information

MDPI

Article

Evaluating Intralingual Machine Translation Quality: Application of an Adapted MQM Scheme to German Plain Language

Silvana Deilen [1,*], Sergio Hernández Garrido [2], Ekaterina Lapshinova-Koltunski [2], Chris Maaß [2] and Annie Werner [2]

[1] Institute for Translation Studies and Interpreting, Heidelberg University, 69117 Heidelberg, Germany
[2] Institute for Translation Studies and Specialised Communication, University of Hildesheim, 31141 Hildesheim, Germany; hernandezs@uni-hildesheim.de (S.H.G.); lapshinovakoltun@uni-hildesheim.de (E.L.-K.); maass@uni-hildesheim.de (C.M.); werner001@uni-hildesheim.de (A.W.)
* Correspondence: silvana.deilen@iued.uni-heidelberg.de

Abstract

This paper presents the results of a study in which we conducted a fine-grained error analysis for intralingual machine translations into Plain Language. As there are no established error schemes for intralingual translations, we adapted the MQM scheme to fit the purposes of intralingual translation and expanded the scheme by error categories that are only relevant to intralingual translation. Our study has revealed that substantial differences exist between general-purpose and domain-specific models, with fine-tuned systems achieving notably higher accuracy and fewer severe errors across most categories. We found that across all four models, most errors occurred in the "Accuracy" category, closely followed by errors in the "Linguistic conventions" category and that all evaluated models produced persistent issues, particularly in terms of accuracy, linguistic conventions, and alignment with the target audience. In addition, we identified subcategories from the MQM scheme that are primarily relevant to interlingual translation, such as "Textual conventions". Furthermore, we found that manual error annotation is resource-intensive and subjective, highlighting the urgent need for the development of automatic or semi-automatic error annotation tools. We also discuss difficulties that arose in the annotation process and show how methodological limitations might be overcome in future studies. Our findings provide practical directions for improving both machine translation technology and quality assurance frameworks for intralingual translation into Plain Language.

Keywords: translation quality evaluation; large language models; intralingual translation; health communication; error analyses

Check for updates

Academic Editor: Katsuhide Fujita

Received: 6 November 2025
Revised: 17 December 2025
Accepted: 2 January 2026
Published: 6 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the rise of Artificial Intelligence (AI) used for machine translation purposes, e.g., in form of interactive chatbots or other tools [1] and the rapid developments in the field since the public release of ChatGPT at the end of 2022 [2], the topic of translation quality evaluation is more important than ever. However, the majority of studies that deal with translation quality evaluation have focused on automatic metrics and computational approaches. These include BLEU [3], METEOR [4], or COMET [5], among others, for

interlingual translation. In the area of intralingual translation, which is the focus of this study, comprehensibility indexes (e.g., Flesch Reading Ease, i.e., FRE [6]) and automated metrics such as SARI [7] are used as common approaches derived from automatic text simplification. While these metrics provide fast and scalable ways to assess machine-translated output, they can only be calculated if a reference translation is available, and in addition to that, they often do not align with human judgment. Also, automatic evaluation is said to be "less comprehensive than manual evaluation and does not readily indicate the type of problems that the translated text contains" [8] (p. 25). In addition to that, automatic error annotation using AI often fails to capture a large number of crucial errors in intralingual translation. Therefore, in our study, we decided to rely on manual annotation and focus on a fine-grained human evaluation to ensure a high-quality and comprehensive analysis. This was also because annotation with LLMs (or other means) works only if there is a gold standard available or if the method has been approved already. However, we test and use MQM for the first time for intralingual translation, which is why we decided to rely on manual annotation and error classification. Our focus is on intralingual translation, i.e., translating a text from standard German into German Plain Language. Although not unrelated, our view differs from traditional studies on automatic text simplification, since we see a change from standard German into Plain Language not as merely simplifying a text within one language on a certain linguistic level. Instead, we view this process as a transfer and adaptation for a different target group, a process which is similar to those happening in interlingual translation when a text is translated from one language and culture into another.

In our study, we conduct a detailed error classification of intralingual translations produced with different machine translation (MT) systems. While for interlingual translation, there is a wide range of established error schemes (see, e.g., multidimensional quality metrics—MQM, [9]), such an established error scheme is still missing in the field of intralingual translation. This is mainly because in intralingual translation, errors do not only refer to aspects like mistranslations or terminology (see, e.g., [10]), but they also include issues such as readability, perceptibility, acceptability or rule violations [11,12]. These specific issues are also addressed by [13], who analyses the appropriateness of Easy Language texts produced by Large Language Models (LLMs) for the respective target audiences, as well as the usability of the texts in web form. Aside from research on the ethics of using LLMs to produce texts in Easy Language [14], there is currently very little research available on other possible translation errors in the field of intralingual translation.

The aim of our study is to address this research gap by adapting the established MQM error scheme to intralingual translation. More specifically, we aim to assess the extent to which the MQM scheme can be applied and has to be adapted to fit the purposes of intralingual translation. We present a systematic description of the main error types and expand the scheme by error categories that are additionally relevant in intralingual translation. These new categories, relevant in intralingual translation, are discussed and explained by using data from a thorough and detailed error analysis of different systems used for intralingual machine translation. We also aim to compare error distribution between models that were specifically fine-tuned and trained for intralingual translation, those fine-tuned for a specific domain under analysis, as well as a general LLM. In this way, we also aim to investigate whether fine-tuning a model with domain-specific data and human gold standard translations helps to avoid or reduce errors in intralingual translation.

