Evaluation of Machine Translation Errors in German Plain Language Texts in the Domain of Health Information

Sarah Ahrens, Silvana Deilen, Sergio Hernández Garrido,
Ekaterina Lapshinova-Koltunski, Christiane Maaß
University of Hildesheim
ahrenss,deilen,hernandezs,lapshinovakoltun,maass@uni-hildesheim.de

Abstract

This paper presents the findings of a study evaluating intralingual machine translation of health information texts. In Germany, the National Action Plan Health Literacy addresses the poor performance of large population groups in this area by means of accessible communication measures, including the use of Plain Language. Machine translation can help here by making more Plain Language texts available. We thus evaluate the machine-generated Plain Language translations produced by two different state-of-the-art models and present the results of our error analysis. Our study reveals that in all categories, both models contain a high number of errors, however, the error distribution differs across the two models under analysis. The paper presents the elaborated evaluation scheme in detail, gives an overview of quantitative results and provides examples of most frequent errors occurring in both translation outputs.

1 Introduction

In the last few years, several studies have examined the quality of Machine Translation (MT) tools for intralingual translation (Anschütz et al., 2023; Deilen et al., 2023, 2024a). However, most studies on intralingual MT have analysed the quality of the output merely on the textual level, using automatic quality metrics. While some studies also mention and discuss the correctness of the generated content, so far none of them has specifically focused on a fine-grained error analysis. Understanding the types and frequency of errors and error categories in MT-generated texts is crucial, particularly in domains where clarity and accessibility are essential. Research has shown that large sections of the population in Germany have very low levels of health literacy (Schaeffer et al., 2017). The National Action Plan for Health Literacy (Schaeffer et al., 2018b) addresses this problem through accessibility measures. Plain Language is explicitly mentioned here as one of the measures to be systematically applied. However, as the need and demand for Plain Language texts are extremely high and continue to grow, it is becoming increasingly difficult to provide high-quality human translations for all the existing source texts. Consequently, the question arises whether the increasing demand for accessible Plain Language texts could be met by using MT systems and whether MT tools are capable of producing accurate, target-oriented, and functional intralingual translations. Such texts would have to meet the needs of the low literacy target groups and the institutional personnel in the health sector that is interacting with them.

This paper investigates the quality of machine translations of health information texts into Plain German, especially focusing on error typology. In our study, we analyse the machine translations from two different MT systems (SUMM AI and ChatGPT-4o) and classify the identified errors using an adapted version of the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014).

2 Related Work

2.1 Translation error analysis

Existing frameworks for error analysis in interlingual translation (both human and machine translation) are widely applied, such as the MQM framework proposed by Lommel et al. (2014). These frameworks typically classify errors based on linguistic categories and assess them according to scoring systems and severity levels. However, to our knowledge, no widely established error evaluation scheme currently exists for intralingual translation. Rodríguez Vázquez et al. (2022) use the MQM scheme to evaluate comprehensible text output, but their study design still focuses on automated interlingual translation of Easy Language texts (for a distinction between Easy Language and

Plain Language, see 2.2). Existing studies that evaluate automated intralingual translation known to us focus only on the surface features of translations, comparing their readability scores or syntactic complexity (Deilen et al., 2024a,b). Although the authors investigate the context correctness or incorrectness of the automatically produced translations, they report only on individual errors. No systematic description of error types is given.

2.2 Plain German

In Germany, both *Easy Language* (Leichte Sprache) and *Plain Language* (Einfache Sprache) belong to the field of intralingual translation (Hansen-Schirra et al., 2020; Maaß, 2020; Maaß, 2024b). Even though both language varieties aim to improve comprehensibility of texts, they are used in different communication settings and have different levels of complexity and therefore also different target groups. Easy Language is characterized by a maximally reduced complexity on all linguistic levels and is mainly intended for people with communication impairments. In contrast, Plain Language is a flexible, dynamic variety. The linguistic features of Plain Language are more complex and mainly intended for non-experts with average or slightly below average language or reading skills (Maaß, 2020; Maaß, 2024b). Also, Plain Language does not have the stigmatizing features that are often associated with Easy Language, which is one of the reasons why it is more acceptable to most users than Easy Language (Maaß, 2020). Plain Language has also gained increased recognition in recent years, with an increasing number of initiatives being implemented that propose Plain Language as a preferred means of accessible (specialized) communication. Currently, Plain Language is, for example, used to increase accessibility in the medical domain and overall in the health sector in Germany (Ahrens and Maaß, 2024).

