that can customise the outputs and interactions with an LLM." (White et al. 2023).

Similar to Fine Tuning, for Prompt Engineering as well rewriting the text to JSONL format was necessary (Table 3.1, Table 4.5). In summary, the ‘role’ attribute contains tokens provided by the system versus those provided by the user, while the ‘content’ attribute contains the actual text of those tokens. This approach enables more precise control over the prompt and context provided to the LLM during Prompt Engineering. In practice, the system prompt sets the stage and provides guidance for the model, while the user input directs the model’s response towards the desired outcome. By separating them, we have more control over the context and can influence the generated output accordingly.

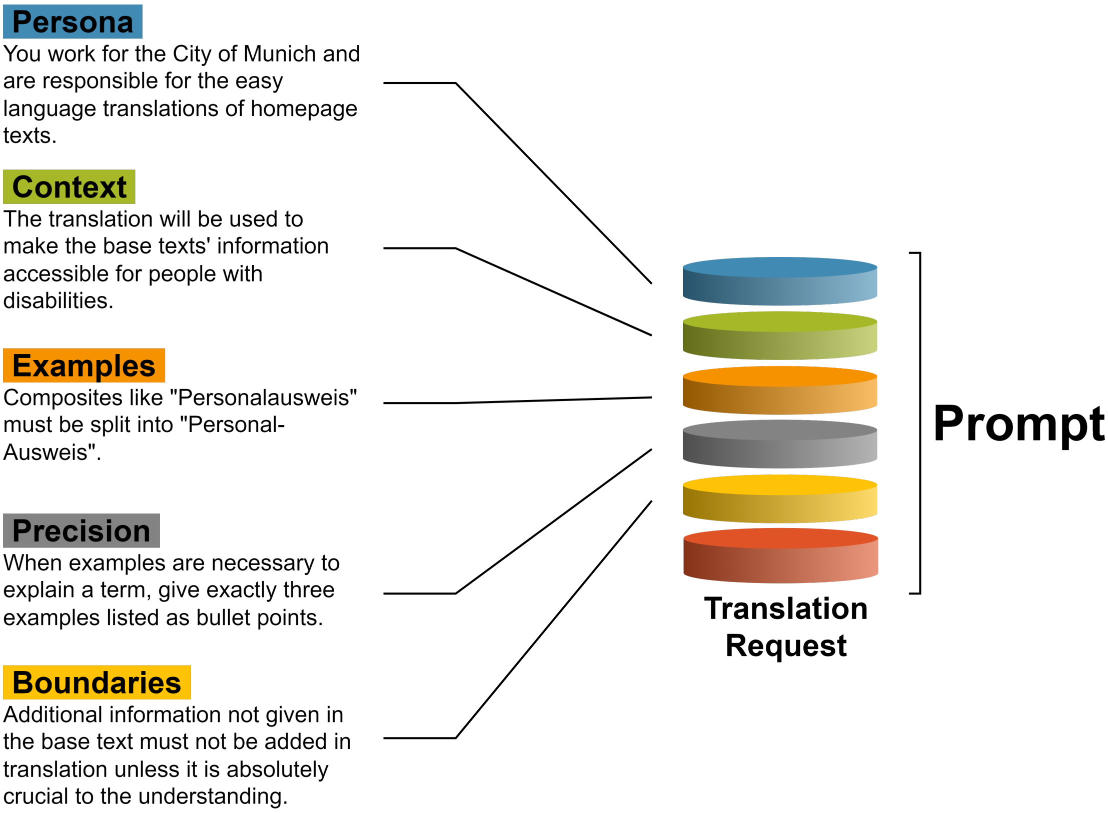
Rule Benutzen Sie einfache Wörter. Beispiel Schlecht: genehmigen Gut: erlauben
Prompt Format `{"role": "system", "content": "Benutzen Sie einfache Wörter.\nBeispiel\nSchlecht: genehmigen\nGut: erlauben"}`

Table 4.5: NLS rules used for Prompt Engineering transformed into JSONL format.

To enhance the clarity and comprehensibility of the prompt, we approached it as a composition of distinct parts rather than a continuous prose (Figure 4.5). We defined a persona for the LLM and provided contextual information to render the results more domain-specific. Examples were used to fine tune the prompt and ensure precision while defining boundaries to restrict the LLM from generating irrelevant content. The following graphic illustrates fictional examples for each category.

In our case, NLS (2022) rules, that were added to the prompt, serve as examples, precision and boundaries. Currently, OpenAI’s ChatGPT API does not include a built-in prompt ID feature. Each API call is considered independent and does not have access to the context or prompts of previous calls. As a result, we had to send the complete prompt before each translation call. The translation prompt was as shown in Table 4.6. During runtime, the variables *‘formatted_rules’* and *‘message’* were replaced accordingly.

In simple terms, Prompt Engineering can be thought of as having a supervisor standing right next to you, providing clear and structured instructions on what to do and how to do it in real-time.



Persona
You work for the City of Munich and are responsible for the easy language translations of homepage texts.

Context
The translation will be used to make the base texts' information accessible for people with disabilities.

Examples
Composites like "Personalausweis" must be split into "Personal-Ausweis".

Precision
When examples are necessary to explain a term, give exactly three examples listed as bullet points.

Boundaries
Additional information not given in the base text must not be added in translation unless it is absolutely crucial to the understanding.

Figure 4.5: A schematic representation of a structured prompt that follows the guidelines of Prompt Engineering.

Initial Prompt

Sie sind ein Übersetzer, welcher Texte, im Kontext der Landeshauptstadt München, in Leichte Sprache übersetzt.

Die Übersetzungen werden genutzt um den Basistext Menschen mit Behinderung zugänglich zu machen.

Komposita wie "Personalausweis" müssen in z.B. "Personal-Ausweis" aufgeteilt werden.

Falls Beispiele nötig sind um einen Sachverhalt zu erklären, geben Sie maximal drei Beispiele an.

Bei der Übersetzung bleiben Sie so nahe wie möglich am Inhalt der ursprünglichen Nachricht.

Sie erfinden nichts dazu was keine Relation mehr zur ursprünglichen Nachricht hat!

Ich nenne Ihnen zunächst alle Regeln die für die Übersetzung in Leichte Sprache eingehalten werden müssen.

Danach werde ich Ihnen einen Text geben. Dieser muss in Leichte Sprache übersetzt werden.

Nun die Regeln:

{formatted_rules}

Folgendes bitte in Leichte Sprache übersetzen und dabei die genannten Regeln einhalten:

{message}

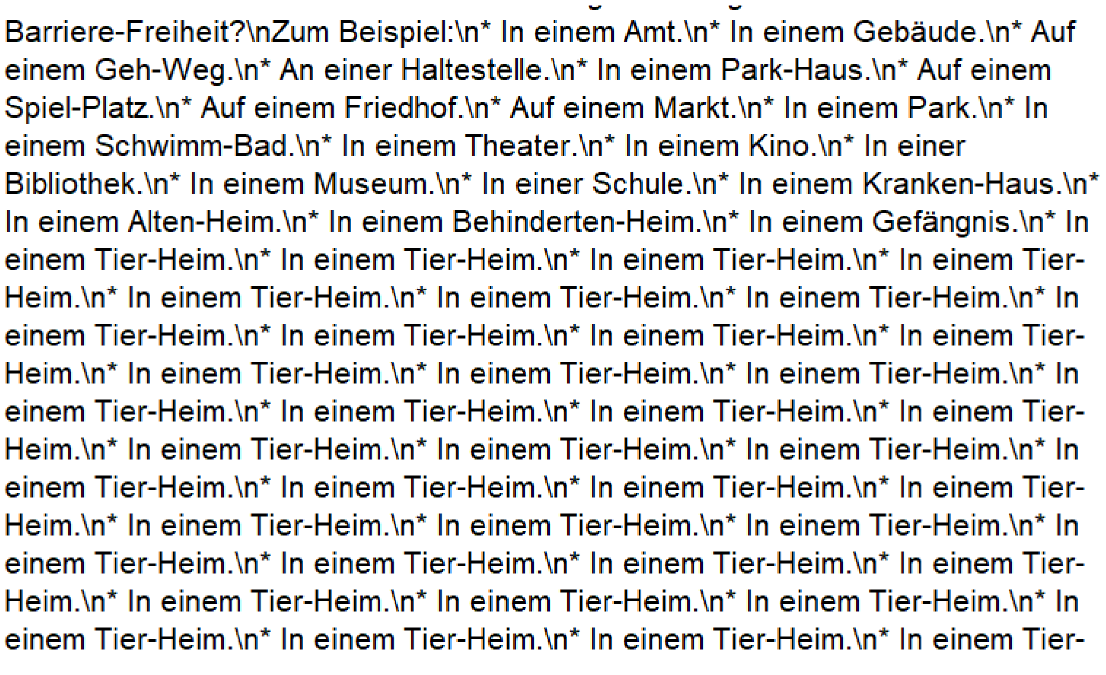
Table 4.6: The prompt we used in our Prompt Engineering experiments. The two variables within the prompt are replaced at runtime.

4.6 Combinations

One advantage of the three selected enhancement methods is their compatibility with each other. After programming and running the three basic versions, we also combined **Em**bedding, **F**ine **T**uning and **P**rompt **E**ngineering. Our aim is to gain a comprehensive understanding of which enhancements are most suitable for our specific text simplification and how they interact with each other by exploring all possible combinations. The combinations are listed alphabetically and in ascending order of the number of methods combined.

- EM-FT
- EM-PE
- FT-PE
- EM-FT-PE

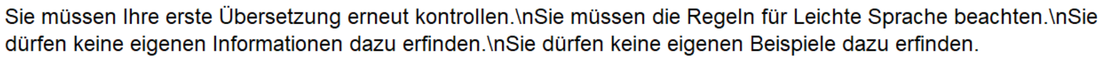
Using EM-FT-PE, we first fine tuned a model with parallel text pairs from our corpora. We then embedded the parallel corpora pages into the fine tuned model. This allowed us to trigger a similarity search in those embeddings using an engineered prompt.



```
Barriere-Freiheit?\nZum Beispiel:\n* In einem Amt.\n* In einem Gebäude.\n* Auf
einem Geh-Weg.\n* An einer Haltestelle.\n* In einem Park-Haus.\n* Auf einem
Spiel-Platz.\n* Auf einem Friedhof.\n* Auf einem Markt.\n* In einem Park.\n* In
einem Schwimm-Bad.\n* In einem Theater.\n* In einem Kino.\n* In einer
Bibliothek.\n* In einem Museum.\n* In einer Schule.\n* In einem Kranken-Haus.\n*
In einem Alten-Heim.\n* In einem Behinderten-Heim.\n* In einem Gefängnis.\n* In
einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-
Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In
einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-
Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In
einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-
Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In
einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-
Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In
einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-
Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In
einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-Heim.\n* In einem Tier-
```

Figure 4.6: A shortened example of a failed translation containing 'endless' repetitions of the same text.

The combination of enhancements resulted in hallucinations, sentence repetition (Figure 4.6), or direct exposure of the given prompt (Figure 4.7), despite explicit instructions to the contrary. To address this issue, we reiterated over the initial results a second time. We used a specifically crafted prompt (Table 4.7) to encourage the LLM to correct its errors during the second processing. However, there are instances where refinement is ineffective and human intervention is required to address the issues. This is likely to occur when the refinement prompt becomes excessively long due to the length of the initial translation.



Sie müssen Ihre erste Übersetzung erneut kontrollen.\nSie müssen die Regeln für Leichte Sprache beachten.\nSie dürfen keine eigenen Informationen dazu erfinden.\nSie dürfen keine eigenen Beispiele dazu erfinden.

Figure 4.7: An example of a failed translation that contains fragments of the actual prompt (Table 4.7), even though the prompt explicitly advises the LLM not to expose it.

Second Prompt
Sie sind ein Übersetzer, welcher Texte, im Kontext der Landeshauptstadt München, in Leichte Sprache übersetzt. Die Übersetzungen werden genutzt, um den Basistext Menschen mit Behinderung zugänglich zu machen. Der folgende Text wurde von Ihnen bereits in Leichte Sprache übersetzt: {translation_rough} Der ursprüngliche Text vor der Übersetzung war: {message} Sie müssen Ihre erste Übersetzung erneut kontrollieren und dabei auf folgende Punkte speziell achten: * Sie entfernen alle mehrfachen identischen Wiederholungen eines Satzes! * Sie beschränken die Übersetzung auf das Wesentliche und kürzen diese, falls sinnlose Text-Patterns oder Beispiele wiederholt werden, die keine wesentlich neuen Informationen beitragen! * Sie stellen sicher, dass die Übersetzung nicht mehr als drei Beispiele bzw. Bullet Points je zu erklärendem Begriff enthält! * Sie geben auf keinen Fall die Regeln zur Übersetzung direkt aus! * Sie übersetzen erneut, falls der ursprüngliche Text mit der Übersetzung identisch ist! * Sie erfinden keine Informationen dazu, die nicht im Ausgangstext stehen!

Table 4.7: The prompt used during the second iteration of the initial translation for experiments that involved combined enhancement methods. The variables *'translation_rough'* and *'message'* are replaced during runtime.

Chapter 5

Results

This chapter presents a detailed analysis of the outcomes achieved by implementing Embedding techniques, Fine Tuning methods and innovative Prompt Engineering strategies to enhance an LLM for translating complex German text into Easy Language (Section 2.2). It aims to demonstrate the effectiveness of state-of-the-art NLP methods in addressing the challenges of Easy Language translation (Chapter 3). The empirical results presented are based on a rigorous experimental design, using both qualitative and quantitative analyses to provide a comprehensive view of the models' performance (Chapter 4). This chapter is a crucial part of the thesis as it provides insight into the feasibility of using LLMs to make complex information more accessible through Easy Language. This helps to bridge the gap identified in existing research and contributes to the ongoing discourse in the field of NLP (Section 2.6). Our results represent a step forward in the pursuit of making information universally understandable.