Finally, we also present challenges, such as non-classifiable errors, and discuss approaches on how to deal with them.

2. Related Work

2.1. Plain Language

Translating a text from standard German into Plain Language belongs to the field of intralingual translation, which, according to [15] is "an interpretation of verbal signs by means of other signs of the same language". Plain Language is a complexity-reduced variety of German that is used to make expert information accessible to non-expert readers [16]. Unlike Easy Language, which is a rule-based variety and characterized by a maximally reduced complexity on all linguistic levels, Plain Language is a dynamic variety and does not have a fixed set of rules. Instead, the complexity of a Plain Language text is usually adapted to the needs of the intended target audience [16]. While Easy Language is primarily intended for people with communication impairments and disabilities, Plain Language is aimed at people with limited prior knowledge in a specific field (non-experts and lay people). According to [17], the language level required to understand a Plain Language text is set at B2. At the time of writing this paper, two domain-independent standards from the German Institute for Standardisation (=Deutsches Institut für Normung, DIN) are available for the German language: DIN ISO 24495-1 "Plain Language—Part 1: Governing principles and guidelines" [18] and DIN 8581-1 "Plain Language—Application for the German language—Part 1: Language-specific provisions" [19]. The former represents the German translation of the international standard ISO 24495-1:2023 on Plain Language [20], while the latter provides language-specific recommendations for Plain Language in German. Both standards are non-binding and serve as guidelines for best practice.

In recent years, we have seen a growing need for Plain Language translations in various domains, not least because of the European Accessibility Act (EAA), an EU directive set up to ensure greater consistency in the accessibility of products and services within the single market. The EAA addresses obstacles arising from differing national accessibility standards by establishing unified requirements, making it easier for businesses to offer accessible solutions across borders and enabling consumers throughout Europe to benefit from improved access. It is aimed at removing regulatory fragmentation and supporting inclusive design; measures have to be implemented by June 2025. However, this rising demand for Plain Language translations cannot be met by human translators alone, opening opportunities for the application of AI in intralingual translation. Even though there are some tools on the market that offer machine translations into Plain Language, only a few studies have evaluated the quality of these AI-generated translations [21]. This lack of research can be attributed to several reasons: Firstly, intralingual translation into Plain Language is still an emerging field, and secondly, unlike in the field of interlingual translation, many of the available intralingual translation tools are commercial tools, which limits access for research purposes.

2.2. Plain Language in Automated Tasks

For automated tasks, such as automatic text simplification or automated text summarisation for non-expert audiences (Plain Language Summarisation or PLS), it is common to use either automatic metrics, such as the above mentioned SARI and LENS [22], or comprehensibility indexes including the above mentioned FLE and further metrics such as Dale Chall Readability Score (DCRS, [23]), Flesh-Kincaid Grade Level (FKGL, [6]), Spache [24], Coleman Liau Index (CLI, [25]) and Gunning Fog Index (GFI, [26]). However, they primarily address shallow lexical features, such as the number of syllables in a word to measure readability and do not always correspond to the targeted results. For instance, as reported by [27], texts containing numerous abbreviations would be ranked as more readable, which, however, does not mean that they would be more comprehensible for the targeted group of readers in Plain Language. Besides that, the authors also report that those metrics do

not correlate with human judgements. Automatic text simplification evaluation metrics like SARI and LENS also lack in informativeness on more specific problems. These issues partly arise because some text simplification systems are based on general assumptions about what makes a text simple [28].

Studies on PLS also report such evaluation criteria as (1) informativeness (the extent to which the Plain Language summary covers essential information from the source text), (2) simplification (the degree to which information is conveyed in a form that non-expert audiences can readily understand), (3) coherence (logical arrangement of a Plain Language summary) and (4) faithfulness (how well the summary aligns factually with the source text), see [29] for more details. Recent works also use LLMs for prompt-based evaluation. For instance, ref. [30] use the above mentioned criteria in a prompt-based scenario with LLMs, showing that LLMs cannot assess all the categories equally. Ref. [27] also utilise LLMs, however, without looking into specific evaluation criteria. Ref. [31] created a corpus of parallel, professionally written and manually aligned simplifications in plain German. They show that using their corpus to train a transformer-based text simplification model can achieve promising simplification results. However, they also highlight the importance of verifying the results by manual evaluation.

Although research on simplification methods is extensive, the systematic evaluation of the generated outputs has received comparatively little attention [32]. Ref. [32] discuss some challenges for evaluation in the field of text simplification. One of these challenges is that simplification needs often depend on the aimed target groups and may differ significantly, which may lead to heterogeneous text demands. For example, a young person with an intellectual disability needs explanations of other terms in a text than an older person with early-stage dementia. In addition, it has to be kept in mind that the evaluation of the output is subjective and depends on the individual background of the annotator and their opinion on what is simple and what is not. This subjectivity often leads to low reproducibility and low inter-annotator agreement [33]. Besides that, it is hard to define target groups as there are no simplified-language speakers [34] (p. 39), as this is the case of interlingual translation. The automated approaches approximate the target groups by age or education level, as well as language proficiency, which does not always correspond to the real users of Plain Language. Moreover, the problem is aggravated by the fact that both the text recommendations and the evaluation guidelines are quite vague.