2.3 Studies on (machine) translation and health communication

Kröger and Maaß (2024) address the problem of accessibility of health information. The authors analyse a small corpus of texts from the apotheken-umschau.de website[1]. They find the standard texts to lack comprehensibility, particularly focusing on readability scores and the use of terminology. They conclude that to comprehend existing standard texts, the reader needs to have prior knowledge of the respective topic of the text. Like Deilen et al. (2024a), Kröger and Maaß (2024) find moderate readability scores among the Apotheken Umschau standard texts that do not reach Plain Language readability levels. More and more texts containing health information are manually translated into Plain German. The use of machine translation, and specifically Large Language Models (LLMs) in the translation process of health communication texts may speed up the intralingual translation process. This would allow for a larger amount of health communication texts in Plain Language improving in this way access to health information (Schaeffer et al., 2018a), particularly to groups that score lower on health literacy surveys (like lower socio-economic status, a different first language or lower education levels, see for example (Schaeffer et al., 2021)). Improved access to health information may in turn increase health literacy (Rossmann and Hastall, 2019). However, studies have revealed that people have particular trouble appraising the quality, trustworthiness and commercial interest and how health information applies to their own situation (Schaeffer et al., 2021). The use of LLMs – if not properly monitored – may exacerbate this issue. Weidinger et al. (2022) discuss different risk areas concerning Language Models. One risk area is "misinformation harms": LLM-generated false information may threaten the reader's autonomy, increase their trust in false beliefs, amplify distrust in shared knowledge, and even cause bodily harm. These issues make unedited LLM-generated texts unsafe for users (Maaß, 2024a). Wilhelm et al. (2023) compare the health information output of four LLMs (like GPT-3.5-Turbo) considering completeness, correctness and harmfulness. The LLMs are prompted with the question "How to treat [disease]". In general, the outputs contain balanced and unbiased information and therefore seemed promising. Yet, they do not mention risks and potential harms of suggested treatments (completeness) and also give wrong or harmful information. The only reported system that does not generate harmful health information in this study is GPT-3.5-Turbo. Overall, the authors conclude that professional proof-reading of the outputs is necessary. However, a systematic description of common errors is still missing.

[1] *Apotheken Umschau* is a German health magazine, which publishes healthcare articles and health information.

3 Research Design

3.1 Data collection

We use parts of the dataset from Deilen et al. (2024a,b), which is based on the text corpus of the *Apotheken Umschau*. The corpus from Deilen et al. (2024a,b) consists of 30 source texts in standard German which were automatically translated into Plain Language using three different models of the SUMM AI[2] machine translation system. SUMM AI is a tool specifically developed for translating texts into Easy German and Plain German. This intralingual translation tool is based on an LLM, but also employs Easy and Plain Language parameters. Unlike ChatGPT, which is a generalized LLM not specifically developed for intralingual translation, the SUMM AI models are fine-tuned with rules and domain specific data, such as intralingual human gold-standard translations. However, unlike ChatGPT, the model cannot be prompted by the user. For Plain Language translation, SUMM AI started with a baseline model, which was a generically trained LLM. As explained by Deilen et al. (2024a) they then fine-tuned two other models (which are referred to as model 1 and model 2) using domain-specific source texts and human gold-standard translations. Model 1 and model 2 had different underlying LLMs but were fine-tuned with the same data. An evaluation of the output of the three models (baseline, model 1, model 2) by Deilen et al. (2024a) revealed that in terms of correctness, readability and text simplification scores (SARI), model 2 performed best. The SARI score (Xu et al., 2016), which is a quantitative measure to evaluate automatic text simplification systems, "compares system output against references and against the input sentence" (Xu et al., 2016).

As Deilen et al. (2024a) report that model 2 was the best of the three SUMM AI models, we compare it with ChatGPT. We aim to test whether a model that was specifically developed and fine-tuned for intralingual translation of health information outperforms a general LLM.

The intralingual translations with ChatGPT are generated using a two-step prompt. The prompt was developed based on findings from Deilen et al. (2023) who showed that assigning a role, setting a context and asking for background information seemed to improve the model's output. In the first step, ChatGPT was asked to define Plain Language in the field of accessible communication and inclusion:

"ChatGPT, wie wird im Bereich der Barrierefreien Kommunikation und der Inklusion „Einfache Sprache" definiert?" [ChatGPT, how is "Plain Language" defined in the field of Accessible Communication and inclusion?].