We did not find any particular demand or special need in data preparation. For all methods standards such as using parallel corpora (Section 3.2) and transforming it into JSON(L) format for processing, were sufficient. Page-based parallel corpora (Section 3.2.1) for Embedding and pair-based parallel corpora (Section 3.2.2) for Fine Tuning were sufficient. The Easy Language rules transformed to JSONL format (Section 4.2, Appendix C) were effectively integrated into the engineered prompt.

It was found that no enhancement method achieves first place in all metric evaluations. Therefore, it is not possible to identify a single perfect method to address our research question. In fact, all methods produce valuable results, but there is a clear distinction between which methods outperform the others in terms of the amount of valuable translations and the amount of Easy Language rules they adhere to. Fine Tuning stands out as the best standalone enhancement method, providing the best results across all metrics (Table 5.1). As for combined methods, Embedding combined with Fine Tuning proved to be the best approach (Table 5.2). In the category of Structural Simplicity, both FT and EM-FT outperformed the human translations (Section 5.2).

As combined methods require a second iteration of translation refinement and additional human control for a certain number of results (Section 4.6), standalone and combined enhancements are ranked separately (Table 5.1, Table 5.2 and Section 5.1).

	GPT	EM	FT	PE	GPT	EM	FT	PE
BERTScore	0.67	0.68	0.74	0.68	3rd	2nd	1st	2nd
FRE	56.69	48.85	50.89	46.16	1st	3rd	2nd	4th
LIX	49.83	48.74	34.03	43.76	4th	3rd	1st	2nd
METEOR	0.14	0.17	0.21	0.15	4th	2nd	1st	3rd
SARI	0.44	0.45	0.59	0.46	4th	3rd	1st	2nd
SNLR	0.66	0.61	0.91	0.63	2nd	4th	1st	3rd
TER	224	1323	140	516	2nd	4th	1st	3rd
TTR	0.87	0.77	0.82	0.78	4th	1st	3rd	2nd
AVG RANKING					3.00	2.75	1.38	2.63

Table 5.1: The standalone enhancement methods are presented with their respective scores in the corresponding metric. They are ranked from first to fourth place. For each method the average rank was calculated by the sum of all ranks divided by the number of metrics.

	EM-FT	EM-PE	FT-PE	EM-FT-PE	EM-FT	EM-PE	FT-PE	EM-FT-PE
BERTScore	0.75	0.71	0.72	0.73	1st	4th	3rd	2nd
FRE	34.09	39.38	34.96	34.24	4th	1st	2nd	3rd
LIX	33.83	40.08	39.23	35.05	1st	4th	3rd	2nd
METEOR	0.34	0.20	0.18	0.32	1st	3rd	4th	2nd
SARI	0.64	0.52	0.55	0.62	1st	4th	3rd	2nd
SNLR	0.94	0.95	0.99	0.93	3rd	2nd	1st	4th
TER	906	935	537	1003	2nd	3rd	1st	4th
TTR	0.64	0.68	0.63	0.61	3rd	4th	2nd	1st
AVG RANKING					2.00	3.13	2.38	2.50

Table 5.2: The combined enhancement methods are presented with their respective scores in the corresponding metric. They are ranked from first to fourth place. For each method the average rank was calculated by the sum of all ranks divided by the number of metrics.

5.1 Ranking

This chapter uses the base text presented in Table 5.3 and its translations to illustrate the differences in results between the methods. The purpose of this comparison is to highlight the strengths and weaknesses of each method. Additionally, the reader can compare the sample results of the methods with the corresponding human translation. Section 5.2 offers a more detailed analysis of the ranking of the human translations compared to the enhancement methods across applicable metrics.

The enhancement methods were ranked according to a simple grading system, from first to fourth place, based on each metric. An overall average ranking was then calculated for both standalone and combined enhancement methods (Table 5.4). All metrics were given equal weight in this calculation.

Base Text Example
Öffentliche Stellen sind gemäß der Richtlinie (EU) 2016/2102 des Europäischen Parlaments dazu verpflichtet, Webauftritte barrierefrei zugänglich zu machen.
Human Translation
Diese Internet-Seite soll barriere-frei werden. Das heißt: Alle Menschen sollen die Internet-Seite ohne Hilfe nutzen können. Also zum Beispiel auch: * blinde Menschen * gehörlose Menschen * Menschen, die nicht alle Finger bewegen können.

Table 5.3: An example of a base text and its corresponding human Easy Language translation from our parallel corpora.

Although we derived clear rankings of methods, with a distinct order of ratings from first to fourth place within each metric (Table 5.1, Table 5.2), even the lowest ranked methods produced results worthy of further improvement. This can be seen through the fact that some methods are closer than their overall rankings suggest (Table 5.4). For example, the average rankings of PE (2.63) compared to EM (2.75), or FT-PE (2.38) compared to EM-FT-PE (2.50). Additionally the metrics' rankings can sometimes be very narrow, such as the SARI scores for standalone methods (FT 0.59 / PE 0.46 / EM 0.45 / GPT 0.44) or the BERTScores for combined methods (EM-FT 0.75 / EM-FT-PE 0.73 / FT-PE 0.72 / EM-PE 0.71). This illustrates the similarity of the methods' results and indicates that there is no clear winner or loser in this research.

Therefore the overall rankings of the methods for enhancing an LLM's ability to translate text into Easy Language are listed in order of suitability, from most to least (Table 5.4), rather than a classical ranking from first to fourth place as used for metric rankings of methods (Table 5.1, Table 5.2). As Section 5.1.3 shows, the second and third rankings are very similar, with only slight differences in their results. Therefore, we name these levels equally, only differing by a plus and minus sign to indicate their quality tendencies.

SUITABILITY	**STANDALONE**	**COMBINATION**
most	FT	EM-FT
fairly$^{(+)}$	PE	FT-PE
fairly$^{(-)}$	EM	EM-FT-PE
least	GPT	EM-PE

Table 5.4: The ranking shows the suitability of standalone and combined enhancement methods, listed from the most to the least suitable approaches.

The radar graphs below show the rankings of enhancement methods for each metric. The best ranking (1st place) is represented on the outermost line, with metrics arranged on the corners. Therefore, the larger the highlighted area, the more suitable and versatile the enhancement method.

Unlike the overall rating (Table 5.4), the rankings for each level of suitability (Sections 5.1.1 - 5.1.4) are averaged per type of simplicity (Section 2.3) in order to identify their strengths and weaknesses for automatic text simplification (Section 2.2). Various characteristics are analysed at different levels of suitability to gain a comprehensive understanding of the results and their impact on method suitability.

5.1.1 Most Suitable

FT

*"Öffentliche Stellen müssen ihre Internet-Seiten barriere-frei machen.
Was bedeutet barriere-frei?
Alle Menschen sollen die Internet-Seiten gut nutzen können.
Auch Menschen mit Behinderung."*

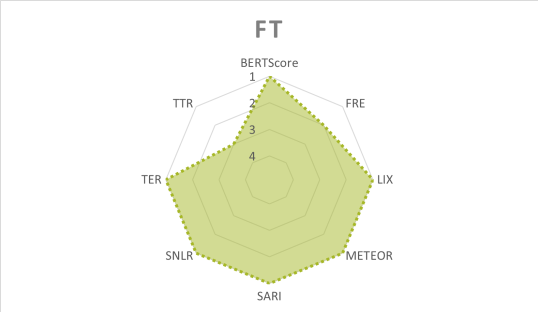


Figure 5.1: A radar chart displaying strengths and weaknesses of the FT method.

EM-FT

*"Öffentliche Stellen müssen ihre Internet-Seiten barriere-frei machen.
Das steht in der Barrierefreiheits-Richtlinie."*

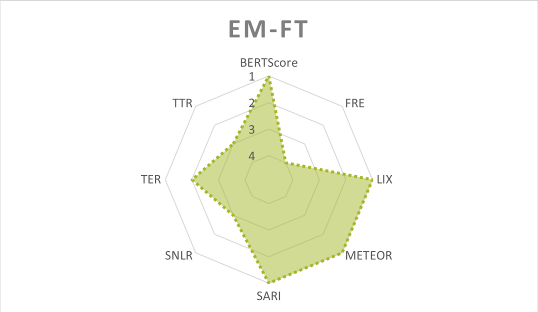


Figure 5.2: A radar chart displaying strengths and weaknesses of the EM-FT method.

Figure 5.1 and Figure 5.2 demonstrate that these two methods are the most versatile and achieve the best rankings across all metrics.

	Direct Assessment	Simplicity Gain	Structural Simplicity
FT	1.00	1.00	1.75
EM-FT	1.00	1.25	2.75

Table 5.5: A dedicated analysis of the rankings of the most suitable methods, averaged by categories of text simplicity, as presented in Section 2.3.

Table 5.5 demonstrates that both FT and EM-FT achieve outstanding average ranks of 1.25 or better for Direct Assessment and Simplicity Gain. These categories are clearly their **strengths**. In the category of Structural Simplicity, the average rank is 1.75 for FT and 2.75 for EM-FT, indicating a **weakness** for both metrics. It is important to notice, that 1.75 for FT is still the best result of all methods for SS. To improve the methods to production-ready quality, it is crucial to address this weakness.

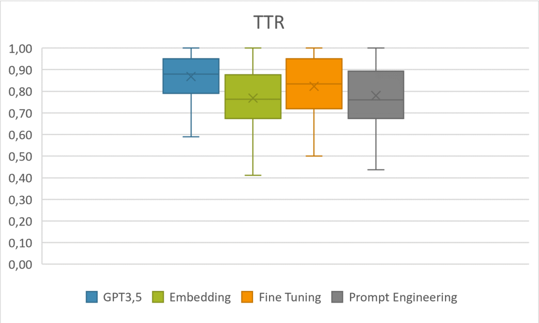


Figure 5.3: The TTR box plot for all standalone enhancement methods, allowing for analysis and comparison of median, minimum, and maximum scores, as well as outliers.

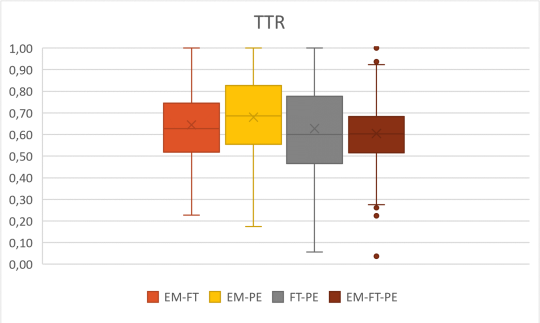


Figure 5.4: The TTR box plot for all combined enhancement methods, allowing for analysis and comparison of median, minimum, and maximum scores, as well as outliers.

	Lowest	**Median**
EM	0.41	0.76
PE	0.44	0.76
FT	**0.50**	**0.84**
GPT	0.59	0.88

Table 5.6: The standalone methods' lowest consistent score and the median scores sorted in ascending order.

	Lowest	**Median**
EM-FT-PE	0.28	0.60
FT-PE	0.06	0.60
EM-FT	**0.23**	**0.63**
EM-PE	0.17	0.69

Table 5.7: The combined methods' lowest consistent score and the median scores sorted in ascending order.

Although both of the most suitable enhancement methods share this weakness, they are not as far apart from their competitors as it may seem. As both methods only placed 3rd in the TTR metric, we use this metric to further analyse their weak point in Structural Simplicity. Figure 5.3 and Figure 5.4 show that all methods are quite close in this metric.

Across all rankings, all methods produced translations with a TTR score of 1.0, which is the worst possible score. FT is outperformed by its respective leader EM only by 0.09 regarding the lowest (best) score and only by 0.08 regarding the median score (Table 5.6). EM-FT outperforms its respective leader EM-FT-PE by 0.05 regarding the lowest (best) score and is outperformed itself only by 0.03 regarding the median score (Table 5.7).

This reduces the identified weakness, as both methods only need to improve by less than 0.10 points in terms of the TTR score or already outperform the leading method. Therefore, we consider Structural Simplicity to be a weakness only within the metrics themselves, but not in comparison with the other metrics.

5.1.2 Fairly$^{(+)}$ Suitable

PE

"Alle öffentlichen Stellen müssen ihre Internet-Seiten barrierefrei machen.
Das steht in der Richtlinie (EU) 2016/2102 des Europäischen Parlaments."

FT-PE

"Die Landes-Haupt-Stadt München arbeitet daran,
dass alle Menschen die Internet-Seiten gut benutzen können.
Das nennt man:
barriere-freie Internet-Seiten.
Barriere-frei bedeutet:
Alle Menschen können alles gut benutzen.
Auch Menschen mit Behinderung.
Zum Beispiel:
** Menschen, die nicht gut sehen können.*
** Menschen, die nicht gut hören können.*
** Menschen, die nicht gut sprechen können.*
** Menschen, die nicht gut laufen können."*

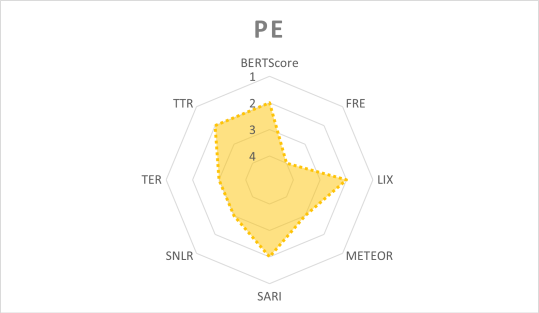


Figure 5.5: A radar chart displaying strengths and weaknesses of the PE method.

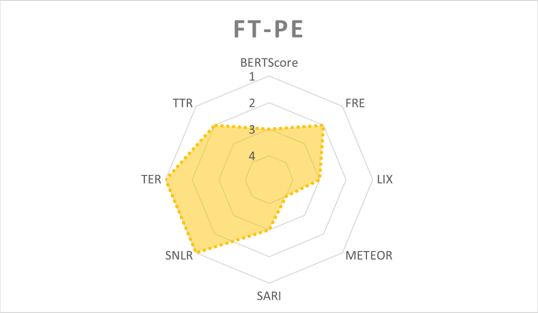


Figure 5.6: A radar chart displaying strengths and weaknesses of the FT-PE method.

Figure 5.5 and Figure 5.6 demonstrate the versatility of these methods. However, the radar method has a smaller expanse compared to the most suitable methods.