2.3. Error Analysis for Intralingual Translation

In our work, we address the transfer from standard texts into Plain Language as a case of intralingual translation. The main objective of intralingual translation remains similar to text simplification [35], i.e., to reduce linguistic complexity and increase comprehensibility while preserving the intended meaning of the source text. This is achieved by applying simplification strategies such as shortening and splitting sentences, choosing more frequent and familiar vocabulary, avoiding complex lexical and syntactic structures, and restructuring and selecting information.

Methodologically, the field of intralingual translation faces research gaps in evaluation, both in human and machine translation, as there are no dedicated evaluation metrics or error typologies for intralingual translation. Existing interlingual frameworks, such as MQM (Multidimensional Quality Metrics; ref. [9] are not directly applicable, as intralingual translation requires broader error definitions, including readability, perceptibility, and rule violations [16]. Adapting these typologies requires the integration of new, reader-oriented categories and strategies.

In the above mentioned work on evaluation for text simplification [32], the authors also investigate to what extent metrics in the evaluation of interlingual machine translations

can be applied to text simplification. However, in this context, they also point out that these metrics often do not correlate with simplicity and are therefore only of limited use. They conclude that for a thorough, detailed evaluation, different metrics are to be combined.

As human expert evaluation of machine-generated translations is very time-consuming and expensive, ref. [36] propose an oblivious, unsupervised method for estimating machine translation quality that does not require a large bilingual corpus of source-target pairs. Testing this method for interlingual translations, they showed that the system was comparable to the performance of non-oblivious, supervised systems and could therefore be equally used to estimate translation quality, especially for low-resource language pairs that lack a large bilingual corpora for training. Even though their method has not been tested for intralingual translations, their results are very promising as they show that quality estimation might even be feasible without an aligned corpus of source and target texts.

Based on [13], one can argue that the usage of AI-powered language tools that can be used, e.g., for correction of grammatical errors or AI-generated suggestions to improve upon the writing, could also be considered to fall into the category of intralingual AI translation tools. To the best of our knowledge, the only studies known to us that conducted error analyses for intralingual translation are those by [13] and by [37,38]. While [13] focuses on a broader discussion of the usefulness of LLMs for translation into Easy Language in a university context, ref. [37,38] focus on the use of LLMs for translation into Easy Language in the healthcare context.

Ref. [37] addressed Easy Language translations generated by GPT-3.5. Later on, they evaluated Plain Language translations generated by different versions of the commercial machine translation tool SUMM AI. However, they only reported whether the content was correct or incorrect, i.e., as soon as they encountered one content-related error, the text was already considered incorrect. This binary approach (correct vs. incorrect) was chosen because they aimed to determine whether the intralingual machine translation tools were suitable for end users and the mere existence of content-related mistakes in the translation allowed them to conclude that the tools could not safely be used by end-users. Still, for an extensive evaluation of the translation quality, this binary approach is not sufficient, as the amount, categorization and severity of identified mistakes also have a significant impact on the overall translation quality and the usefulness of these tools for intralingual translation. They also elaborate on how they tried to adapt the MQM scheme to intralingual translation and which annotation frame they used. The present paper builds on that work but aims to go into further detail in the error categorization, integrating the analysis of four different commercial models and comparing their performance in intralingual translation.

3. Materials and Methods

3.1. Corpus Data

As mentioned above, we analyse outputs of four different commercial models. We use three models of the SUMM AI tool (summ-ai.com, last accessed on 14 October 2025, the company SUMM AI offers different licenses for freelancers, authorities and companies), and one general LLM, i.e., ChatGPT-4o (openai.com, last accessed on 15 December 2024, translations produced between October 2024 and December 2024 using the university's AI platform UHiKI which operates on university servers). The motivation for these tools is based on the fact that ChatGPT is the most popular AI tool used by a broader audience and the SUMM AI tool is the most well-known tool for Easy/Plain German translation. SUMM AI provides translations for both Easy and Plain German and also offers further functions, such as image generation or glossaries. On the user interface, the source text is entered on the left and after selecting at the top whether the text should be translated into Easy Language or Plain Language, the target text appears on the right.

The main difference between the first three models of SUMM AI and GPT-4o is that the former are dedicated to translating texts into Easy German and Plain German, as they were specifically trained and fine-tuned for this task. ChatGPT-4o, on the contrary, is a general LLM that was not specifically fine-tuned for intralingual translation. Another main difference is that ChatGPT is an interactive chatbot that can be prompted, whereas the SUMM AI tool does not contain any prompting functions. ChatGPT translations were generated using a two-step prompt [39].

Moreover, there is a difference between the three SUMM AI models themselves: while the baseline model is a general model for intralingual translation, the other two – model 1 (M1) and model 2 (M2) were fine-tuned for translation within a specific domain, i.e., health communication. While being fine-tuned on the same parallel data, model 1 and model 2 differ in the underlying LLM: model 1 shares the underlying LLM with the baseline system. With respect to LLM architectures, these are the only known technical details available to the research team. As all three models are proprietary and not open-source, we adopted a black box-approach, focusing exclusively on the interaction between model inputs and outputs without access to further specifications concerning LLM architectures.

For our study, we use the dataset provided by [21,39]. This dataset consists of 30 source texts from the website of the German health publisher Apotheken Umschau (https://www.apotheken-umschau.de, last accessed on 30 October 2025) and the four outputs of the models mentioned above. The 30 source texts cover a broad range of different topics, such as food poisoning, vaccination or bladder infection. The 30 source texts are representative of the entire Plain Language corpus of the Apotheken Umschau (which consists of approx. 200 texts) as they cover all major supercategories present in the corpus, including diseases, medication, first aid, and contraception. Table 1 illustrates the size of the subcorpora in terms of token and sentence number. As seen from the table, although we deal with the translation variants of the same source texts, their size varies a lot, especially if the baseline and the GPT-4o outputs are concerned.