In a second step, ChatGPT was prompted to take on the role of a Plain Language expert:

"ChatGPT, du bist jetzt ein Experte für Einfache Sprache. Wir brauchen Unterstützung in der Übersetzung eines Textes der Gesundheitskommunikation in Einfache Sprache. In dem Text geht es um [hier Thema einfügen]. Die Übersetzung soll für Menschen mit Leseschwierigkeiten leicht verständlich sein. Übersetze bitte den folgenden Text:" [ChatGPT, you are now an expert in Plain Language. We need help translating a health communication text into Plain Language. The text is about [insert topic]. The translation should be easily comprehensible for people who have difficulty reading. Please translate the following text:]

Using the text corpus from Deilen et al. (2024a), we apply the interlingual translation error framework MQM (Multidimensional Quality Metrics, Lommel et al. (2014) to this intralingual corpus. MQM contains a variety of common translation error types. Depending on the project, certain error types are chosen and analysed in the source and target text. We intended to identify content-related errors (terminology and accuracy), errors related to common linguistic conventions like spelling, and the manner in which the audience is addressed, as well as how health messages are conveyed (audience appropriateness). The analysis did not evaluate the adherence to Plain Language guidelines. While SUMM AI adheres to a certain set of rules, ChatGPT does not. Additionally, as stated in section 2.2, Plain Language is flexible and dynamic. It is therefore of little interest to analyse the adherence to any specific guideline.

3.2 Data analysis

We analysed error types manually according to an adapted MQM scheme (Council, n.d.). We used four error types that we exemplify with the most meaningful sub-categories.

Terminology The use of a term does not fit to the field conventions, is incorrectly used in the target text or is not equivalent to the term in the source

[2] The company SUMM AI (`https://summ-ai.com/en/`) offers different licenses for freelancers, authorities and companies.

text. Also, multiple terms are used when just one term is appropriate.

Inconsistent terminology: Multiple terms are used to describe the same concept when just one term is needed or appropriate and consistency is desirable. For example, *birth control pill* and *contraceptive pill* used in the same text.

Wrong term: Use of term that it is not the term a domain expert would use or that gives rise to a conceptual mismatch. For instance *Hautprobleme* (skin problems) is used instead of *Hautveränderung* (skin change).

Accuracy Content in the target text does not match the propositions from the source text. This category has numerous subcategories that can be divided into further subgroups.

Mistranslation: Errors occurring when the target content does not accurately represent the source content. (1) Mistranslation of technical relations: In the target text, the relation between two elements is not presented in a technically appropriate way, even though the translation solution sometimes seems to be plausible at first glance. We provide examples of this case in Section 4.2. (2) Ambiguous target content: Ambiguity is introduced in the target where specificity is needed, e.g. the use of *Training* (training) instead of *Ausdauersport* (endurance sports). (3) Ambiguous source content: Ambiguous source content rendered in the target content inappropriately is illustrated in Section 4.2. (4) Hallucination: Output is decoupled from the source text or contains information that is not in the source text. We illustrate some cases in Section 4.2. Further subcategories include also incorrect conversion of numeric values, numbers, dates, or times.

Wrong addition: Necessary explanation is added, but does not represent the information from the source text or is wrong.

Missing addition: An explanation is needed, but was not added. For instance, various bacteria, such as Salmonella, Campylobacter, Listeria, were mentioned in the target text, but their differences were not explained.

Completeness: Relevant information from the source text is missing in the output. This may include incomplete details or also lists, e.g. for *Acetylsalicylsäure, Ibuprofen, Diclofenac und Paracetamol* (Acetylsalicylic acid, ibuprofen, diclofenac and paracetamol) in the source text, we find only *Aspirin, Ibuprofen und Paracetamol* (Aspirin, ibuprofen and paracetamol) in the target text.

Linguistic conventions *Grammar, punctuation, spelling:* grammatical errors, punctuation errors, and spelling errors. For instance, the phrase *manche psychische Probleme* (some psychological problems) has the wrong form of adjective (*psychische* instead of *psychischen*). Another example is the wrong spelling in *Reisenimpfungen* instead of *Reiseimpfungen* (Travel vaccinations).

Further cases: errors related to the textual cohesion and coherence, such as unclear references (see examples in Section 4.2). Apart from unclear references, texts may miss portions of text needed to connect the text into an understandable whole (connectives or lexical cohesion). This may result in a text which lacks a clear semantic relation between its parts, and in extreme cases it even does not make sense. For instance, in the question *Hat die Frau schon einmal schwere Erkrankungen in der Familie gehabt?* (Has the woman ever had serious illnesses in her family?), the semantic link between the woman, illnesses and family is missing. It should be rather *Hat die Frau schwere Erkrankungen gehabt* (Has the woman had any serious illnesses?) or *Hat die Familie der Frau schwere Erkrankungen gehabt?* (Has the woman's family had serious illnesses?).