Table 5.8 demonstrates that PE achieves an average ranking of 2.xx across all categories. It excels in Direct Assessment but falls short in Structural Simplicity. The strengths and weaknesses of FT-PE are reversed compared to PE. While it achieves

	Direct Assessment	Simplicity Gain	Structural Simplicity
PE	2.33	2.50	2.75
FT-PE	3.33	2.75	2.00

Table 5.8: A dedicated analysis of the rankings of the fairly$^{(+)}$ suitable methods, averaged by categories of text simplicity, as presented in Section 2.3.

2.xx rankings for Simplicity Gain and Structural Simplicity, it only scores 3.33 for Direct Assessment.

For PE, the categories show a minimal difference of only 0.21 points on average. Therefore, we consider PE to have balanced results without any significant strengths or weaknesses. Although FT-PE performs slightly worse on average across the three categories, with a categorical difference spread of about 0.67 points, we also consider it a balanced method without significant peaks or lows.

5.1.3 Fairly$^{(-)}$ Suitable

EM

"Öffentliche Stellen müssen ihre Websites so gestalten, dass sie für alle Menschen leicht zugänglich sind. Das steht in einer Regel, die das Europäische Parlament gemacht hat."

EM-FT-PE

"Die Stadt München muss ihre Internet-Seiten barriere-frei machen.
Das heißt:
Alle Menschen müssen die Internet-Seiten gut benutzen können.
Auch Menschen mit Behinderung.
Zum Beispiel:
** Menschen, die nicht gut sehen.*
** Menschen, die nicht gut hören.*
** Menschen, die nicht gut sprechen.*
** Menschen, die nicht gut lesen.*
** Menschen, die nicht gut schreiben.*
** Menschen, die nicht gut mit der Hand arbeiten können."*

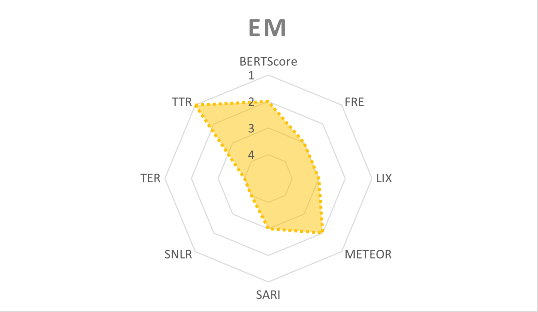


Figure 5.7: A radar chart displaying strengths and weaknesses of the EM method.

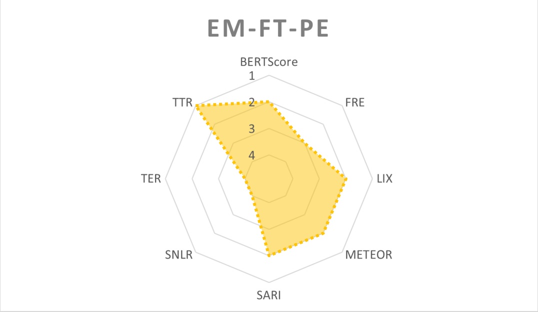


Figure 5.8: A radar chart displaying strengths and weaknesses of the EM-FT-PE method.

Figure 5.7 and Figure 5.8 are quite similar to those of the slightly superior methods from Section 5.1.2 regarding the highlighted area. Both EM and EM-FT-PE are of lower quality compared to the most suitable methods, similar to the fairly$^{(+)}$ suitable methods.

	Direct Assessment	Simplicity Gain	Structural Simplicity
EM	2.33	2.75	2.75
EM-FT-PE	2.00	2.50	2.50

Table 5.9: A dedicated analysis of the rankings of the fairly$^{(-)}$ suitable methods, averaged by categories of text simplicity, as presented in Section 2.3.

Similarly to the small difference in the radar graphs, there are only minor differences in the quality of the control sample translations between the fairly$^{(+)}$ and fairly$^{(-)}$ suitable methods. The average scores achieved by the two levels' methods support this claim. The superior PE method has an average score of 2.63, while the EM method has a score of 2.75 (Table 5.1). In Table 5.2, the FT-PE achieves a score of 2.38 and the EM-FT-PE achieves a score of 2.50. Notably, the standalone and combined methods both have an average difference of only 0.12 points to their superior counterparts.

5.1.4 Least Suitable

GPT

"Öffentliche Stellen müssen gemäß der Richtlinie (EU) 2016/2102 des Europäischen Parlaments ihre Websites barrierefrei gestalten."

EM-PE

"Öffentliche Stellen müssen ihre Internet-Seiten barrierefrei machen.
Das steht in der Richtlinie (EU) 2016/2102 des Europäischen Parlaments."

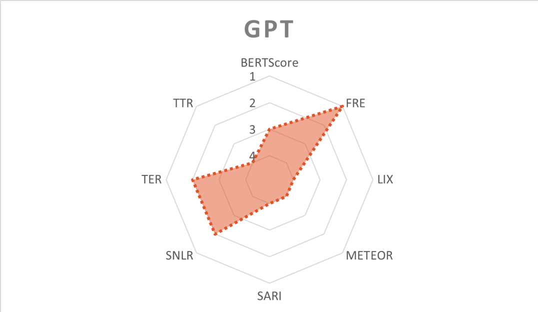


Figure 5.9: A radar chart displaying strengths and weaknesses of the GPT baseline.

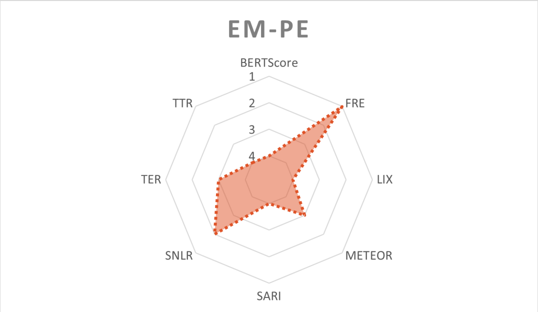


Figure 5.10: A radar chart displaying strengths and weaknesses of the EM-PE method.

Figure 5.9 and Figure 5.10 clearly visualise that GPT and EM-PE are the least suitable methods, with the lowest rankings across all metrics. They show the smallest extent of the highlighted area.

	Direct Assessment	Simplicity Gain	Structural Simplicity
GPT	3.67	3.25	2.75
EM-PE	3.67	3.50	2.75

Table 5.10: A dedicated analysis of the rankings of the least suitable methods, averaged by categories of text simplicity, as presented in Section 2.3.

The two methods with the lowest suitability are overwhelmingly ranked by 3.xx scores, indicating their inferiority to the other metrics (Table 5.10). The control sample translations also demonstrate their lower quality compared to the samples from the other methods.

We consider both GPT and EM-PE to have no strengths compared to the other metrics due to their best score of 2.75 in the Structural Simplicity category being regarded as a weakness in the better suited methods. The results of these methods indicate the most significant deficit of all methods in terms of being production-ready.

5.2 Human Translation Ranking

Human translations (HU) cannot be compared to the machine-generated translations across all metrics, as they serve as the reference translation themselves. However, we can compare all translations equally when evaluating Structural Simplicity (Section 2.3), as these metrics do not require a reference translation.

	GPT	EM	FT	PE	HU	GPT	EM	FT	PE	HU
FRE	56.69	48.85	50.89	46.16	37.56	1st	3rd	2nd	4th	5th
LIX	49.83	48.74	34.03	43.76	33.50	5th	4th	2nd	3rd	1st
SNLR	0.66	0.61	0.91	0.63	0.96	3rd	5th	2nd	4th	1st
TTR	0.87	0.77	0.82	0.78	0.78	4th	1st	3rd	2nd	2nd
AVG RANKING						3.00	2.75	1.38	2.63	2.25

Table 5.11: The standalone enhancement methods and the human translations are presented with their respective scores in the corresponding metric. They are ranked from first to fourth place. For each method the average rank was calculated by the sum of all ranks divided by the number of metrics.

	EM-FT	EM-PE	FT-PE	EM-FT-PE	HU	EM-FT	EM-PE	FT-PE	EM-FT-PE	HU
FRE	34.09	39.38	34.96	34.24	37.56	5th	1st	3rd	4th	2nd
LIX	33.83	40.08	39.23	35.05	33.50	2nd	5th	4th	3rd	1st
SNLR	0.94	0.95	0.99	0.93	0.96	4th	3rd	1st	5th	2nd
TTR	0.64	0.68	0.63	0.61	0.78	3rd	4th	2nd	1st	5th
AVG RANKING						2.00	3.13	2.38	2.50	2.50

Table 5.12: The combined enhancement methods and the human translations are presented with their respective scores in the corresponding metric. They are ranked from first to fourth place. For each method the average rank was calculated by the sum of all ranks divided by the number of metrics.

Table 5.11 and Table 5.12 demonstrate that on average, human translations achieve an **FRE** score of 37.56. Notably, this places them last when compared to the standalone methods, behind PE (46.16). However, when compared to the combined methods, human translations would rank on the second place, behind EM-PE (39.38) and ahead of FT-PE (34.96).

Human translations achieve an average **LIX** score of 33.50, placing them in first position among standalone methods, ahead of FT (34.03) and also in first position among combined methods, ahead of EM-FT (33.83).

On average, human translations achieve an **SNLR** score of 0.96, which would place them first in standalone ranking before FT (0.91). For combined ranking, HU would be ranked second, after FT-PE (0.99) but before EM-PE (0.95).

Human translations achieve an average **TTR** score of 0.78, placing them in second position among standalone methods alongside PE and narrowly behind the first placed EM (0.77). However, in contrast, for combined methods, HU ranks last behind EM-PE (0.68).

SUITABILITY	STANDALONE	COMBINATION
most	FT	EM-FT
fairly$^{(+)}$	HU	FT-PE
fairly	PE	EM-FT-PE / HU
fairly$^{(-)}$	EM	
least	GPT	EM-PE

Table 5.13: The ranking shows the suitability of standalone and combined enhancement methods compared to human translations in terms of Structural Simplicity, listed from the most to the least suitable approaches. A fifth level 'fairly' was added for this ranking.

When considering **Structural Simplicity rankings** (Table 5.13), the human translations achieve **fairly$^{(+)}$ suitable** for standalone with an average ranking of 2.25. For the combination methods they achieve **fairly suitable** with an average ranking of 2.50. They lose the top spots to FT (1.38) and EM-FT (2.00). Given that EM-PE achieved a 3.xx rating, which is comparable to GPT, we elected to rank it in the least suitable level, rather than the now skipped fairly$^{(-)}$ suitable level.

It is worth noting that while EM-FT-PE only achieved fairly$^{(-)}$ suitability in the overall ranking (Section 5.1), it is on par with the human reference translations in terms of Structural Simplicity. This statement supports the proposition that all enhancement methods are worthy of further improvement. This will be discussed further in Section 6.4. Moreover, this comparison does not provide a clear position for human translations nor a clear superior or inferior method for each metric. These results provide evidence for the assertion that various enhancement techniques possess distinct advantages and disadvantages (Section 5.1). Further research in this area could yield valuable insights. Section 6.4.4 presents our conclusion on this topic.

Chapter 6

Conclusion & Outlook

This chapter summarises the key research findings in relation to the research aims, reviews the limitations of the study and proposes opportunities for future research. Before delving to the conclusions, it is useful to review the original research topic and its associated questions.

Which technique is most suitable for improving an LLM's ability to translate complex text into Easy Language?

- *How does the existing public data need to be prepared for this purpose?*
- *How does the existing language model need to be enhanced with the prepared data?*

6.1 Findings

The objective of this study is to explore effective techniques for enhancing an LLM's capacity to translate intricate text into Easy Language. This is a prerequisite for a future extension of the City of Munich's internal LLM service "MUCGPT".

Therefore, we investigated how to prepare the existing public data for this purpose. We showed that the public websites of the city of Munich and their Easy Language counterparts are sufficient as base data for our parallel corpora. They can be used as documents for Embedding or prompt/completion-pairs can be derived for Fine Tuning. Furthermore the Easy Language rules NLS (2022) positively affect the translation quality when being used in an engineered prompt. For these tasks, standard file formats such as JSON(L) are adequate.

We also looked at how an LLM needs to be enhanced with the prepared data. We examined the three enhancement methods: Embedding, Fine Tuning and Prompt Engineering. We evaluated both their standalone and their combined setups and their translations of 170 examples of normal language text. The methods were ranked from best to least suitable, as all of them produced valuable results to a certain extent. The three types of Text Simplification: Direct Assessment, Simplicity Gain and Structural Simplicity, were used to locate these methods' strengths and weaknesses. We could provide data to lessen the weak point of the overall best ranked methods Fine Tuning and EM-FT concerning Structural Simplicity. We declared these methods as the most suitable for enhancing an existing LLM for the purpose of Easy Language translation.

However, there is still room for improvement in all of the methods. Vicariously for all translations, we analysed various control samples. We highlighted the existing problems and the necessary improvements to enhance the methods' quality on their way to becoming production ready.

6.1.1 Method comparison

Pre-trained word **Embeddings**, such as Word2Vec or GloVe, can enhance LLMs by providing a deeper understanding of language semantics. This can help distinguish between complex and simplified language constructs (Mikolov et al. 2013a; Pennington et al. 2014). This method utilises contextual knowledge, which may improve the model's ability to understand the nuances of language simplification.

Fine Tuning involves training pre-existing LLMs on a tailored dataset comprising pairs of standard German and its Easy Language equivalents. This method aligns the LLM's capabilities with the task at hand. Howard and Ruder (2018) demonstrated the efficacy of Fine Tuning in adapting models to specific textual styles or complexities. Fine Tuning enables the model to adjust its internal parameters to more accurately capture the subtleties of language simplification, which may result in more precise and contextually relevant translations.

Prompt Engineering presents a distinct approach by creating precise prompts that direct the LLM to produce the desired output without requiring extensive retraining. This technique utilises the inherent abilities of LLMs to comprehend and generate text based on the context provided in the prompt (Brown et al. 2020). Although it requires fewer computational resources than Fine Tuning, the effectiveness of prompt-based methods heavily depends on the quality of the prompts. This may limit the model's ability to adapt to the wide range of linguistic structures encountered in the translation task.