Table 1. Corpus size in tokens and sentences.

	Sources	**Baseline**	**M1**	**M2**	**GPT4o**
tok	62,050	66,641	43,140	40,475	13,486
sent	3167	5639	3587	3726	949

3.2. Evaluation Framework for Intralingual Translation

For our analysis, we use the adapted MQM framework introduced by [39]. The authors first selected relevant parts of the MQM framework, which were then adapted to fit the purpose of intralingual translation (see Table 2). The adapted evaluation framework includes four main categories that can be subdivided into further subcategories. In our study, we revise this error categorization relying on the issues that arise during a consensual annotation process. In the following, we report in detail on how annotation was performed, which issues arose and how we revised the evaluation framework, i.e., removed or introduced new error categories.

3.3. Annotation Process

3.3.1. First Round of Annotation

The first round of annotation aimed to deliver the first insights into the categories developed as presented in Table 2. After an initial briefing by the project leads, five student assistants worked on the consensual annotation. All five student assistants had been trained as intralingual translators. They had also been thoroughly introduced to the intralingual version and the adapted version of the MQM framework. For the first round of annotation,

three texts were chosen. Three of the student assistants worked on the three texts produced with model 1.

Table 2. Categories and subcategories used for error classification.

Error Type	Definition
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 text.**
Inconsistent terminology	Multiple terms are used to describe the same concept when just one term is needed or appropriate.
Wrong term	Use of a term that is not the term a domain expert would use or that gives rise to a conceptual mismatch
Accuracy	**Content in the target text does not match the propositions from the source text.**
Mistranslations	Target content does not accurately represent the source content.
Ambiguous content	Ambiguity is introduced where specification is needed.
Hallucination	Machine translation produces an output that is totally decoupled from the source text.
Wrong or missing explanation	Explanation is necessary and added, but does not represent the information from the source text (wrong) or an explanation is needed but is not present in the target text (missing).
Incomplete information	Relevant information from the source text is missing in the target text.
Linguistic conventions	**Errors related to the linguistic level of the source text.**
Grammar	Grammatical rules are violated in the source text.
Punctuation	Punctuation is used incorrectly.
Spelling	Words are misspelled.
Cohesion and coherence	Connectors necessary to understand the text as a whole are missing or incorrect (cohesion). Semantic relationships within the text are not clear (coherence)
Audience appropriateness	**Content in the target text is not valid, appropriate or acceptable for the target audiences.**
Inaccurate advice	Target text contains advice that is not in the source text or that is not suitable for the situation in question.
Stigmatising content	Content can lead to stigmatization of end users.

The second group annotated the three target texts that had been produced with ChatGPT-4o. Each text was worked on by at least two people independently to prevent bias and to allow for a more thorough analysis and evaluation of the error categories. The annotators were supervised by the project leads, who were available for questions and issues arising during the annotation process.

The annotators were asked to meticulously compare the source text with the machine-translated text to identify and mark errors. These errors were then classified based on the categories explained in Table 2. The first round of annotation also allowed for an error to be included in two or more relevant categories, i.e., resulting in an ambiguous annotation. This was especially the case for grammatical errors, which could be interpreted both as MT hallucinations and errors in the cohesion of the text. This initial strategy allowed for comparison between different categories and potential difficulties faced by annotators of both ChatGPT-4o and SUMM AI texts. Errors that were not clearly categorizable were put into the respective "Not classifiable" category (for more details see Section 4.2 below). These were then further analyzed and discussed in subsequent rounds of annotation. After annotation of the first six texts, the two teams met up with the project leads to discuss issues and difficulties that arose during the annotation process.

More specifically, they discussed the following questions:

- Which parts of the categorization did not work and why?
- Where does the categorization miss clear definitions?

- Where are better examples needed?
- Which errors could not be fit into the existing categories and why?
- What kinds of categories are missing?
- Which categories have not been used at all?

For items in the "Not classifiable" categories, the two teams worked together with the project leads to find suitable categories if possible. New categories were created where necessary (i.e., if none of the already existing categories fit the specific error types), and the annotation categories were updated and extended accordingly. Categories that had fewer than 10 annotated items after the first round were deleted since discussion within the team showed that they were not relevant for the task at hand. More on these deleted categories can be found in Section 4.3.

3.3.2. Updated Annotation Categories

Following the consensual annotation in the first round of annotation (see Section 3.3.1 above), we updated the annotation categories to reflect the changes and adaptations identified. Then, further consensual annotation took place. This includes mainly the decision that any errors within the category "Audience appropriateness" were to be classified as general errors within this category without classifying them into further subcategories. More on this can be found in Section 4.3.

We will provide the annotated data, as well as annotation guidelines, in a dedicated github repository.

4. Results

In the following, we present both the annotation results based on the updated evaluation framework and the issues that arose during the annotation process.