Audience appropriateness Content in the target text is not valid, appropriate or acceptable for the target audiences. This may include (1) inaccurate advice – the target text contains advice that is not in the source text or that is not suitable for the target situation, or (2) stigmatising content – content can lead to stigmatization of end users. For instance the use of informal forms is a common stigmatising error[3].: *Ab 35 Jahren kannst du alle zwei Jahre eine Hautuntersuchung bei deiner Krankenkasse machen lassen.* (From the age of 35, you can have a skin examination every two years with your health insurance company). Another error is related to (3) the modality, if the same instruction is given in the source and target text, but with different intensity. For instance, the translation *Kinder sollten das nicht allein machen* (Children should not do this alone.) does not have the same intensity of advice as the source text *...das heiße Gebräu von Kindern fernhalten* (keep the hot brew away from children).

[3] In German, there is a distinction between informal and formal second-person pronouns: "du" is used for informal address (e.g., with children, friends or family), while "Sie" is the formal form, used in professional or public contexts. In official communications, such as those with health institutions, the formal form of address ("Sie") is expected.

Before the annotation of the whole corpus, the elaborated error analysis scheme was proved in a consensual validation process. All our human annotators are trained linguists with good knowledge of Plain German. Three annotators independently analysed a text generated with the MT system by SUMM AI, which we refer to as SUMM MT [4], and further two researchers did the same on the ChatGPT-generated text. The analysis of ChatGPT-generated texts was faster, since these translations are much shorter (see Section 4.1 for details). The results of the independent analyses were compared qualitatively and the following cases were discussed:

- One annotator categorised an error, but another did not;
- One annotator identified the same error as another annotator, but categorised it into a different category.

In this process of consensual validation, the error analysis scheme was modified and updated. The definition of labels were more clearly described. This made them more distinguishable from each other, and more fitting examples were chosen. Afterwards, the same annotators worked on the corpus, with, however, different texts assigned to them. For each text, one annotator analysed a text for errors. The same text was then cross-checked by another annotator. As a result, more errors were identified or the existing annotations were consistently checked for their categories.

In this paper, we present the quantitative comparison of error types across the two translation variants. We also qualitatively compare the outputs providing some examples of frequent errors.

In addition, we also calculated the SARI score for the ChatGPT output and compared it to the different models from SUMM AI.

4 Results

4.1 Overall performance

The first observation is the significant difference in text length: On average, the SUMM MT texts are three times as long as the ChatGPT texts (1225 words vs. 418 words). In the first step, we look into distributions of the four error types across the two translation outputs. Since the text length in the two translation variants varies strongly (see Figure 1), we normalise the distributions per total number of tokens in each translation variant. As seen from Figure 2, the outputs by ChatGPT contain more errors overall. At the same time, the distribution across the four error types defined above, varies in both outputs.

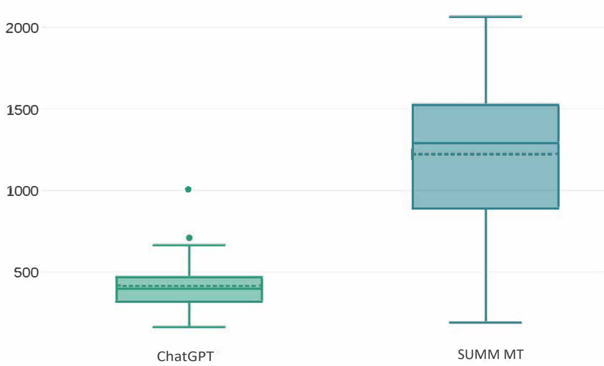

Figure 1: Text length in ChatGPT and SUMM MT outputs measured in tokens.

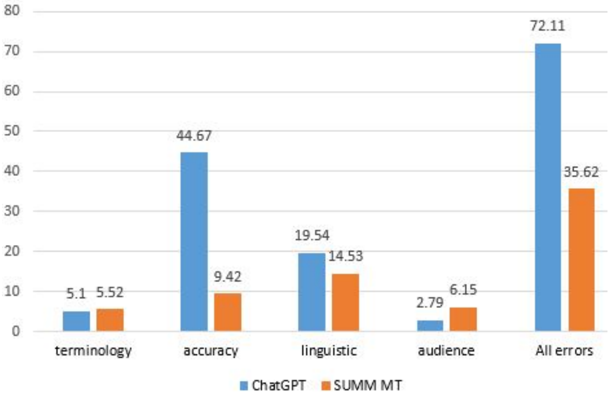

Figure 2: Annotated error distribution in ChatGPT and SUMM MT translations normalised against the total number of tokens (per 1000).

The outputs by ChatGPT contain far more accuracy errors than translations with SUMM MT (44.67 vs. 9.42). SUMM MT outperforms ChatGPT in the accuracy and linguistic convention categories. At the same time, ChatGPT outperforms SUMM MT in terms of audience appropriateness (2.79 vs. 6.15). In the category of terminology, both models show similar results (5.10 vs. 5.52).