In comparison, Embeddings offer a fundamental semantic understanding that is beneficial for distinguishing between complex and simplified language. Fine Tuning tailors the model more precisely to the task, potentially resulting in the most accurate translations. Prompt Engineering offers a cost-effective and resource-light method, but its effectiveness heavily relies on the craft of prompt design and may vary. When selecting a method, it is important to take into account the particular needs of the translation task, the availability of data, computational resources, and the desired balance between accuracy and efficiency.

6.2 Contributions to the Field

As previously stated, the findings of this thesis will be integrated into the City of Munich's internal chatbot, MUCGPT, as a new function. This function will assist the city's employees in creating texts in Easy Language. Additionally, this thesis provides a comprehensive comparison of the mentioned enhancement methods, which we did not encounter during our own research.

To a certain extent, this thesis feels like a potential implementation of the (Djeffal and Horst 2021, pp. 24 - 30) workshop and aims to address unresolved issues identified by Deilen et al. (2023).

We hope this thesis can serve as a blueprint for others to compare enhancement methods for different goals or to specialise in one of the discussed methods.

6.3 Limitations of the Study

The research encountered several limitations. **No standardised evaluation procedure** for Easy Language was found. Furthermore, the **rules contradict themselves** to a certain degree. For instance, rule T2 NLS (2022) suggests avoiding questions, but also acknowledges that questions can be useful in certain cases, such as headings. As a result, establishing appropriate metrics for meaningful evaluation that can be compared to other papers was challenging.

Furthermore, due to time and workforce constraints, we were **only** able to improve and evaluate **one Large Language Model**. We determined that adding another LLM would compromise the overall quality of the thesis to a greater extent than the additional findings would benefit the ranking of the enhancement methods.

This study offers a comprehensive understanding of how different enhancement methods can improve LLM capabilities for text simplification and how they can impact each other when combined. However, due to time constraints, the **metric ranking may be expandable**. To ensure comparability, all methods were trained with equal quality, using generally accepted default settings, such as three epochs for Fine Tuning. However, this leaves open the possibility that metric rankings could change if all metrics were built to perfection.

Moreover, as previously mentioned in Section 3.2, our **parallel corpora are limited**. Firstly, the data was not generated by a certified Easy Language translator. Secondly, our research focuses solely on the domain of Easy Language in the public administration of the City of Munich. This limits our training data to approximately 40 websites and 170 text pairs from only one specific domain.

In addition, the **fast-paced advancement of generative AI** presents challenges for presenting current research. This thesis was written between October 2023 and April 2024, during which time several relevant studies were published. Among others these include work on text simplification (Schomacker et al. 2023), research on Fine Tuning (Moslem et al. 2023), work on text complexity assessment (Ormaechea et al. 2023) and new findings on automatic prompts (Battle and Gollapudi 2024). These studies were published between December 15th, 2023 and February 20th, 2024. Consequently, new results had to be integrated into the thesis midway through the research process, which impacted previously completed steps that had to be revisited.

6.4 Future Work

Although this thesis provides a versatile evaluation of different enhancement methods for automatic text simplification in the context of Easy Language, further research opportunities exist.

6.4.1 Easy Language

Regarding Easy Language, there are several possibilities for improvement.

Rule P1 (NLS 2022) states that texts should always be **reviewed by experts**. Therefore, further research should include a group of certified experts, both with and without learning disabilities, to provide valuable insights into the evaluation.

As our thesis was based on a textual LLM, rule G15 (NLS 2022) was outside the scope of our study. This rule states that **pictures** need to be utilised to enhance the text's comprehensibility. During our research, LLMs have made significant advances in creating pictures and videos. This should be further investigated for the purpose of enhancing visualisation in Easy Language translations.

6.4.2 Parallel Corpora

As the base data websites were never intended to be used in this way, they do not live up to their full potential. The foundational data for enhancing LLMs are the Parallel Corpora, which **should be created by certified Easy Language editors** to increase translation quality.

Moreover, research could benefit from a **larger amount of parallel corpora**. Only 170 sensible example pairs were derived, but it is advisable to have over 200 examples. Furthermore, determining the **optimal amount of examples** for which LLM enhancement reaches its peak in terms of Easy Language translations could be a thesis topic.

6.4.3 Language Model

As previously stated, this research **only examines one LLM** due to limitations in time and workforce capacity. However, including the results of a second or third LLM could enhance the accuracy of evaluations and rankings. Additionally, upgrading the research from GPT to GPT-4 would be worthwile, as version 4 generally outperforms 3.5. However, an initial attempt at translating the example from Section 4.1 using GPT-4 has not yielded better results than GPT thus far: *"Alle Personen, die die deutsche Staatsbürgerschaft haben, müssen ab dem 16. Lebensjahr einen gültigen Personalausweis oder Reisepass besitzen."* (compare to Section 4.1).

6.4.4 Enhancement Methods

Chapter 5 demonstrates that some enhancement methods produce low-quality translations, particularly when used in combination. Therefore, **improving the selected methods** could be a promising approach. Embedding could benefit from higher quality base data, Fine Tuning from adjusting the number of epochs, and Prompt Engineering from prompt chaining (Wu et al. 2022). Section 5.1.2 and Section 5.1.3 demonstrate the similarity of their respective method's results. It may be worth researching how these four methods could be enhanced further and whether this would have an equal impact on all methods or if it would create a more significant difference in translation quality between them.

Human interaction with the LLM still improves translation quality. Future research may consider changing the translation process from a single loop of translating all 170 base texts to **an interactive process** that involves human interaction. This would allow for direct review of translations and the ability to trigger a new translation with detailed information on necessary improvements. In addition to research, this approach could be particularly useful for real-life production use cases of Easy Language translators.

An alternative approach could be to reduce human interaction with the LLM. Battle and Gollapudi (2024) suggest that optimising input prompts is more effective when left to AI models rather than humans. They discovered that **automatic prompt optimisation**, such as instructing AI models to refine prompts themselves, produces superior results to anything a human engineer could achieve. According to the research findings, it is recommended to provide basic instructions for the model to optimize the input prompt instead of writing them by hand.

6.4.5 Evaluation

Another approach could be to improve the evaluation of the translations regarding the **level of simplified language**. A more precise ranking of which translations reached Plain, Simple Plus or Simple Language could indicate where further improvements are necessary and to which extent.

As previously stated, a **standardised automatic evaluation framework for Easy Language** has not yet been found. Currently, evaluation relies on human interaction, which can be costly in terms of both time and budget. The implementation of a standardised automatic evaluation framework would greatly benefit research by allowing for the processing of larger amounts of data. Furthermore, this would enable the comparison of different research studies that are based on the same evaluation framework.

6.4.6 Multilingualism

Munich is a cosmopolitan economic metropolis with international connections. Around 30% of its 1.59 million inhabitants are foreign citizens from over 180 different nations, making it a popular destination for foreigners. The city hosts consular representations from over 100 countries (Munich 2024). Therefore a possible future research topic could be "Which steps are required to enable an LLM to translate **German complex texts into foreign Easy Language** without altering the content?"

6.5 Closing Summary

Although Easy Language translations can be created to a certain extent, we agree with Deilen et al. (2023) *“that in Easy Language translation, human translators are still indispensable”* and that *“there are still some tasks that require the translator’s specialised knowledge, creativity and understanding and awareness of the target group.”* This is particularly true when enhancement methods are combined.

In our opinion, a shift in research focus towards Easy Language and its automatic translation based on NLP has been observed in recent years. However, as stated in Section 6.4, this is just the beginning. Although new possibilities are being developed almost weekly in this field, there is still much to explore.

Despite the vast amount of new technical possibilities, we must not forget that we are developing automated language services for people to enable digital participation!

Acknowledgements

I would like to thank Torsten Zesch and Jeanette Bewersdorff for their constructive feedback and their pleasant guidance throughout the process of writing this thesis. I am also thankful to Elke Wildraut for her expertise in Easy Language and to Leon Lukas, Michael Jaumann and Stefanie Lämmle for introducing me to the world of AIs and LLMs. And last but not least, I would like to thank my family and friends for their support and inspiration.

Appendix

A muenchen.de article URLs

normal_url	**easy_url**
https://stadt.muenchen.de/infos/barrierefreiheit-erklaerung.html	https://stadt.muenchen.de/leichte-sprache/infos/barriere-freiheit-ls.html
https://stadt.muenchen.de/infos/datenschutzbeauftragte-direktorium.html	https://stadt.muenchen.de/leichte-sprache/infos/daten-schutz-ls.html
https://stadt.muenchen.de/infos/ukraine.html	https://stadt.muenchen.de/leichte-sprache/infos/ukraine-hilfe-ls.html
https://stadt.muenchen.de/infos/stadtinformation.html	https://stadt.muenchen.de/leichte-sprache/infos/stadt-information-ls.html
https://stadt.muenchen.de/infos/behoerdennummer-115.html	https://stadt.muenchen.de/leichte-sprache/infos/behoerden-nummer-ls.html
https://stadt.muenchen.de/rathaus/verwaltung/sozialreferat/sozialbuergerhaus.html	https://stadt.muenchen.de/leichte-sprache/infos/sozial-buerger-haus-ls.html
https://stadt.muenchen.de/terminvereinbarung_/terminvereinbarung_bb.html	https://stadt.muenchen.de/leichte-sprache/infos/buerger-buero-terminvereinbarung-ls.html
	https://stadt.muenchen.de/leichte-sprache/infos/buerger-buero-ls.html
https://stadt.muenchen.de/buergerservice/verkehr-mobilitaet/fahrzeuge.html	https://stadt.muenchen.de/leichte-sprache/infos/kfz-zulassung-ls.html
https://stadt.muenchen.de/buergerservice/verkehr-mobilitaet/fuehrerschein.html	https://stadt.muenchen.de/leichte-sprache/infos/fuehrerschein-stelle-ls.html
https://stadt.muenchen.de/infos/heirat-standesamt.html	https://stadt.muenchen.de/leichte-sprache/infos/heirats-buero-ls.html
https://stadt.muenchen.de/infos/portrait-heimaufsicht.html	https://stadt.muenchen.de/leichte-sprache/infos/heim-aufsicht-ls.html
https://stadt.muenchen.de/rathaus/politik/bezirksausschuss.html	https://stadt.muenchen.de/leichte-sprache/infos/bezirks-ausschuss-ls.html
https://stadt.muenchen.de/service/info/buergerinformation-der-bezirksausschuesse/1063621/	https://stadt.muenchen.de/leichte-sprache/infos/geschaefts-stellen-ls.html
https://stadt.muenchen.de/infos/behindertenbeirat.html	https://stadt.muenchen.de/leichte-sprache/infos/behinderten-beirat-ls.html
https://stadt.muenchen.de/infos/migrationsbeirat.html	https://stadt.muenchen.de/leichte-sprache/infos/migrations-beirat-ls.html
https://stadt.muenchen.de/infos/seniorenbeirat.html	https://stadt.muenchen.de/leichte-sprache/infos/senioren-beirat-ls.html
https://stadt.muenchen.de/rathaus/verwaltung/direktorium/lgbtiq-stelle.html	https://stadt.muenchen.de/leichte-sprache/infos/koordinierungs-stelle-ls.html
https://stadt.muenchen.de/infos/frauengleichstellungsstelle_start.html	https://stadt.muenchen.de/leichte-sprache/infos/gleichstellungs-stelle-fuer-frauen-ls.html
https://stadt.muenchen.de/infos/gesundheitstreffs.html	https://stadt.muenchen.de/leichte-sprache/infos/gesundheits-treffs-ls.html
https://stadt.muenchen.de/service/info/wohnsitz-anmelden-oder-ummelden/1063475/n0/	https://stadt.muenchen.de/leichte-sprache/infos/wohn-sitz-anmelden-ls.html
https://stadt.muenchen.de/service/info/wohnsitz-abmelden/1063486/n0/	https://stadt.muenchen.de/leichte-sprache/infos/wohn-sitz-abmelden-ls.html
https://stadt.muenchen.de/service/info/hauptabteilung-i-sicherheit-und-ordnung-praevention/1072014/	https://stadt.muenchen.de/leichte-sprache/infos/park-ausweis-mensch-mit-behinderung-ls.html
https://stadt.muenchen.de/infos/portrait-heimaufsicht.html	https://stadt.muenchen.de/leichte-sprache/infos/hilfe-von-heim-aufsicht-ls.html
https://stadt.muenchen.de/service/info/ska-4-2-grund-zweitwohnung-hundesteuer/1074315/	https://stadt.muenchen.de/leichte-sprache/infos/hund-anmelden-ls.html
https://stadt.muenchen.de/service/info/ska-4-2-grund-zweitwohnung-hundesteuer/1074318/	https://stadt.muenchen.de/leichte-sprache/infos/hund-abmelden-ls.html
https://stadt.muenchen.de/service/info/personalausweis-beantragen/1063441/n0/	https://stadt.muenchen.de/leichte-sprache/infos/personal-ausweis-beantragen-ls.html
https://stadt.muenchen.de/service/info/reisepass-beantragen/1063453/n0/	https://stadt.muenchen.de/leichte-sprache/infos/reise-pass-beantragen-ls.html
https://stadt.muenchen.de/service/info/vorlaeufiger-reisepass/1080582/n0/	https://stadt.muenchen.de/leichte-sprache/infos/vorlaeufigen-reise-pass-beantragen-ls.html
https://stadt.muenchen.de/service/info/kinderreisepass-bis-12-jahre-beantragen/1063464/n0/	https://stadt.muenchen.de/leichte-sprache/infos/kinder-reise-pass-beantragen-ls.html
https://stadt.muenchen.de/service/info/tiefbau/1075172/	https://stadt.muenchen.de/leichte-sprache/infos/muell-melden-ls.html
https://stadt.muenchen.de/service/info/muenchen-pass/1073956/n0/	https://stadt.muenchen.de/leichte-sprache/infos/muenchen-pass-beantragen-ls.html
https://stadt.muenchen.de/service/info/melderechtliche-bescheinigung/1063576/n0/	https://stadt.muenchen.de/leichte-sprache/infos/melde-bescheinigung-beantragen-ls.html
https://stadt.muenchen.de/service/info/medikamentenhilfe/10313859/n0/	https://stadt.muenchen.de/leichte-sprache/infos/medikamenten-hilfe-ls.html
https://stadt.muenchen.de/service/info/landeshauptstadt-muenchen/10313434/	https://stadt.muenchen.de/leichte-sprache/infos/taxi-gutscheine-frauen-ls.html
https://stadt.muenchen.de/infos/schwangerschaft_geburt.html	https://stadt.muenchen.de/leichte-sprache/infos/beratung-schwangere-ls.html
https://stadt.muenchen.de/infos/ungewollte-schwangerschaft.html	https://stadt.muenchen.de/leichte-sprache/infos/beratung-ungewollt-schwangere-ls.html
https://stadt.muenchen.de/service/info/sg-beratungsstellen-sti-prostschg/10322180/	https://stadt.muenchen.de/leichte-sprache/infos/sprech-stunde-frauen-arzt-ls.html
https://stadt.muenchen.de/infos/buergerversammlungen.html	https://stadt.muenchen.de/leichte-sprache/infos/buerger-versammlung-ls.html