4.1. Error Distribution

Across all four models, most errors occurred in the "Accuracy" category (2852 errors in total), closely followed by errors in the "Linguistic conventions" category (2838 errors in total). To enable a comparison of error distributions across the four models and taking into account the varying corpus sizes in tokens reported in Table 1, Figure 1 presents the total error distribution across all categories per tokens normalised to 1000, following the formula:

$$\frac{\text{amount of errors in a specific category}}{\text{corpus size in tokens}} \times 1000 \quad (1)$$

Across all models, a total of 936 terminology errors were identified. When normalised against the total number of tokens as reported in Table 1, the Baseline model exhibited the highest number of such errors, whereas Model 1 showed the lowest incidence of terminology errors (see Figure 2). Overall, however, the incidence of this error category remained relatively low across all models. The subcategory "Wrong term" accounted for a large majority of errors across models (837), far exceeding those in "Inconsistent use of terminology". Examining the normalised error distribution in the subcategory "Wrong term", Model 1 (2.62) and Model 2 (3.78) exhibited the fewest errors. ChatGPT-4o (4.52) also showed a comparatively low incidence of errors in this category, while the Baseline model recorded the highest ratio (12.56).

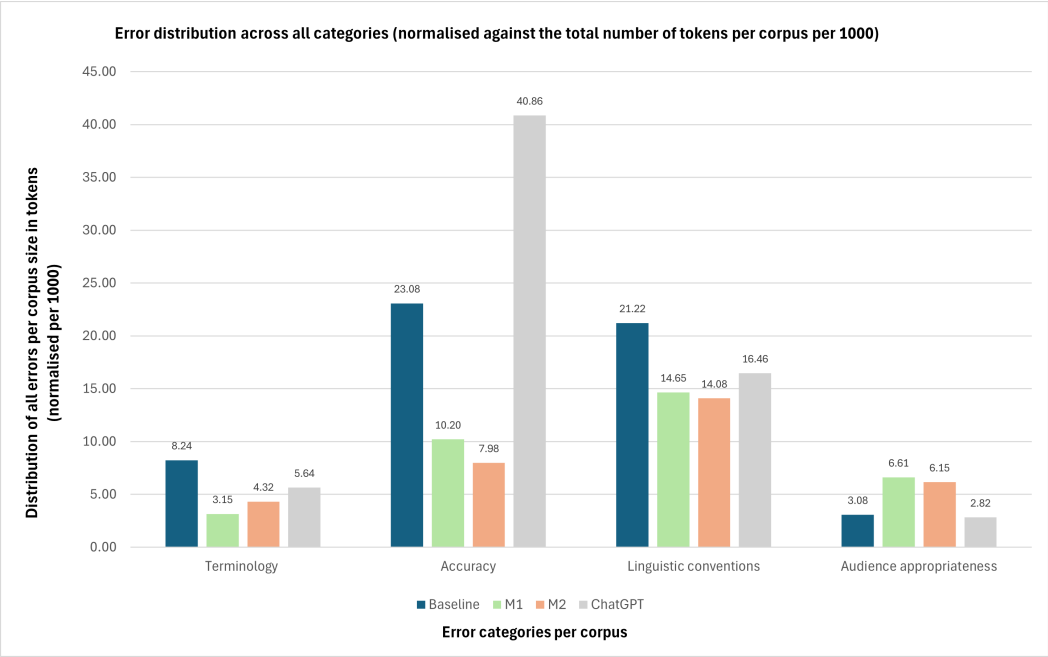


Figure 1. Error distribution across all categories.

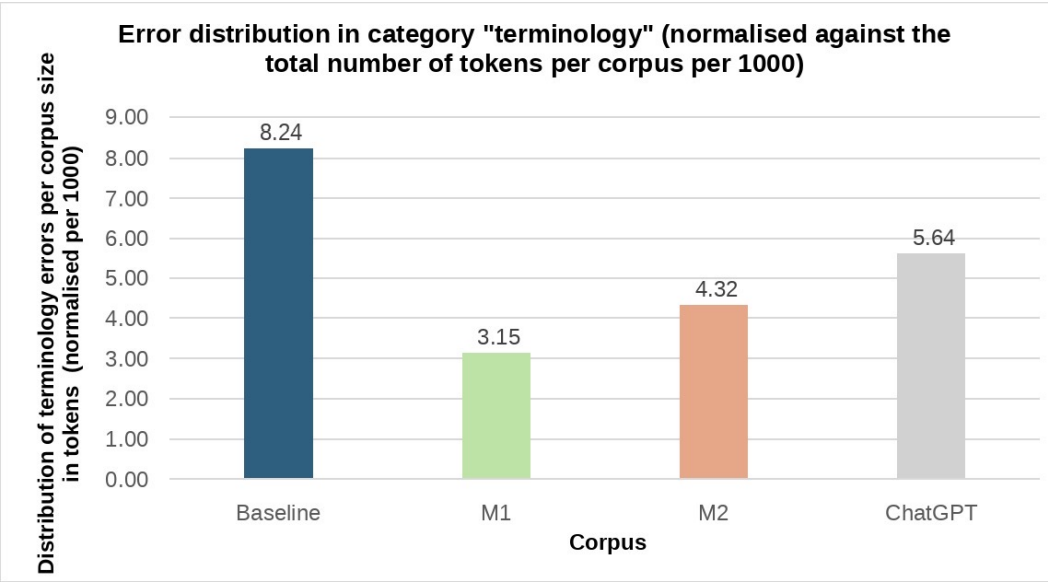


Figure 2. Error distribution in category "Terminology".