To better understand which error categories prevail in ChatGPT and SUMM MT outputs separately, the distribution of error categories normalised per total number of errors within each translation output is illustrated in Figure 3.

It is interesting to see that the distribution of categories differs in the two translation variants.

[4] The term SUMM MT refers to the above-mentioned best-performing model 2 by SUMM AI.

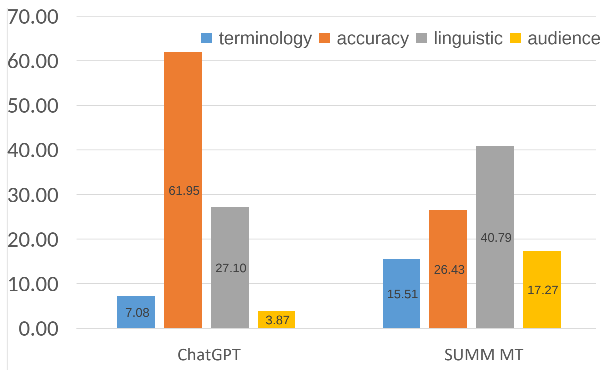

Figure 3: Error category distribution within each translation output normalised against the total number of errors in %.

While translations with ChatGPT definitely suffer from a high number of accuracy errors, translations with SUMM MT have a more even distribution with none of the category achieving even 50%.

4.2 Fine-grained differences

In the following, we analyse some specific examples of errors in the two translation outputs.

Terminology We start with the errors in terminology use. The first error type is ’Inconsistent terminology’ illustrated in example (1). The source text contains two variants of the term *Antibabypille* (birth control pill) and *Pille* (pill). So do both automatic translations.

(1) a Source: *Die Antibabypille schützt gut vor einer ungewollten Schwangerschaft und ist einfach anzuwenden. Wir informieren über die Wirkung, Vorteile und Nachteile der Pille.* (The contraceptive pill is a reliable method of contraception and easy to use. We provide information about the effect, advantages and disadvantages of the pill.)

b ChatGPT: *Die Antibabypille ist ein Medikament, das gut vor Schwangerschaft schützt und einfach angewendet werden kann. Wir erklären, was die Pille bewirkt und ihre Vor- und Nachteile.* (The contraceptive pill is a medication that provides good protection against pregnancy and is easy to use. We explain what the pill does and its advantages and disadvantages.)

c SUMM MT: *Die Antibabypille verhindert eine Schwangerschaft. Sie ist leicht zu benutzen. Wir erklären, wie die Pille wirkt.* (The contraceptive pill prevents pregnancy. It is easy to use. We explain how the pill works.)

Accuracy The next case illustrated in example (2) belongs to the category of accuracy and demonstrates mistranslation of technical relation that occurs in the outputs produced with ChatGPT.

(2) a Source: *Der Zeitraum, in dem wechseljahrsbedingte Hitzewallungen oder Nachtschweiß häufig vorkommen, dauert einer Studie zufolge durchschnittlich 7,4 Jahre* (According to a study, the period in which menopause-related hot flashes or night sweats frequently occur lasts an average of 7.4 years.)

b ChatGPT: *Im Durchschnitt dauern Hitzewallungen etwa 7,4 Jahre.* (On average, hot flashes last about 7.4 years.)

c SUMM MT: *Die Zeit, in der Hitzewallungen und Nacht-Schweiß vorkommen, dauert durchschnittlich 7,4 Jahre.* (The period in which hot flushes and night sweats occur lasts an average of 7.4 years.).

Here, the mistranslation in the ChatGPT output also includes a change in meaning. While the source text states that the phase in which these symptoms commonly occur lasts on average 7.4 years, the translation by ChatGPT implies that hot flashes themselves last that long, which is misleading and incorrect. Although the SUMM MT translation does not contain this kind of error, we observe here a syntactic problem instead: due to the relative clauses inserted, the sentence is too complex.

Another example of a mistranslation of a technical relation is the use of a wrong verb, e.g. *Warndreieck aufsetzen* (put on a warning triangle) instead of *Warndreieck aufstellen* (put up a warning triangle). This error subtype is more frequent in the ChatGPT translations than in the SUMM MT ones (65.4 vs. 27.5, normalised per 10,000 against the total number of words in translations).

In example (3), we deal with ambiguous source content, another subcategory of the accuracy errors.