B Parallel Corpora

Normal Language	Easy Language
Öffentliche Stellen sind gemäß der Richtlinie (EU) 2016/2102 des Europäischen Parlaments dazu verpflichtet, Webauftritte barrierefrei zugänglich zu machen.	Diese Internet-Seite soll barriere-frei werden.\nDas heißt:\nAlle Menschen sollen die Internet-Seite ohne Hilfe nutzen können.\nAlso zum Beispiel auch:\n* blinde Menschen\n* gehörlose Menschen\n* Menschen, die nicht alle Finger bewegen können.
Die Stadt München arbeitet beständig an der Verbesserung ihres Webauftritts und konnte aufgrund der Fülle des Materials und der Komplexität der Site noch nicht alle Inhalte und Services digital barrierefrei gestalten.	Die Stadt München arbeitet immer weiter daran.\n\nDie Internet-Seite soll noch weniger Barrieren haben.
Sie können Mängel bei der Einhaltung der Anforderungen an die Barrierefreiheit mitteilen oder Informationen, die nicht barrierefrei dargestellt werden müssen, barrierefrei anfordern.	Vielleicht brauchen Sie Informationen von der Internet-Seite.\nAber Sie können die Informationen nicht richtig bekommen.\nWeil es noch Barrieren gibt.\nZum Beispiel:\nIhr Computer kann ein wichtiges PDF-Dokument nicht vorlesen.\n\nManche Texte bekommt man im Internet nur als PDF-Dokument.\nPDF heißt hier:\nMan kann diese Texte hin und herschicken.\nDiese Texte schauen auf jedem Computer gleich aus.\nMan kann diese Texte aber nicht verändern.\n\nDas ist bei Microsoft Word anders.\nDiese Texte kann man immer verändern.\nDafür sehen Word-Dokumente nicht auf\njedem Computer immer gleich aus.\n\nKönnen Sie ein Dokument nicht lesen?\nDann können Sie uns schreiben.\n\nSchreiben Sie uns dafür eine E-Mail.\nUnsere E-Mail-Adresse ist:\nwebmanagement@muenchen.de
Durchsetzungsverfahren\n\nBleibt eine Anfrage an die Kontaktstelle innerhalb von sechs Wochen ganz oder teilweise unbeantwortet, prüft die zuständige Aufsichtsbehörde auf Antrag der betroffenen Nutzer*innen, ob im Rahmen der Überwachung gegenüber den Betreibern des Webangebots Maßnahmen erforderlich sind. Die für das Durchsetzungsverfahren zuständigen Aufsicht ist das Landesamt für Digitalisierung, Breitband und Vermessung:\n\nLandesamt für Digitalisierung, Breitband und Vermessung\nAlexandrastraße 4\n80538 München\nTelefon: 089 2129-1111\nFax: 089 2129-1113\nE-Mail: service@geodaten.bayern.de	Dauert die Antwort länger als 6 Wochen?\nDann können Sie sich bei dieser Stelle beschweren:\nLandesamt für Digitalisierung, Breitband und Vermessung\n\nSie können einen Brief schreiben.\nDie Adresse ist:\nLandesamt für Digitalisierung, Breitband und Vermessung\nAlexandrastraße 4\n80538 München\n\nSie können auch anrufen.\nDie Telefon-Nummer ist:\n0 89 - 21 29 11 11\n\nSie können auch eine E-Mail schreiben.\nDie E-Mail-Adresse ist:\nservice@geodaten.bayern.de

Wenn Ihre personenbezogenen Daten von der Landeshauptstadt München verarbeitet werden, können Sie als betroffene Person im Rahmen der gesetzlichen Regelungen von der Landeshauptstadt München Auskunft über die Sie betreffenden personenbezogenen Daten erhalten. Sind Daten unrichtig, so können Sie unter Umständen Berichtigung verlangen. Als betroffene Person haben Sie ferner unter bestimmten Voraussetzungen das gesetzliche Recht, die Daten löschen oder sperren zu lassen.	Was passiert mit Ihren Daten?\n\nWir gebrauchen Ihre Daten nur für unsere Arbeit.\n\nIhre Daten sind zum Beispiel:\n\n* Ihr Name\n\n* Ihr Geburts-Datum\n\n* Ihre Adresse\n\n\n\nEs passiert nichts damit, was Sie nicht wollen.\n\nSie können sagen:\n\nIch will nicht, dass andere meine Daten bekommen.\n\nDas nennt man Widerspruch.\n\nDann geben wir niemals Ihre Daten an andere weiter.\n\nSie können den Widerspruch aber auch erst später machen.\n\nWir dürfen wir Ihre Daten nur weitergeben,\n\nwenn Sie das wollen.
In der Stadt-Information im Rathaus erhalten Sie neben Printmedien, Rat und Hilfeleistung verschiedenster Art.	Die Mitarbeiter von der Stadt-Information beantworten Ihre Fragen.\n\n\n\nBei der Stadt-Information bekommen Sie Informations-Hefte.\n\n\n\nBei der Stadt-Information bekommen Sie Formulare.\n\n\n\nDie Stadt-Information finden Sie im Rat-Haus von München.\n\nDas Rat-Haus ist am Marienplatz.\n\nDer Marienplatz ist im Zentrum von München.
In der Stadt-Information im Rathaus erhalten Sie Prospekte, Informationsmaterial und Formulare zu unterschiedlichen Themen. Die Mitarbeiter stehen den Besucherinnen und Besucher mit Rat und Hilfeleistung verschiedenster Art zur Verfügung.	Die Mitarbeiter von der Stadt-Information beantworten Ihre Fragen.\n\nZum Beispiel:\n\n* Wo bekomme ich einen Reise-Pass?\n\n* Wo kann ich mein Auto anmelden?\n\n* Wo ist das Deutsche Museum?\n\n* Welche Konzerte und Ausstellungen gibt es gerade in München?\n\n* Wo ist der Englische Garten?\n\n* Das ist ein großer Park in München.\n\n\n\nBei der Stadt-Information bekommen Sie Informations-Hefte.\n\nZum Beispiel:\n\n* Informations-Hefte über den Stadt-Rat\n\n* Informations-Hefte über den Behinderten-Beirat\n\n* Informations-Hefte über die Bezirks-Ausschüsse\n\n* Informations-Hefte über Schulen in München\n\n* Informations-Hefte über Angebote für ältere Menschen\n\n* Informations-Hefte über Umwelt-Schutz\n\n* Informations-Hefte über Freizeit-Angebote\n\n* Veranstaltungs-Programme\n\nDie Informations-Hefte kosten nichts.\n\n\n\nBei der Stadt-Information bekommen Sie Formulare.\n\nZum Beispiel:\n\n* Patienten-Verfügung\n\n* Vorsorge-Vollmacht\n\n* Betreuungs-Verfügung\n\n* Antrag auf geförderte Wohnung\n\n* Wohn-Geld-Antrag
Wohnsitz	Der Wohn-Sitz ist der Ort, wo Sie wohnen.\nZum Beispiel:\n* eine Wohnung\n* ein Haus\n* ein Wohn-Heim
Gebührenrahmen\nkostenlos	Sie müssen nichts bezahlen.
Hundesteuer	Hunde-Steuer
Personalausweis	Personal-Ausweis

Bei der Behördennummer wird der persönliche Kontakt groß geschrieben. Von Montag bis Freitag, 8 bis 18 Uhr, können den Mitarbeitenden der 115-Servicecenter Fragen gestellt werden. Die meisten Anliegen werden direkt beantwortet. Falls dies einmal nicht möglich ist, verspricht die 115 eine schnelle Rückmeldung.	\nSie haben eine Frage an die Stadt-Verwaltung von München?\n\nDann können Sie die Behörden-Nummer 115 anrufen.\nEgal wo Sie sind.\nEgal welche Frage Sie haben.\n\nDie 115 ist eine Telefon-Nummer.\nMan sagt auch: Service-Nummer.\nService ist ein englisches Wort.\nService spricht man Sör-wis.\nService bedeutet Hilfe.\nDie Hilfe ist für alle Menschen.
Mit einer leicht zu merkenden Rufnummer erhalten Bürger*innen bundesweit einen direkten Draht zur Verwaltung. Wird eine Frage wie z.B. zum Wohnsitz anmelden, Ausweis beantragen oder Fahrzeug ummelden, gestellt, beantworten die Mitarbeitenden des Service-Centers der Landeshauptstadt München gerne Ihre Anliegen.	Sie haben eine Frage an die Stadt-Verwaltung von München?\nDann können Sie die Behörden-Nummer 115 anrufen.\nEgal wo Sie sind.\nEgal welche Frage Sie haben.\nDie 115 ist eine Telefon-Nummer.\nMan sagt auch: Service-Nummer.\nService ist ein englisches Wort.\nService spricht man Sör-wis.\nService bedeutet Hilfe.\nDie Hilfe ist für alle Menschen.\nSie bekommen meistens sofort eine Antwort.\n Am Telefon.\nZum Beispiel auf diese Fragen:\n* Ich ziehe nach München.\n Wo kann ich mich anmelden?\n* Wo bekomme ich einen neuen Reise-Pass?\n* Ich habe ein neues Auto.\n Ich muss zur Zulassungs-Stelle.\n Wann hat die Zulassungs-Stelle geöffnet?
Komplexere Anfragen werden durch ein eigens dafür eingerichtetes Back-Office bearbeitet oder an die Fachbereiche weitergeleitet und es erfolgt eine Rückmeldung.	Manchmal ist die Frage besonders schwierig.Dann sagt der Mitarbeiter zu Ihnen:Soll ich mich darum kümmern,dass Sie eine Antwort bekommen?Dafür braucht der Mitarbeiter Ihre E-Mail-Adresse.Oder Ihre Telefon-Nummer.Sie bekommen die Antwort.Als E-Mail.Oder ein Mitarbeiter ruft Sie an.Von der Stadt-Verwaltung.Das kann 1 Tag dauern.
Die 115 erteilt ausschließlich behördliche Auskünfte. Fragen zu wie z.B. Wegbeschreibungen, Wetterauskünften oder touristischen Informationen können unter der 115 nicht erteilt werden.	Sie können die Behörden-Nummer 115 anrufen.Sie bekommen Informationen zur Stadt-Verwaltung von München.