This category is further divided into the subcategories "Inconsistent use of terminology" and "Wrong term". The latter makes up 89.42 percent of all errors in the category "Terminology". Typical errors include the usage of multiple different terms like "Medikamente" (EN: "medication") and "Präparate" (EN: "medicinal preparations") across a single text. This can be confusing for the readers if they are not familiar with the language and therefore do not know that these terms can be used interchangeably. The consistent use of a single term for one specific instance is recommended as to not confuse the readers. In this case, it would be advisable to stick with the more commonly known and more frequent term "Medikamente" (EN: "medication"). This type of error also occurs with the names of illnesses. One text about "Fersensporn" (EN: "heel spur") used the term "Fersensporn" and the shortened "Sporn" (EN: "spur") interchangeably. Another text used both the old spelling ("Hämorrhoiden") as well as the more modern spelling of the word ("Hämorriden") when referring to the illness (EN: "hemorrhoids"). This could confuse the readers,

especially if they are not familiar with different names or spellings of certain illnesses. In the subcategory "Wrong term", commonly found errors include a wrong transfer of terms from the original text to the Easy Language text. This includes errors like "Physiotherapie" (EN: "physiotherapy"), which was translated by the AI model to "Sporttherapie" (EN: "sport therapy") or "starke Sehbehinderung" (EN: "severe visual impairment"), which was machine translated to "schlechte Sehfähigkeit" (EN: "poor eyesight"). Especially in the medical field, errors like this can cause harm to potential patients, which is why an accurate translation of terminology is important.

In the category of Accuracy, across all models, a total of 2852 errors were documented. When normalised (see Figure 3), ChatGPT-4o exhibited the highest incidence of accuracy-related errors (40.86), while Model 2 demonstrated the best performance with the lowest error rate (7.98). Examining error types, the "Mistranslation" subcategory was clearly predominant, accounting for 1596 errors—outnumbering by far the subcategories "Wrong addition" (91 errors), "Missing addition" (109 errors) and "Completeness" (743 errors). Notably, issues with completeness were more prevalent than errors related to added or missing content, particularly in outputs from both ChatGPT-4o and the Baseline model. One possible explanation for the high incidence of accuracy errors in the ChatGPT-4o outputs is that ChatGPT-4o, as a general-purpose language model with generalized training, is more prone to producing phrases and hallucinations that introduce inaccuracies than a model that was fine-tuned with domain-specific data and Plain Language human gold standard translations. In interlingual translation, LLMs are also prone to accuracy errors, as they are usually more focused on fluency, which often results in a high number of hallucinations. Thus, our study has shown that the same applies to intralingual translation.

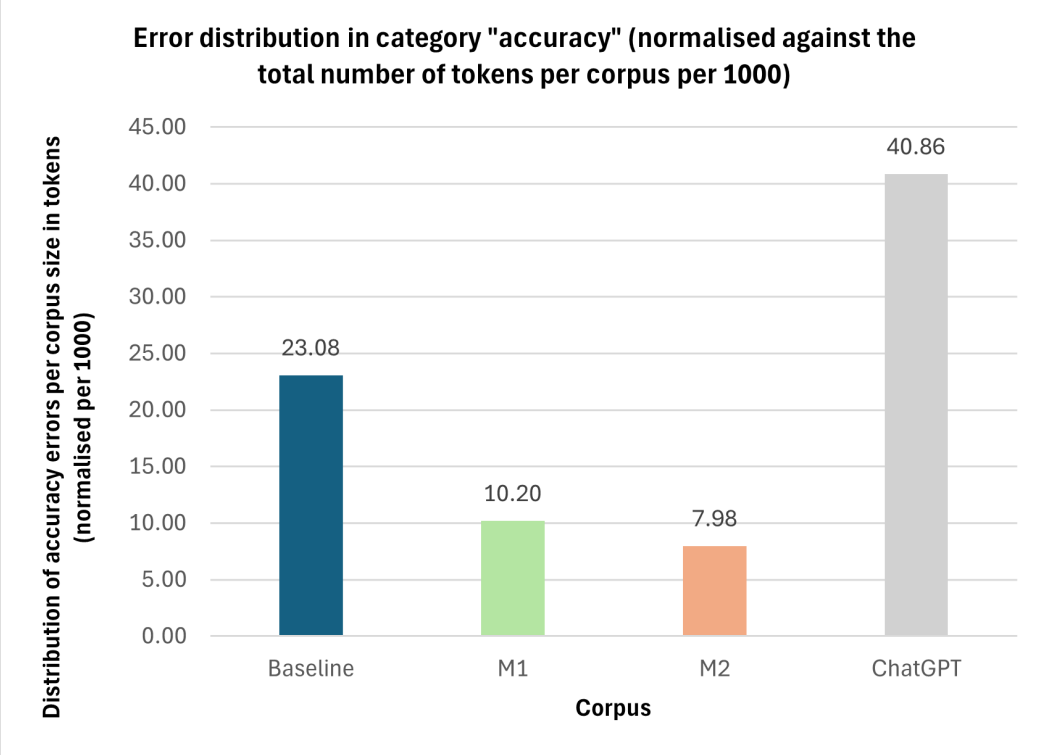


Figure 3. Error distribution across category "Accuracy".