(3) a Source: *Setzen Sie Minzöl bei Spannungskopfschmerzen ein, massieren Sie das verdünnte Öl direkt auf der schmerzenden Stelle ein oder träufeln es vorher auf ein Tuch, um das dann zu verwenden.* (Use mint oil for tension headaches. Massage the diluted oil

directly onto the painful area or drip it onto a cloth first and use that.)

b ChatGPT: *Gegen Spannungskopfschmerzen massieren Sie das verdünnte Öl direkt auf die schmerzende Stelle oder benutzen ein Tuch.* (Against tension headaches, massage the diluted oil directly onto the painful area or use a cloth.)

c SUMM MT: *Massieren Sie das verdünnte Öl auf die schmerzende Stelle. Oder geben Sie das Öl auf ein Tuch und legen Sie das Tuch auf die schmerzende Stelle.* (Massage the diluted oil onto the painful area. Or apply the oil to a cloth and place the cloth on the painful area.)

There is information missing in the ChatGPT translation illustrated in (3-b): It remains unclear what exactly to do with the cloth. The translation with SUMM MT in (3-c) instead provides the necessary details.

Another case of accuracy errors includes hallucinations. Example (4) is is a translation with ChatGPT that confuses the source concepts *gesunder Lebensstil* (healthy lifestyle) and *schlaffördernde Verhaltensweisen* (sleep-improving behaviours). It hallucinates the sleep-improving acts of going to bed at regular times and not having any heavy meals before bed time as being examples of a healthy lifestyle. All the while not mentioning the sleep-improving behaviours, leaving the reader with faulty action-orientation.

(4) a Source: *Selbsthilfemaßnahmen wie gesunder Lebensstil und den Schlaf fördernde Verhaltensweisen* (Self-help measures like a healthy lifestyle and sleep-improving behaviours)

b ChatGPT: *Gesunder Lebensstil: Regelmäßig schlafen gehen, keine schweren Mahlzeiten vor dem Schlaf.* (Healthy lifestyle: Go to bed regularly, no heavy meals before sleep).

In general, ChatGPT contains much more accuracy errors of the hallucination subtype than SUMM MT: 33.5 vs. 11.7 (frequencies per 10,000 normalised against the total number of words).

Linguistic conventions This category contains many spelling and punctuation errors, as well as grammar errors such as missing definite articles. Cohesion and coherence, as well as cases of unclear reference are also common. An example of unclear reference is illustrated in examples (5) and (6). The word *Gerät* (device) does not have any antecedents in the preceding sentence(s). It is a result of translating the subheading "*Mit dem Blutdruckmessgerät verbunden*" (Connected to the blood pressure monitor), and is, therefore, wrong.

(5) a Source: *Mit Medikamenten und einem gesunden Lebensstil gelingt es den meisten, zu hohe Werte langfristig zu senken. Digitale Helfer können die Therapie unterstützen und den Umgang mit der Krankheit erleichtern. Mit dem Blutdruckmessgerät verbunden.* (Most people may reduce high values with medication and a healthy lifestyle over the long term. Digital aids can support therapy and make it easier to deal with the disease. Connected to the blood pressure monitor.

b SUMM MT: *Mit Medikamenten und einem gesunden Lebens-Stil können Sie den Blut-Druck senken. Digitale Helfer können dabei unterstützen. Das Gerät ist mit dem Blut-Druck-Messer verbunden.* (You can lower your blood pressure with medication and a healthy lifestyle. Digital helpers can provide support. The device is connected to the blood pressure meter.)

(6) a Source: *Telemedizin ersetzt nicht den persönlichen Kontakt. Eine gewisse Offenheit für den Einsatz von Computer und Smartphone sowie eine sichere und stabile Internetverbindung vorausgesetzt, ist die Telemedizin aber eine sinnvolle Ergänzung.* (Telemedicine is no substitute for face-to-face contact. However, telemedicine is a useful addition, provided you are open to the use of computers and smartphones and have a secure and stable internet connection.)

b SUMM MT: *Aber: Telemedizin ersetzt nicht den Arzt-Besuch. Sie brauchen dafür:* (But: Telemedicine does not replace a visit to the doctor. You need for this:)

Example (6) contains the referring expression *dafür* (for this), a pronominal adverb in German which may refer to a prepositional phrase or to a clause or sentence. This reference in the context in (6-b) is ambiguous, since it may refer to both *Telemedizin* (telemedicine) and also to *Arzt-Besuch* (visit to the doctor).

Audience appropriateness This type of error was mostly related to the use of advice that was not contained in the source text, and that was not appropriate for the target readership at the same time. Another frequent type of errors were stigmatising elements.

Example (7-a) contains the advice *suchen Sie im Internet nach Informationen* (search the web for more information) which is not in the source text and is not appropriate for the target audience. Example (7-b) demonstrates stigmatising content, as the text contains the 2nd person plural *ihr*, which is the informal address in German where the formal address would be required. Moreover, the expression *passt auf!* (Watch out!) is also informal and a way patronising or even infantilising, which is definitely not approrpiate in the current context.