C NLS Easy Language Rules

```
{"role": "system", "content": "Benutzen Sie einfache Wörter.\nBeispiel\nSchlecht:
↪ genehmigen\nGut: erlauben"}
{"role": "system", "content": "Benutzen Sie Wörter, die etwas genau
↪ beschreiben.\nBeispiel\nSchlecht: Öffentlicher Nahverkehr\nGut: Bus und Bahn"}
{"role": "system", "content": "Benutzen Sie bekannte Wörter.\nVerzichten Sie auf
↪ Fachwörter und Fremdwörter.\nBeispiel\nSchlecht: Workshop\nGut:
↪ Arbeits-Gruppe\nErklären Sie schwere Wörter.\nKündigen Sie schwere Wörter
↪ an.\nSie können am Ende vom Text ein Wörterbuch machen.\nBeispiel\nGut:\nHerr
↪ Meier hatte einen schweren Unfall.\nJetzt lernt er einen anderen Beruf.\nDas
↪ schwere Wort dafür ist:\nberufliche Rehabilitation."}
{"role": "system", "content": "Benutzen Sie immer die gleichen Wörter für die
↪ gleichen Dinge.\nZum Beispiel:\nSie schreiben über ein Medikament.\nBenutzen Sie
↪ immer ein Wort. Zum Beispiel: Tablette.\nWechseln Sie nicht zwischen Tablette und
↪ Pille."}
{"role": "system", "content": "Benutzen Sie kurze Wörter.\nBeispiel\nSchlecht:
↪ Omnibus\nGut: Bus\nWenn das nicht geht:\nTrennen Sie lange Wörter mit einem
↪ Bindestrich.\nDann kann man die Wörter besser lesen.\nBeispiel\nSchlecht:
↪ Bundesgleichstellungsgesetz\nGut: Bundes-Gleichstellungs-Gesetz"}
{"role": "system", "content": "Verzichten Sie auf Abkürzungen.\nBeispiel
↪ d.h.\nSchlecht:\nGut: das heißt\nEs gibt aber Ausnahmen:\nManche Abkürzungen sind
↪ sehr bekannt.\nZum Beispiel:\n* WC\n* LKW\n* Dr.\n* ICE"}
{"role": "system", "content": "Benutzen Sie Verben.\nVerben sind
↪ Tu-Wörter.\nVermeiden Sie Haupt-Wörter.\nBeispiel\nSchlecht: Morgen ist die Wahl
↪ zum Heim-Beirat.\nGut: Morgen wählen wir den Heim-Beirat."}
{"role": "system", "content": "Benutzen Sie aktive Wörter.\nBeispiel\nSchlecht:
↪ Morgen wird der Heim-Beirat gewählt.\nGut: Morgen wählen wir den Heim-Beirat."}
{"role": "system", "content": "Vermeiden Sie den Genitiv.\nDen Genitiv erkennt man
↪ oft an dem Wort: des.\nBenutzen Sie lieber die Wörter:\nvon, von dem oder
↪ vom.\nBeispiel\nSchlecht: Das Haus des Lehrers.\nDes Lehrers Haus.\nGut: Das Haus
↪ von dem Lehrer.\nDas Haus vom Lehrer."}
{"role": "system", "content": "Vermeiden Sie den Konjunktiv.\nDen Konjunktiv erkennt
↪ man an diesen Wörtern:\nhätte, könnte, müsste, sollte, wäre,
↪ würde.\nBeispiel\nSchlecht: Morgen könnte es regnen.\nGut: Morgen regnet es
↪ vielleicht."}
```

```
{"role": "system", "content": "Benutzen Sie möglichst eine positive Sprache.\nPositive Sprache ist zum Beispiel: Peter steht.\nNegative Sprache ist zum Beispiel: Peter sitzt nicht.\nNegative Sprache erkennt man oft am Wort: nicht.\nBeispiel\nSchlecht: Reg Dich nicht auf.\nGut: Bleib ruhig.\nBeispiel\nSchlecht: Der Schlüssel ist nirgendwo.\nGut: Der Schlüssel ist weg.\nBeispiel\nSchlecht: Das war nicht schlecht.\nGut: Das war gut.\nAber manchmal braucht man negative Sprache.\nDann benutzen Sie möglichst diese Wörter:\n* nein\n* nicht\n* nichts\n* nie\n* kein\n* keine\n* keiner"}
{"role": "system", "content": "Vermeiden Sie Redewendungen und bildliche Sprache.\nViele Menschen verstehen das falsch.\nSie verstehen diese Sprache wörtlich.\nZum Beispiel:\nDas Wort Raben-Eltern ist bildliche Sprache.\nRaben-Eltern sind nicht die Eltern von Raben-Küken.\nMit Raben-Eltern meint man: schlechte Eltern."}
{"role": "system", "content": "Schreiben Sie Zahlen so, wie die meisten Menschen sie kennen.\nBeispiel\nSchlecht: römische Zahlen. Zum Beispiel: IX\nGut: arabische Zahlen. Zum Beispiel: 9"}
{"role": "system", "content": "Vermeiden Sie alte Jahres-Zahlen.\nBeispiel\nSchlecht: 1867\nGut: Vor langer Zeit.\nOder: Vor mehr als 100 Jahren."}
{"role": "system", "content": "Vermeiden Sie hohe Zahlen und Prozent-Zahlen.\nBenutzen Sie Vergleiche oder ungenaue Angaben.\nBeispiel\nSchlecht: 14.795 Menschen\nGut: Viele Menschen\nWenn es genauer sein soll, schreiben Sie dazu:\nFast 15 Tausend Menschen.\nBeispiel\nSchlecht: 14%\nGut: Einige oder wenige"}
{"role": "system", "content": "Wie sollen Sie Zahlen schreiben?\nMeistens sind Ziffern leichter als Worte.\nBeispiel\nSchlecht: Fünf Frauen\nGut: 5 Frauen"}
{"role": "system", "content": "Wie sollen Sie ein Datum schreiben?\nBeispiel\nSchlecht: 03.03.12\nGut: 3. März 2012 oder 3.3.2012"}
{"role": "system", "content": "Wie sollen Sie Uhr-Zeiten schreiben?\nZum Beispiel:\n* 11:00 Uhr\n* 11 Uhr\n* 11.00 Uhr\n* 11:45 Uhr\n* 11 Uhr 45\n* 11.45 Uhr\n* 6 Uhr abends\n* 18:00 Uhr\n* 18.00 Uhr"}
{"role": "system", "content": "Wie sollen Sie Zeit-Angaben schreiben?\nZum Beispiel:\n* Am Ende vom Monat\n* Am 31. Dezember\n* Zum Monats-Ende"}
{"role": "system", "content": "Schreiben Sie Telefon-Nummern mit Leerzeichen.\nBeispiel\nSchlecht: Tel.: (05544) 332211\n05544 / 332211\nGut: Telefon: 0 55 44 33 22 11\n0 55 44 - 33 22 11"}
{"role": "system", "content": "Vermeiden Sie Sonder-Zeichen.\nBeispiel\nSchlecht:\n■ ■ Anführungs-Striche\n% Prozent\n... Punkt Punkt Punkt\n;\nStrich-Punkt\n& Und\n( ) Klammern\n■ Paragraf\nWenn Sie ein Sonder-Zeichen benutzen müssen:\nDann erklären Sie das Zeichen.\nBeispiel\nGut:\nEin Paragraf ist ein Teil in einem Gesetz.\nDas Zeichen für Paragraf ist: ■\nJeder Paragraf hat eine Nummer.\nSie können auch das Wort und das Zeichen schreiben:\nZum Beispiel:\nParagraf ■1"}
```

```
{"role": "system", "content": "Benutzen Sie kurze Sätze.\nMachen Sie in jedem Satz nur eine Aussage.\nBeispiel\nSchlecht: Ich habe meinem guten Freund Leo ein Buch über die Geschichte von Berlin geliehen.\nGut: Leo ist ein guter Freund von mir.\nIch habe ihm ein Buch geliehen.\nDas Buch ist über die Geschichte von Berlin.\nBeispiel\nSchlecht: Das Buch, das auf dem Tisch liegt, habe ich schon gelesen.\nGut: Auf dem Tisch liegt ein Buch.\nIch habe das Buch schon gelesen."}
{"role": "system", "content": "Benutzen Sie einen einfachen Satzbau.\nSchreiben Sie die Wörter in dieser Reihenfolge:\n* Erst steht:\nWer macht etwas?\n* Dann steht:\nWas macht die Person?\nBeispiel\nSchlecht: Die Rechnung bezahlt Frau Weber.\nGut: Frau Weber bezahlt die Rechnung."}
{"role": "system", "content": "Sie dürfen verkürzte Sätze benutzen.\nVerkürzt heißt:\nDer Satz muss nicht vollständig sein.\nDer Satz darf mit diesen Wörtern anfangen:\n* Oder\n* Und\n* Aber\nBeispiel Oder\nSchlecht: Wollen Sie nach Berlin oder nach Hamburg fahren?\nGut: Wollen Sie nach Berlin fahren?\nOder nach Hamburg?\nBeispiel Und\nSchlecht: Menschen mit Behinderung wollen mitreden und mitbestimmen.\nGut: Menschen mit Behinderung wollen mitreden.\nUnd mitbestimmen.\nBeispiel Aber\nSchlecht: Ali war müde vom Sport, aber auch glücklich.\nGut: Ali war müde vom Sport.\nAber glücklich."}
```

```
{"role": "system", "content": "Trennen Sie lange Sätze.\nVermeiden Sie Neben-Sätze.\nBenutzen Sie besser mehrere Haupt-Sätze.\nWie erkenne ich einen Neben-Satz?\n1. Neben-Sätze erkennen Sie an einem Komma.\nDas ist ein Komma: ,\n2. Neben-Sätze erkennen Sie oft an diesen Wörtern:\n* damit\n* obwohl\n* weil\n* dass\n* ob\n* wenn\n* falls\n* als\n* während\n* bevor\n* nachdem\n* sobald\n* solange\n* so dass\nWir haben einige Beispiele aufgeschrieben.\nBeispiel für damit-Sätze\nSchlecht: Wir gehen heute früh ins Bett, damit wir morgen ausgeschlafen sind.\nGut: Wir wollen morgen ausgeschlafen sein.\nDeshalb gehen wir heute früh ins Bett.\nBeispiel für obwohl-Sätze\nSchlecht: Ich mache einen Spaziergang, obwohl es regnet.\nGut: Es regnet.\nIch mache trotzdem einen Spaziergang.\nBeispiel für weil-Sätze\nSchlecht: Die Suppe schmeckt schlecht, weil in der Suppe zu viel Salz ist.\nGut: In der Suppe ist zu viel Salz.\nDeshalb schmeckt die Suppe schlecht.\nBeispiel für Sätze mit so dass\nSchlecht: Tobias hat so lange geschlafen, dass er zu spät bei der Arbeit war.\nGut: Tobias hat sehr lange geschlafen.\nDeshalb war er zu spät bei der Arbeit.\nBeispiel für Zeit-Sätze mit während\nSchlecht: Während wir den Film sahen, haben wir Popcorn gegessen.\nGut: Wir haben den Film gesehen.\nDabei haben wir Popcorn gegessen.\nBeispiel für Zeit-Sätze mit bevor\nSchlecht: Sie müssen diese Medizin nehmen, bevor Sie etwas essen.\nGut: Nehmen Sie erst die Medizin.\nDann essen Sie etwas.\nBeispiel für Zeit-Sätze mit nachdem\nSchlecht: Ich gehe ins Kino, nachdem ich die Küche aufgeräumt habe.\nGut: Ich räume die Küche auf.\nDanach gehe ich ins Kino.\nBeispiele für Sätze mit wenn:\nEs gibt 2 Arten von wenn-Sätzen.\nDie erste Art\nBei der ersten Art geht es darum:\nDie Sachen passieren vielleicht.\nBeispiele für die erste Art:\nBeispiel für Sätze mit der Bedeutung vielleicht\nSchlecht: Wenn Sie ins Theater gehen möchten, dann müssen Sie Karten kaufen.\nGut: Möchten Sie ins Theater gehen?\nDann müssen Sie Karten kaufen.\nBeispiel für Sätze mit der Bedeutung vielleicht\nSchlecht: Wenn morgen die Sone scheint, geht Klara ins Freibad.\nGut: Vielleicht scheint morgen die Sone.\nDann geht Klara ins Freibad.\nBei der zweiten Art\nBei der zweiten Art geht es darum:\nDie Sachen passieren auf jeden Fall.\nBeispiel für die zweite Art:\nWenn ich erwachsen bin, dann möchte ich im Zoo arbeiten.\nBeispiel für Sätze mit der Bedeutung auf jeden Fall\nSchlecht: Vielleicht bin ich erwachsen.\nDann möchte ich im Zoo arbeiten.\nAuch schlecht: Bin ich erwachsen?\nDann möchte ich im Zoo arbeiten.\nGut: Ich möchte später im Zoo arbeiten.\nDafür muss ich erst erwachsen sein.\nOder Sie können einen Wenn-Satz benutzen:\nWenn ich erwachsen bin, dann möchte ich im Zoo arbeiten.\nManchmal müssen Sie die Sätze auch ganz anders schreiben.\nBeispiel\nSchlecht: Wenn Sie mir sagen, was Sie wünschen, kann ich Ihnen helfen.\nGut: Ich kann Ihnen helfen.\nBitte sagen Sie mir:\nWas wünschen Sie?"}
```

```
{"role": "system", "content": "Sprechen Sie die Leser und Leserinnen persönlich an.\nBeispiel\nSchlecht: Morgen ist die Wahl.\nGut: Sie dürfen morgen wählen.\nBenutzen Sie die Anrede Sie.\nWann geht die Anrede Du?\n* Bei Kindern\n* Oder Sie kennen die Leser und Leserinnen.\nUnd Sie duzen diese Person auch sonst.\nVielleicht benutzen Sie die weibliche und männliche Form.\nDann schreiben Sie immer zuerst die männliche Form.\nSo kann man es leichter lesen.\nBeispiel\nSchlecht: Mitarbeiterinnen und Mitarbeiter\nGut: Mitarbeiter und Mitarbeiterinnen"}
{"role": "system", "content": "Vermeiden Sie Fragen im Text.\nManche Menschen fühlen sich dadurch belehrt.\nManche Menschen denken:\nSie müssen darauf antworten.\nAber: Fragen als Überschrift sind manchmal gut."}
{"role": "system", "content": "Schreiben Sie alles zusammen, was zusammen gehört.\nVermeiden Sie Verweise.\nVerweisen Sie nicht auf andere Stellen im Text.\nVerweisen Sie nicht auf andere Texte.\nDas schwere Wort dafür heißt: Quer-Verweis.\nWenn Sie doch einen Verweis machen:\n* Heben Sie ihn gut hervor.\n* Erklären Sie ihn genau.\nBeispielnSchlecht: (siehe: Heft 3)\nGut: In Heft 3 steht mehr dazu."}
{"role": "system", "content": "Sie dürfen einen Text beim Schreiben in Leichter Sprache verändern.\nInhalt und Sinn müssen aber stimmen.\nZum Beispiel:\n* Sie dürfen Dinge erklären.\nDann versteht man sie besser.\n* Sie dürfen Hinweise geben.\n* Sie dürfen Beispiele schreiben.\n* Sie dürfen die Reihenfolge ändern.\n* Sie dürfen das Aussehen ändern.\n* Sie dürfen Teile vom Text weg lassen, wenn diese Teile nicht wichtig sind."}
{"role": "system", "content": "Benutzen Sie eine einfache Schrift.\nDie Schrift muss gerade sein.\nBeispiel\nSchlecht: Times New Roman\nArial Kursiv\nCourier\nLucida Handwritimg\nGut: Calibri\nMyriad Pro\nOpen Sans\nVerdana\nBenutzen Sie am besten nur eine Schriftart.\nZu viele Schriftarten verwirren."}
{"role": "system", "content": "Benutzen Sie eine große Schrift.\nWenn Sie zum Beispiel die Schriftart Open Sans nehmen:\nBenutzen Sie die Schriftgröße 14 oder größer.\nManche Schriftarten sind sehr klein.\nDann müssen Sie eine größere Schriftgröße nehmen."}
{"role": "system", "content": "Lassen Sie genug Abstand zwischen den Zeilen.\nSchlecht: Dieser Satz hat einen Zeilen-Abstand von 1.\nMan sagt auch: Einfacher Zeilen-Abstand.\nDas ist sehr eng.\nGut: Dieser Satz hat einen Zeilen-Abstand von 1,5.\nMan sagt auch: 1,5-facher Zeilen-Abstand.\nDas ist besser."}
{"role": "system", "content": "Schreiben Sie immer links-bündig.\nSchreiben Sie nicht Blocksatz.\nSchreiben Sie nicht rechts-bündig.\nSchreiben Sie nicht zentriert.\nAusnahme:\nDie Überschrift darf vielleicht in der Mitte stehen."}
{"role": "system", "content": "Schreiben Sie jeden neuen Satz in eine neue Zeile.\nBeispiel\nSchlecht: Das Spiel ist ab 18.00 Uhr und geht bis 22.00 Uhr. Die Halle öffnet um 16.00 Uhr.\nGut: Die Halle öffnet um 16.00 Uhr.\nDas Spiel ist ab 18.00 Uhr.\nEs geht bis 22.00 Uhr."}
```

```
{"role": "system", "content": "Trennen Sie keine Wörter am Ende einer Zeile.\nBeispiel\nSchlecht: Der letzte Urlaub auf Mallorca war ein Er-\nlebnis.\nGut: Der letzte Urlaub auf Mallorca\nwar ein Erlebnis."}
{"role": "system", "content": "Schreiben Sie alle Wörter in eine Zeile, die vom Sinn her zusammen gehören.\nBeispielnSchlecht: Wir sagen: Leichte\nSprache ist für alle gut.\nGut: Wir sagen:\nLeichte Sprache ist für alle gut."}
{"role": "system", "content": "Lassen Sie den Satz zusammen.\nManchmal ist die Seite voll.\nDer Satz ist aber noch nicht zu Ende.\nSchreiben Sie den ganzen Satz auf die nächste Seite.\nNoch besser: Lassen Sie den Absatz zusammen."}
{"role": "system", "content": "Machen Sie viele Absätze und Überschriften.\nBeispiel\nSchlecht: Im Winter fällt Schnee.\nUnd es ist kalt.\nIm Sommer scheint die Sonne.\nDann ist es wärmer.\n\nGut: Winter:\nIm Winter fällt Schnee.\nUnd es ist kalt.\nSommer:\nIm Sommer scheint die Sonne.\nDann ist es wärmer."}
{"role": "system", "content": "Schreiben Sie eine Adresse so wie auf einem Brief.\nSo kann man die Adresse besser verstehen.\nUnd abschreiben.\nBeispiel\nSchlecht: Frau Tanja Muster, Alte Mustergasse 10, 12345 Musterstadt, Musterland\nGut: Frau\nTanja Muster\nAlte Mustergasse 10\n12345 Musterstadt\nMusterland"}
{"role": "system", "content": "Heben Sie wichtige Dinge hervor.\nBeispiel\nSchlecht:\n* NUR GROßE BUCHSTABEN\n* Kursive oder schräg gestellte Schrift\n* Größerer Zeichen-Abstand\nGut:\n* Setzen Sie Aufzählungs-Punkte.\n* Machen Sie ein Wort fett.\n* Nehmen Sie eine andere dunkle Schrift-Farbe.\n* Hinterlegen Sie den Text mit einer hellen Farbe.\n* Aber man soll die Schrift trotzdem gut lesen können.\n* Auch nach dem Kopieren.\n* Machen Sie um einen Satz einen Rahmen.\n* Unterstreichen Sie so wenig wie möglich."}
{"role": "system", "content": "Benutzen Sie dunkle Schrift.\nUnd helles Papier."}
{"role": "system", "content": "Benutzen Sie dickes Papier.\nNehmen Sie Papier mit der Stärke von 80 Gramm oder mehr.\nDas Papier darf nicht dünner sein.\nBei dünnem Papier kann die Schrift durchscheinen."}
{"role": "system", "content": "Nehmen Sie mattes Papier.\nGlänzendes Papier spiegelt.\nDann kann man den Text schlechter lesen."}
{"role": "system", "content": "Benutzen Sie Bilder.\nBilder helfen Texte zu verstehen.\nDie Bilder müssen zum Text passen."}
{"role": "system", "content": "Benutzen Sie scharfe und klare Bilder.\nMan muss die Bilder gut erkennen.\nZum Beispiel nach dem Kopieren."}
{"role": "system", "content": "Benutzen Sie Bilder nicht als Hintergrund.\nDann kann man den Text schlecht lesen."}
```