This category is further subdivided into "Mistranslation", "Wrong addition", "Missing addition" and "Completeness". The categories are ranked as follows (from most errors to least errors):

1. Mistranslation (1909 errors).

2. Completeness (743 errors).
3. Missing addition (109 errors).
4. Wrong addition (91 errors).

Mistranslation includes various types of mistranslation errors, such as mistranslation of technical relationships. These errors refer to instances where the machine-translated texts misinterpreted the original texts and produced texts that showed wrong relationships between certain elements of the texts. Some examples of this are:

- "Mikropillen mit neuen Gestagenen wie zum Beispiel Gestoden, Desogestrel oder Drospirenon haben ein höheres Risiko für eine Venenthrombose als andere Mikropillen." (EN: Micro-pills containing new progestogens such as gestodene, desogestrel or drospirenone carry a higher risk of venous thrombosis than other micro-pills.)
  The way this is phrased in the German MT text implies that the micro-pills themselves are at risk of venous thrombosis, when in fact the people who take these pills are at a higher risk of venous thrombosis. This could potentially be confusing to readers with lower reading skills.
- "Wenn ein großer Teil der Bevölkerung gegen eine Krankheit immun ist, gibt es auch weniger Krankheiten." (EN: When a large proportion of the population is immune to a disease, there are less diseases overall.)
  The German MT translation wrongly implies that the more people are vaccinated and thus immune to a certain disease, the fewer diseases exist overall. This is a wrong interpretation of the fact that if more people are immune to a disease, fewer cases of this disease should (in theory) occur.
- "Die Studien müssen gut sein, wenn sie für die evidenzbasierte Medizin gelten sollen." (EN: Studies must be of good quality if they are to be considered valid for evidence-based medicine.)
  The machine translation misinterprets the fact that studies must be conducted according to the criteria of evidence-based medicine in order to be reliable.

Other instances include mistranslations where the MT system hallucinated information that was not present in the original text. This includes wrong additions to sentences, whole sentences and paragraphs that were not included in the original text, as shown in the following examples:

- "An den Kniekehlen und dem Ende des Darmes ist ein Gelenk." (EN: There is a joint at the back of the knees and at the end of the intestine.)
  This is simply not true, as there is no joint at the end of the human intestine.
- In one case, the MT system even hallucinated information in English: "OPIATES, COX inhibitors and other medicines." The MT system put this English fragment as the heading to a German text on painkillers.

A total of 2838 linguistic convention errors were observed across all models (see Figure 4), with the Baseline model exhibiting the highest relative frequency per token (21.22). Models 1 and 2 showed nearly identical error rates (14.65 and 14.08, respectively). ChatGPT-4o displayed a higher incidence than Models 1 and 2, but a lower incidence than the Baseline model (16.46). The largest share of linguistic errors resulted from difficulties with subordinate clauses, with 874 instances overall. It is notable that ChatGPT-4o did not produce any errors in this subcategory; the majority of them came from the Baseline model (6.02) and SUMM AI model 2 (8.00), while Model 1 accounted for the fewest subordinate clause errors among the non-ChatGPT systems (3.66).

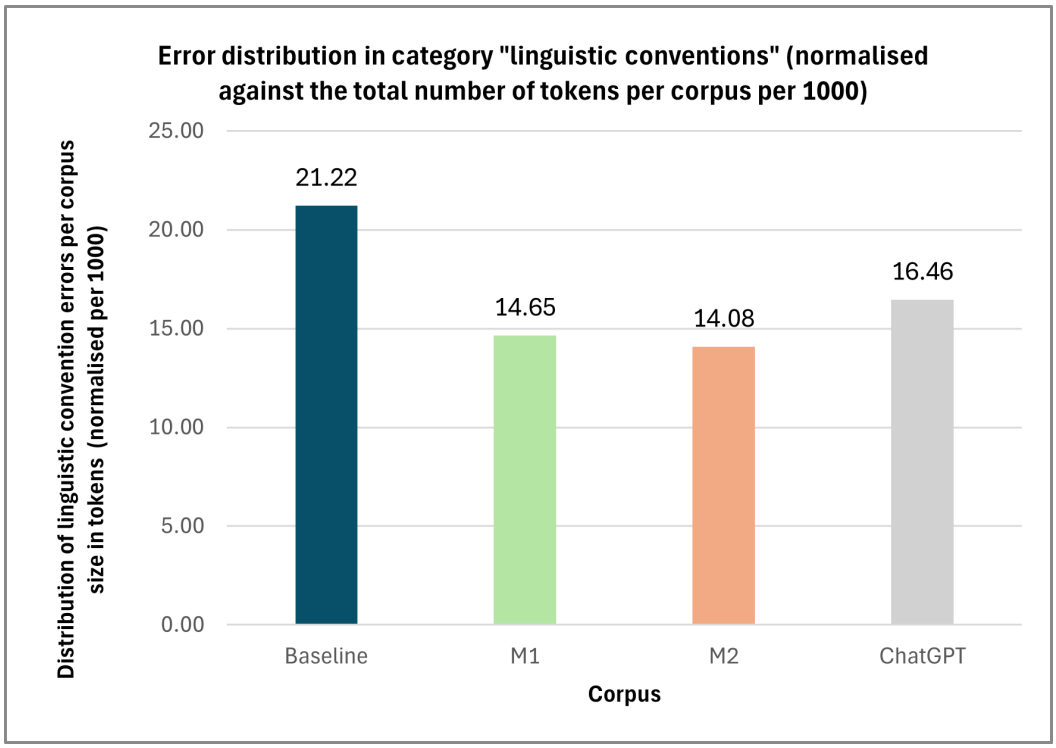


Figure 4. Error distribution in category "Linguistic Conventions".