(7) a SUMM MT: *Sie wollen mehr wissen? Dann fragen Sie Ihren Arzt. Oder suchen Sie im Internet nach Informationen.* (Do you want to know more? Then ask your doctor. Or search the web for information).

b ChatGPT: *Wenn ihr keine Inhalationsgeräte habt und eine Schüssel benutzt, passt auf!* (If you do not have an inhaler and use a bowl, watch out!)

4.3 Automatic evaluation measures

In addition to the error analyes, we also calculated the SARI score, which is a text simplification metric. Figure 4 shows the SARI score of the machine translation output from SUMM MT as well as the SARI score of the ChatGPT texts, with higher SARI values indicating better machine translated outputs. Comparing ChatGPT with SUMM AI reveals that ChatGPT performs worse than SUMM MT. This result is in line with the results by Ahrens et al. (2025) who also found that in terms of readability and syntactic complexity, the models by SUMM AI outperformed ChatGPT. However, it has to be kept in mind that the SARI score is only calculated on the text surface and does not take into account how well a text adheres, for example, to official guidelines and Plain Language rules.

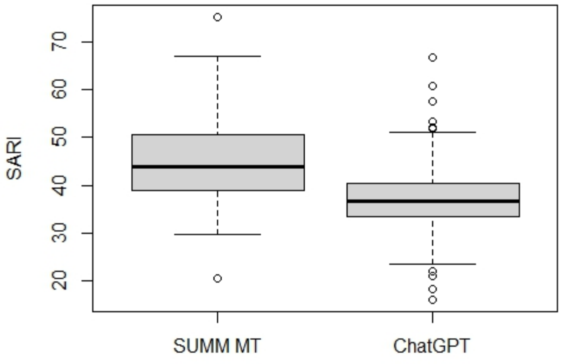

Figure 4: SARI score of the machine translation output from SUMM MT and ChatGPT.

5 Discussion and Future Work

In our study, we compared the performance of two machine translation systems, ChatGPT and SUMM MT, on a dataset of German health information texts. We carried out a fine-grained error analysis of the translations, using a scheme based on the MQM framework that was applied to intralingual translation. We analyzed the errors in the MT outputs and categorized them, following the MQM scheme, into four types: terminology, accuracy, linguistic conventions, and audience appropriateness. The results show that ChatGPT contains more accuracy errors than SUMM MT and that SUMM MT also outperforms ChatGPT in the linguistic convention category. ChatGPT, on the other hand, outperforms SUMM MT in terms of audience appropriateness. The errors in the translations include mistranslations, ambiguous content, and hallucinations. The results suggest that SUMM MT has a more even distribution of error categories than ChatGPT, while the latter suffers from an exceedingly high number of accuracy errors.

The results can be explained by differences between the models. The SUMM AI model outperforms ChatGPT in almost all analysed criteria and error categories. This suggests that fine-tuning a model both with task-specific data (i.e., the task of intralingual translation into Plain Language), and with domain-specific data (health information texts from *Apotheken Umschau*), improves MT into Plain German. Therefore, for our study we can conclude that fine-tuned models that are trained specifically for intralingual translation tasks outperform the general model that uses prompting.

The analysis of error categories and specific errors has revealed a high number of errors with respect to accuracy and audience appropriateness. Health information is a safety-critical domain, so these errors can be regarded as more severe than, for example, spelling errors. Our study thus confirms findings by existing studies (Deilen et al., 2024b; Maaß, 2024a; Wilhelm et al., 2023) that the automatically produced outputs cannot be used safely by end users. Although machine translation does facilitate intralingual translation, without human intervention in form of post-editing, it is not suitable for achieving the goal of increasing health literacy.

In our future studies, we also aim to assess error severity, following MQM settings and recommendations. Besides that, we plan to investigate whether the high error rate, especially in terms of accuracy, might be reduced by using best-practice prompting strategies that specifically aim to mitigate accuracy errors.

In addition, we plan to analyse the cases of disagreement between the annotators which may indicate interesting ambiguous cases.

Limitations

The error analysis was conducted with a larger team of annotators with one team annotating SUMM MT and separate team annotating the ChatGPT corpus. It is possible that error categories have been assessed differently between the sub-corpora. An annotation process with only two annotators who both annotate both corpora would have been more reliable. Another limitation is that we have not yet assessed error severity. However, doing so is important for the overall evaluation, as the impact of different error types can vary significantly. For instance, content-related errors (e.g., mistranslations or omissions) can alter the meaning of a sentence and pose a much greater risk than more superficial issues such as spelling or punctuation mistakes. Future work will integrate a severity scale to differentiate between critical and minor errors, which will allow for a more detailed evaluation of translation quality. Also, this study is limited by the amount of data. Our evaluation was only conducted on a selected set of translations from one domain, which limits the generalization of our findings. Therefore, in future studies, we will also extend the analysis to texts from other domains. A study such as this contains error-prone texts in a field where misinformation is can be dangerous. Therefore, user testing cannot ethically be performed.