List of Figures

List of Tables

Bibliography

Agrawal and Carpuat 2023
AGRAWAL, Sweta ; CARPUAT, Marine: *Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension.* http://dx.doi.org/10.48550/arXiv.2312.10126. Version: dec 2023. – arXiv:2312.10126 [cs]

Allen 2009
ALLEN, David: A study of the role of relative clauses in the simplification of news texts for learners of English. In: *System* 37 (2009), Nr. 4, p. 585–599

Alva-Manchego et al. 2021
ALVA-MANCHEGO, Fernando ; SCARTON, Carolina ; SPECIA, Lucia: The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. In: *Computational Linguistics* 47 (2021), dec, Nr. 4, 861–889. http://dx.doi.org/10.1162/coli_a_00418. – DOI 10.1162/coli$_{a0}$0418. – $-ISSN$0891 – –2017

Anschütz et al. 2023
ANSCHÜTZ, Miriam ; OEHMS, Joshua ; WIMMER, Thomas ; JEZIERSKI, Bartłomiej ; GROH, Georg: Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training. In: *Findings of the Association for Computational Linguistics: ACL 2023.* Toronto, Canada : Association for Computational Linguistics, jul 2023, 1147–1158

Banerjee and Lavie 2005
BANERJEE, Satanjeev ; LAVIE, Alon: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, p. 65–72

Bang et al. 2023
BANG, Yejin ; CAHYAWIJAYA, Samuel ; LEE, Nayeon ; DAI, Wenliang ; SU, Dan ; WILIE, Bryan ; LOVENIA, Holy ; JI, Ziwei ; YU, Tiezheng ; CHUNG, Willy ; DO, Quyet V. ; XU, Yan ; FUNG, Pascale: *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity.* http://dx.doi.org/10.48550/arXiv.2302.04023. Version: nov 2023. – arXiv:2302.04023 [cs]

Barthélemy et al. 2022
BARTHÉLEMY, Florian ; GHESQUIÈRE, Nathan ; LOOZEN, Nicolas ; MATHA, Louis ; STANI, Emidio: Natural Language Processing for Public Services. (2022)

Barzilay and Elhadad 2003
Barzilay, Regina ; Elhadad, Noemie: Sentence Alignment for Monolingual Comparable Corpora. In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, 25–32

Battle and Gollapudi 2024
Battle, Rick ; Gollapudi, Teja: The Unreasonable Effectiveness of Eccentric Automatic Prompts. In: *arXiv preprint arXiv:2402.10949* (2024)

Beauchemin et al. 2023
Beauchemin, David ; Saggion, Horacio ; Khoury, Richard: MeaningBERT: assessing meaning preservation between sentences. In: *Frontiers in Artificial Intelligence* 6 (2023), sep, 1223924. `http://dx.doi.org/10.3389/frai.2023.1223924`. – DOI 10.3389/frai.2023.1223924. – ISSN 2624–8212

Björnsson 1968
Björnsson, Carl-Hugo: *Läsbarhet: Lesbarkeit durch Lix.(Aus dem Schwedischen)*. Liber, 1968

Brown et al. 2020
Brown, Tom ; Mann, Benjamin ; Ryder, Nick ; Subbiah, Melanie ; Kaplan, Jared D. ; Dhariwal, Prafulla ; Neelakantan, Arvind ; Shyam, Pranav ; Sastry, Girish ; Askell, Amanda ; Agarwal, Sandhini ; Herbert-Voss, Ariel ; Krueger, Gretchen ; Henighan, Tom ; Child, Rewon ; Ramesh, Aditya ; Ziegler, Daniel ; Wu, Jeffrey ; Winter, Clemens ; Hesse, Chris ; Chen, Mark ; Sigler, Eric ; Litwin, Mateusz ; Gray, Scott ; Chess, Benjamin ; Clark, Jack ; Berner, Christopher ; McCandlish, Sam ; Radford, Alec ; Sutskever, Ilya ; Amodei, Dario: Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems* Bd. 33, Curran Associates, Inc., 2020, 1877–1901

Carroll et al. 1999
Carroll, John A. ; Minnen, Guido ; Pearce, Darren ; Canning, Yvonne ; Devlin, Siobhan ; Tait, John: Simplifying text for language-impaired readers. In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, 1999, p. 269–270

Chen et al. 2021
Chen, Shijie ; Zhang, Yu ; Yang, Qiang: Multi-task learning in natural language processing: An overview. In: *arXiv preprint arXiv:2109.09138* (2021)

De Belder and Moens 2010
De Belder, Jan ; Moens, Marie-Francine: Text simplification for children. In: *Prroceedings of the SIGIR workshop on accessible search systems*, ACM; New York, 2010, p. 19–26

Deilen et al. 2023
Deilen, Silvana ; Garrido, Sergio H. ; Lapshinova-Koltunski, Ekaterina ; Maass, Christiane: *Using ChatGPT as a CAT tool in Easy Language translation*. http://dx.doi.org/10.48550/arXiv.2308.11563. Version: aug 2023. – arXiv:2308.11563 [cs]

Devaraj et al. 2022
Devaraj, Ashwin ; Sheffield, William ; Wallace, Byron C. ; Li, Junyi J.: Evaluating Factuality in Text Simplification. In: *Proceedings of the conference. Association for Computational Linguistics. Meeting* 2022 (2022), may, 7331–7345. http://dx.doi.org/10.18653/v1/2022.acl-long.506. – DOI 10.18653/v1/2022.acl–long.506. – ISSN 0736–587X

Devlin et al. 2019
Devlin, Jacob ; Chang, Ming-Wei ; Lee, Kenton ; Toutanova, Kristina: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *arXiv preprint arXiv:1810.04805* (2019)

Djeffal and Horst 2021
Djeffal, Christian ; Horst, Antonia: Übersetzung und künstliche Intelligenz in der öffentlichen Verwaltung. In: *Berichte des NEGZ* 17 (2021), p. 1–40

Dodge et al. 2020
Dodge, Jesse ; Ilharco, Gabriel ; Schwartz, Roy ; Farhadi, Ali ; Hajishirzi, Hannaneh ; Smith, Noah: Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. In: *arXiv preprint arXiv:2002.06305* (2020)

Ebling et al. 2022
Ebling, Sarah ; Battisti, Alessia ; Kostrzewa, Marek ; Pfütze, Dominik ; Rios, Annette ; Säuberli, Andreas ; Spring, Nicolas: Automatic Text Simplification for German. In: *Frontiers in Communication* 7 (2022). https://www.frontiersin.org/articles/10.3389/fcomm.2022.706718. – ISSN 2297–900X

EU 2023
EU, Council: ChatGPT in the Public Sector - overhyped or overlooked? (2023)

Evans et al. 2014
Evans, Richard ; Orasan, Constantin ; Dornescu, Iustin: An evaluation of syntactic simplification rules for people with autism Association for Computational Linguistics, 2014

Feng 2008
Feng, Lijun: Text simplification: A survey. In: *The City University of New York, Tech. Rep* (2008)

Flesch 1948
Flesch, Rudolph: A new readability yardstick. In: *Journal of applied psychology* 32 (1948), Nr. 3, p. 221

Garbacea and Mei 2022
GARBACEA, Cristina ; MEI, Qiaozhu: Adapting Pre-trained Language Models to Low-Resource Text Simplification: The Path Matters. In: *Proceedings of The 1st Conference on Lifelong Learning Agents*, PMLR, nov 2022, 1103–1119. – ISSN: 2640-3498

Ghosh 2009
GHOSH, Siddhartha: Application of natural language processing (NLP) techniques in E–governance. In: *E-Government Development and Diffusion: Inhibitors and Facilitators of Digital Democracy.* IGI Global, 2009, p. 122–132

Gille et al. 2023
GILLE, Michael ; SCHOMACKER, Thorben ; HÜLLS, Jörg von d. ; TROPMANN-FRICK, Marina: Der Einsatz von Neural Language Models für eine barrierefreie Verwaltungskommunikation: Anforderungen an die automatisierte Vereinfachung rechtlicher Informationstexte, Gesellschaft für Informatik e.V., 2023. – ISBN 978–3–88579–735–7, 144–158

Gou et al. 2021
GOU, Jianping ; YU, Baosheng ; MAYBANK, Stephen J. ; TAO, Dacheng: Knowledge distillation: A survey. In: *International Journal of Computer Vision* 129 (2021), Nr. 6, p. 1789–1819

Graham et al. 2017
GRAHAM, Yvette ; BALDWIN, Timothy ; MOFFAT, Alistair ; ZOBEL, Justin: Can machine translation systems be evaluated by the crowd alone. In: *Natural Language Engineering* 23 (2017), jan, Nr. 1, 3–30. `http://dx.doi.org/10.1017/S1351324915000339`. – DOI 10.1017/S1351324915000339. – ISSN 1351–3249, 1469–8110. – Publisher: Cambridge University Press

Howard and Ruder 2018
HOWARD, Jeremy ; RUDER, Sebastian: Universal language model fine-tuning for text classification. In: *arXiv preprint arXiv:1801.06146* (2018)

Ji et al. 2023
JI, Ziwei ; LEE, Nayeon ; FRIESKE, Rita ; YU, Tiezheng ; SU, Dan ; XU, Yan ; ISHII, Etsuko ; BANG, Ye J. ; MADOTTO, Andrea ; FUNG, Pascale: Survey of hallucination in natural language generation. In: *ACM Computing Surveys* 55 (2023), Nr. 12, p. 1–38

Kajiwara and Komachi 2018
KAJIWARA, Tomoyuki ; KOMACHI, Mamoru: Text simplification without simplified corpora. In: *The Journal of Natural Language Processing* 25 (2018), p. 223–249

Kauchak 2013
KAUCHAK, David: Improving text simplification language modeling using unsimplified text data. In: *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2013, p. 1537–1546

Klaper et al. 2013

Klaper, David ; Ebling, Sarah ; Volk, Martin: Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In: Williams, Sandra (Hrsg.) ; Siddharthan, Advaith (Hrsg.) ; Nenkova, Ani (Hrsg.): *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Sofia, Bulgaria : Association for Computational Linguistics, aug 2013, 11–19

Kopp et al. 2023

Kopp, Tobias ; Rempel, Amelie ; Schmidt, Andreas P. ; SpieSS, Miriam: Towards machine translation into Easy Language in public administrations. Version: 2023. `http://dx.doi.org/10.57088/978-3-7329-9026-9_14`. In: *Emerging Fields in Easy Language and Accessible Communication Research*. Frank & Timme, Berlin, 2023. – DOI 10.57088/978–3–7329–9026–9$_1$4.$--ISBN978--3--7329--9026--9,371--406$