References

Sarah Ahrens, Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Hernández Garrido, and Christiane Maaß. 2025. Evaluation of translations into plain german produced by humans and mt systems including chatgpt. *SKASE Journal of Translation and Interpretation*. In press.

Sarah Ahrens and Christiane Maaß. 2024. Der Einfluss von Vorwissen und sprachlichen Texteigenschaften auf die Anwendung des Fragebogens in einer gynäkologischen Patientenaufklärung: eine leitfadengestützte Interviewstudie. *Prävention und Gesundheitsförderung*, 19:504–511.

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for german text simplification: Overcoming parallel data scarcity through style-specific pre-training. *arXiv preprint arXiv:2305.12908*.

The MQM Council. n.d. The MQM error typology. Last accessed on Feb 18th 2025.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using chatgpt as a cat tool in easy language translation. *arXiv preprint arXiv:2308.11563*.

Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Garrido, Julian Hörner, Christiane Maaß, Vanessa Theel, and Sophie Ziemer. 2024a. Evaluation of intralingual machine translation for health communication. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 469–479.

Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Hernández Garrido, Christiane Maaß, Julian Hörner, Vanessa Theel, and Sophie Ziemer. 2024b. Towards AI-supported Health Communication in Plain Language: Evaluating Intralingual Machine Translation of Medical Texts. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 44–53.

Silvia Hansen-Schirra, Walter Bisang, Arne Nagels, Silke Gutermuth, Julia Fuchs, Liv Borghardt, Silvana Deilen, Anne-Kathrin Gros, Laura Schiffl, and Johanna Sommer. 2020. Intralingual translation into Easy language–or how to reduce cognitive processing costs. *Easy Language Research: Text and User Perspectives. Berlin: Frank & Timme*, pages 197–225.

Janina Kröger and Christiane Maaß. 2024. Mangelnde verständlichkeit durch fachsprache in gesundheitsinformationen zu chronischen erkrankungen – eine qualitative korpusanalyse. *Prävention und Gesundheitsförderung*, 19:497–503.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:0455–463.

Christiane Maaß. 2020. *Easy language–Plain language–Easy language plus: Balancing comprehensibility and acceptability*. Frank & Timme.

Christiane Maaß. 2024a. Hi ChatGPT, translate this text into Easy Language. Is the new Easy Language translator a machine? In *Proceedings of the XLIV International VAKKI Symposium*.

Christiane Maaß. 2024b. Intralingual Translation in Easy Language and Plain Language. In *Routledge Handbook of Intralingual Translation*, pages 234–251. Routledge.

Silvia Rodríguez Vázquez, Abigail Kaplan, Pierrette Bouillon, Cornelia Griebel, and Razieh Azari. 2022. La traduction automatique des textes faciles à lire et à comprendre (falc): une étude comparative. *Meta*, 67(1):18–49.

Constanze Rossmann and Matthias R Hastall. 2019. *Handbuch der Gesundheitskommunikation*. Springer.

Doris Schaeffer, Eva-Maria Berens, Svea Gille, Lennert Griese, Klinger Julia, Steffen de Sombre, Dominique Vogt, and Klaus Hurrelmann. 2021. *Gesundheitskompetenz der Bevölkerung in Deutschland vor und während der Corona Pandemie: Ergebnisse des HLS-GER 2*. Universität Bielefeld, Interdisziplinäres Zentrum für Gesundheitskompetenzforschung.

Doris Schaeffer, Klaus Hurrelmann, Ullrich Bauer, and Kai Kolpatzik. 2018a. *National Action Plan Health Literacy. Promoting health literacy in Germany*. Kompart, Berlin.

Doris Schaeffer, Klaus Hurrelmann, Ulrich Bauer, Kai Kolpatzik, and Attila Altiner. 2018b. *Nationaler Aktionsplan Gesundheitskompetenz: die Gesundheitskompetenz in Deutschland stärken: Informationen bewerten, finden, anwenden, verstehen*. Hertie School of Governance.

Doris Schaeffer, Dominique Vogt, Eva-Maria Berens, and Klaus Hurrelmann. 2017. Gesundheitskompetenz der bevölkerung in deutschland: ergebnisbericht.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Theresa Isabelle Wilhelm, Jonas Roos, and Robert Kaczmarczyk. 2023. Large language models for therapy recommendations across 3 clinical specialties: Comparative study. *J Med Internet Res*, 25:e49324.

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