Kuang et al. 2023

Kuang, Weirui ; Qian, Bingchen ; Li, Zitao ; Chen, Daoyuan ; Gao, Dawei ; Pan, Xuchen ; Xie, Yuexiang ; Li, Yaliang ; Ding, Bolin ; Zhou, Jingren: *FederatedScope-LLM: A Comprehensive Package for Fine-tuning Large Language Models in Federated Learning*. 2023

Kühnhenrich and Michalik 2020

Kühnhenrich, Daniel ; Michalik, Susanne: Verwaltungssprache, schwere Sprache?–Ergebnisse zur Verständlichkeit von behördlichen Formularen und Schreiben aus der Lebenslagenbefragung 2019. In: *Verständliche Verwaltungskommunikation in Zeiten der Digitalisierung* Nomos Verlagsgesellschaft mbH & Co. KG, 2020, p. 47–62

Kumar et al. 2020

Kumar, Varun ; Choudhary, Ashutosh ; Cho, Eunah: Data Augmentation using Pre-trained Transformer Models. In: *CoRR* abs/2003.02245 (2020). `https://arxiv.org/abs/2003.02245`

Lewis et al. 2020

Lewis, Patrick ; Perez, Ethan ; Piktus, Aleksandra ; Petroni, Fabio ; Karpukhin, Vladimir ; Goyal, Naman ; Küttler, Heinrich ; Lewis, Mike ; Yih, Wen-tau ; Rocktäschel, Tim et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Advances in Neural Information Processing Systems* 33 (2020), p. 9459–9474

LHM a

LHM: *Digitale Teilhabe*. `https://ru.muenchen.de/2023/205/Mehr-finanzielle-Mittel-fuer-digitale-Teilhabe-beschlossen-109784`

LHM b

LHM: *Leichte Sprache*. `https://stadt.muenchen.de/leichte-sprache.html`

Liu et al. 2023

Liu, Pengfei ; Yuan, Weizhe ; Fu, Jinlan ; Jiang, Zhengbao ; Hayashi, Hiroaki ;

NEUBIG, Graham: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In: *ACM Computing Surveys* 55 (2023), Nr. 9, p. 1–35

Liu et al. 2019
LIU, Yinhan ; OTT, Myle ; GOYAL, Naman ; DU, Jingfei ; JOSHI, Mandar ; CHEN, Danqi ; LEVY, Omer ; LEWIS, Mike ; ZETTLEMOYER, Luke ; STOYANOV, Veselin: Roberta: A robustly optimized bert pretraining approach. In: *arXiv preprint arXiv:1907.11692* (2019)

Maaß 2020
MAASS, Christiane: *Easy Language – Plain Language – Easy Language Plus: Balancing Comprehensibility and Acceptability.* Frank & Timme, 2020. http://dx.doi.org/10.26530/20.500.12657/42089. http://dx.doi.org/10.26530/20.500.12657/42089. – ISBN 978–3–7329–9268–3. – Accepted: 2020-09-28T09:51:54Z

Martin et al. 2023
MARTIN, Tania J. ; ABREU SALAS, José I. ; MOREDA POZO, Paloma: A Review of Parallel Corpora for Automatic Text Simplification. Key Challenges Moving Forward. In: MÉTAIS, Elisabeth (Hrsg.) ; MEZIANE, Farid (Hrsg.) ; SUGUMARAN, Vijayan (Hrsg.) ; MANNING, Warren (Hrsg.) ; REIFF-MARGANIEC, Stephan (Hrsg.): *Natural Language Processing and Information Systems.* Cham : Springer Nature Switzerland, 2023. – ISBN 978–3–031–35320–8, p. 62–78

Mikolov et al. 2013a
MIKOLOV, Tomas ; CHEN, Kai ; CORRADO, Greg ; DEAN, Jeffrey: Efficient estimation of word representations in vector space. In: *arXiv preprint arXiv:1301.3781* (2013)

Mikolov et al. 2013b
MIKOLOV, Tomas ; SUTSKEVER, Ilya ; CHEN, Kai ; CORRADO, Greg S. ; DEAN, Jeff: Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems* 26 (2013)

Mohammed and Kora 2023
MOHAMMED, Ammar ; KORA, Rania: A comprehensive review on ensemble deep learning: Opportunities and challenges. In: *Journal of King Saud University-Computer and Information Sciences* 35 (2023), Nr. 2, p. 757–774

Moslem et al. 2023
MOSLEM, Yasmin ; HAQUE, Rejwanul ; WAY, Andy: *Fine-tuning Large Language Models for Adaptive Machine Translation.* http://dx.doi.org/10.48550/arXiv.2312.12740. Version: dec 2023. – arXiv:2312.12740 [cs]

Munich 2024
MUNICH, City o.: *Kenndaten zum Standort München.* https://www.munich-business.eu/standort-muenchen/standort-fakten.html. Version: 2024

NLS 2024
NLS: *Die Geschichte der Leichten Sprache.* `https://www.leichte-sprache.org/der-verein/die-geschichte/`. Version: 2024

NLS 2022
NLS, Netzwerk Leichte S.: *Die Regeln für Leichte Sprache.* `https://www.leichte-sprache.org/leichte-sprache/die-regeln/`. Version: 2022

OpenAI
OpenAI: *Models.* `https://platform.openai.com/docs/models/gpt-3-5-turbo`

Ormaechea et al. 2023
Ormaechea, Lucía ; Tsourakis, Nikos ; Schwab, Didier ; Bouillon, Pierrette ; Lecouteux, Benjamin: Simple, Simpler and Beyond: A Fine-Tuning BERT-Based Approach to Enhance Sentence Complexity Assessment for Text Simplification, 2023

Papineni et al. 2002
Papineni, Kishore ; Roukos, Salim ; Ward, Todd ; Zhu, Wei-Jing: Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, p. 311–318

Peng et al. 2023
Peng, Ruoling ; Liu, Kang ; Yang, Po ; Yuan, Zhipeng ; Li, Shunbao: *Embedding-based Retrieval with LLM for Effective Agriculture Information Extracting from Unstructured Data.* 2023

Pennington et al. 2014
Pennington, Jeffrey ; Socher, Richard ; Manning, Christopher D.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, p. 1532–1543

Peters et al. 2018
Peters, Matthew E. ; Neumann, Mark ; Iyyer, Mohit ; Gardner, Matt ; Clark, Christopher ; Lee, Kenton ; Zettlemoyer, Luke: ELMo: Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, p. 2227–2237

Petersen and Ostendorf 2007
Petersen, Sarah E. ; Ostendorf, Mari: Text simplification for language learners: a corpus analysis. In: *SLaTE*, 2007, p. 69–72

PolicyWeb 2024
PolicyWeb: *Government Policy Chatbot.* `https://huggingface.co/PolicyWeb/PolicyWeb`. Version: 2024

Radford et al. 2018
RADFORD, Alec ; NARASIMHAN, Karthik ; SALIMANS, Tim ; SUTSKEVER, Ilya: Improving language understanding by generative pre-training. (2018)

Rello et al. 2013
RELLO, Luz ; BAEZA-YATES, Ricardo ; SAGGION, Horacio: The impact of lexical simplification by verbal paraphrases for people with and without dyslexia. In: *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14* Springer, 2013, p. 501–512

Reynolds and McDonell 2021
REYNOLDS, Laria ; MCDONELL, Kyle: Prompt programming for large language models: Beyond the few-shot paradigm. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, p. 1–7

Ruder et al. 2019
RUDER, Sebastian ; PETERS, Matthew E. ; SWAYAMDIPTA, Swabha ; WOLF, Thomas ; VULIC, Ivan: A survey of transfer learning in NLP. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, p. 333–372

Schomacker 2023
SCHOMACKER, Thorben: Automatic German Easy Language (Leichte Sprache) Simplification: Data, Requirements and Approaches. (2023), 1–10. `https://dl.gi.de/handle/20.500.12116/42401`. – Publisher: Gesellschaft für Informatik e.V.

Schomacker et al. 2023
SCHOMACKER, Thorben ; GILLE, Michael ; HÜLLS, Jörg von d. ; TROPMANN-FRICK, Marina: *Data and Approaches for German Text simplification – towards an Accessibility-enhanced Communication.* 2023

Shin et al. 2020
SHIN, Taylor ; RAZEGHI, Yasaman ; LOGAN IV, Robert L. ; WALLACE, Eric ; SINGH, Sameer: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: *arXiv preprint arXiv:2010.15980* (2020)

Siddharthan and Katsos 2010
SIDDHARTHAN, Advaith ; KATSOS, Napoleon: Reformulating discourse connectives for non-expert readers. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, p. 1002–1010

Snover et al. 2006
SNOVER, Matthew ; DORR, Bonnie ; SCHWARTZ, Rich ; MICCIULLA, Linnea ; MAKHOUL, John: A Study of Translation Edit Rate with Targeted Human Annotation.

In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers.* Cambridge, Massachusetts, USA : Association for Machine Translation in the Americas, aug 2006, 223–231

Sulem et al. 2018a
SULEM, Elior ; ABEND, Omri ; RAPPOPORT, Ari: BLEU is Not Suitable for the Evaluation of Text Simplification. In: RILOFF, Ellen (Hrsg.) ; CHIANG, David (Hrsg.) ; HOCKENMAIER, Julia (Hrsg.) ; TSUJII, Jun'ichi (Hrsg.): *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium : Association for Computational Linguistics, oct 2018, 738–744

Sulem et al. 2018b
SULEM, Elior ; ABEND, Omri ; RAPPOPORT, Ari: *Semantic Structural Evaluation for Text Simplification.* `http://dx.doi.org/10.48550/arXiv.1810.05022`. Version: oct 2018. – arXiv:1810.05022 [cs]

Sulem et al. 2018c
SULEM, Elior ; ABEND, Omri ; RAPPOPORT, Ari: Simple and Effective Text Simplification Using Semantic and Neural Methods. In: GUREVYCH, Iryna (Hrsg.) ; MIYAO, Yusuke (Hrsg.): *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Melbourne, Australia : Association for Computational Linguistics, jul 2018, 162–173

Sun et al. 2021
SUN, Renliang ; JIN, Hanqi ; WAN, Xiaojun: *Document-Level Text Simplification: Dataset, Criteria and Baseline.* `http://dx.doi.org/10.48550/arXiv.2110.05071`. Version: oct 2021. – arXiv:2110.05071 [cs]

Tan et al. 2015
TAN, Liling ; DEHDARI, Jon ; GENABITH, Josef van: An awkward disparity between bleu/ribes scores and human judgements in machine translation. In: *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, 2015, p. 74–81

Toborek et al. 2023
TOBOREK, Vanessa ; BUSCH, Moritz ; BOSSERT, Malte ; BAUCKHAGE, Christian ; WELKE, Pascal: A New Aligned Simple German Corpus. In: ROGERS, Anna (Hrsg.) ; BOYD-GRABER, Jordan (Hrsg.) ; OKAZAKI, Naoaki (Hrsg.): *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Toronto, Canada : Association for Computational Linguistics, jul 2023, 11393–11412

Turney and Pantel 2010
TURNEY, Peter D. ; PANTEL, Patrick: From frequency to meaning: Vector space models of semantics. In: *Journal of artificial intelligence research* 37 (2010), p. 141–188

Vajjala and Meurers 2012
Vajjala, Sowmya ; Meurers, Detmar: On improving the accuracy of readability classification using insights from second language acquisition. In: *Proceedings of the seventh workshop on building educational applications using NLP*, 2012, p. 163–173

Watanabe et al. 2009
Watanabe, Willian M. ; Junior, Arnaldo C. ; Uzêda, Vinícius R. ; Fortes, Renata Pontin de M. ; Pardo, Thiago Alexandre S. ; Aluísio, Sandra M.: Facilita: reading assistance for low-literacy readers. In: *Proceedings of the 27th ACM international conference on Design of communication*, 2009, p. 29–36

White et al. 2023
White, Jules ; Fu, Quchen ; Hays, Sam ; Sandborn, Michael ; Olea, Carlos ; Gilbert, Henry ; Elnashar, Ashraf ; Spencer-Smith, Jesse ; Schmidt, Douglas C.: *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. `http://dx.doi.org/10.48550/arXiv.2302.11382`. Version: feb 2023. – arXiv:2302.11382 [cs]

Wu et al. 2022
Wu, Tongshuang ; Jiang, Ellen ; Donsbach, Aaron ; Gray, Jeff ; Molina, Alejandra ; Terry, Michael ; Cai, Carrie J.: Promptchainer: Chaining large language model prompts through visual programming. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, p. 1–10

Xu et al. 2016
Xu, Wei ; Napoles, Courtney ; Pavlick, Ellie ; Chen, Quanze ; Callison-Burch, Chris: Optimizing statistical machine translation for text simplification. In: *Transactions of the Association for Computational Linguistics* 4 (2016), p. 401–415

Yeung 2019
Yeung, Chin M.: *Effects of inserting domain vocabulary and fine-tuning BERT for German legal language*. `http://essay.utwente.nl/80128/`. Version: November 2019

Zhang et al. 2019
Zhang, Tianyi ; Kishore, Varsha ; Wu, Felix ; Weinberger, Kilian Q. ; Artzi, Yoav: Bertscore: Evaluating text generation with bert. In: *arXiv preprint arXiv:1904.09675* (2019)

Zhou et al. 2023
Zhou, Yongchao ; Muresanu, Andrei I. ; Han, Ziwen ; Paster, Keiran ; Pitis, Silviu ; Chan, Harris ; Ba, Jimmy: *Large Language Models Are Human-Level Prompt Engineers*. `http://dx.doi.org/10.48550/arXiv.2211.01910`. Version: mar 2023. – arXiv:2211.01910 [cs]

Zhu et al. 2010
Zhu, Zhemin ; Bernhard, Delphine ; Gurevych, Iryna: A monolingual tree-based translation model for sentence simplification. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, p. 1353–1361