This category was subdivided into the categories "Grammar", "Punctuation", "Spelling", "Unclear reference" and "Textual conventions". The categories are ranked as follows (from the most errors to the least errors):

1. Textual conventions (1822 errors).
2. Spelling (356 errors).
3. Grammar (294 errors).
4. Unclear reference (229 errors).
5. Punctuation (137 errors).

Most errors in the subcategory "Textual conventions" refer to textual coherence and cohesion, as shown in the following examples:

- "Diese Beschwerden können entstehen." (EN: These complaints may arise.).
  The sentence does not fit into the context of the text due to a missing reference and conjunction.
- "Manche Antibiotika machen Menschen mit Diabetes zu Diabetikerinnen oder Diabetikern mit Zuckerstürzen." (EN: Some antibiotics cause people with diabetes to become diabetic with hypoglycaemia.)
  This sentence seems redundant and wrong, as the information in the German sentence tells the reader that people who already suffer from diabetes might suffer from diabetes caused by the antibiotics.
- "Was hilft Pfefferminze?" (EN: What does peppermint help?)
  This sentence is grammatically wrong, as it should be "Gegen was hilft Pfefferminze?" (EN: What does peppermint help with?)

Most grammar errors refer to the German differentiation between "das" and "dass". "Das" is either used as an article or a pronoun, while "dass" is a conjunction that introduces a subordinate clause. Other grammar errors occurred when the MT translation wrongly hyphenated words like "Lebens-Mittel" (Lebensmittel, EN: food), "Zucker-Stoff-Wechsel" (Zuckerstoffwechsel, EN: glucose metabolism) or "All-Tag" (Alltag, EN: daily

life). These wrong hyphens split up the words into individual parts that make no sense with the hyphen.

Within the error category "Audience appropriateness" (as illustrated in Figure 5), a total of 777 cases were documented across all systems. According to the normalised error distribution, Model 1 exhibited the highest frequency (6.61), followed by Model 2 (6.15) and the Baseline model (3.08). ChatGPT-4o demonstrated the lowest incidence of errors (2.82).

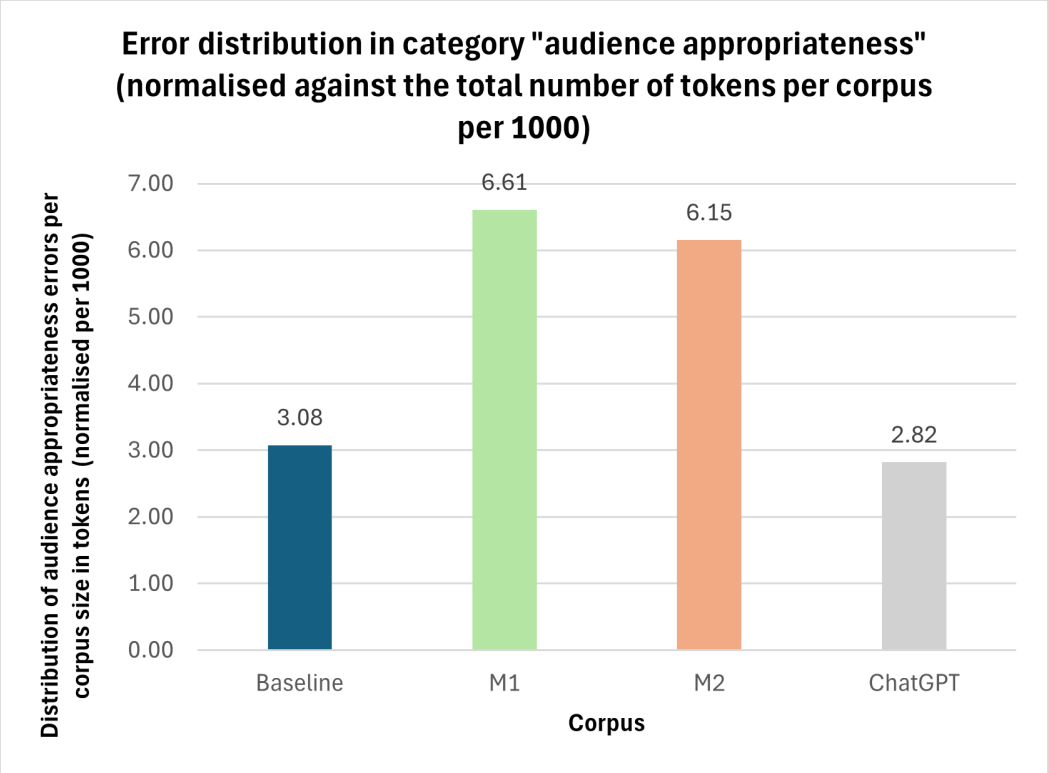


Figure 5. Error distribution in category "Audience Appropriateness".

Within the category of "Audience appropriateness" we found several errors concerning the form of address, i.e., the German reader was addressed using the informal form "Du" ("you") instead of the formal form "Sie" ("you"). Further errors were found in sentences that used stigmatising language and/or unnecessary repetition:

- "Die Schleimhaut von der Gebärmutter geht weg." (EN: The lining of the uterus goes away).
  The German phrasing is very child-like and not appropriate for adult audiences.
- "Die Pille hat Vor- und Nachteile. Die Mikropille hat viele Vorteile". (EN: The pill has advantages and disadvantages. The micro pill has many advantages.)
  This is unnecessary repetition, which can be stigmatising for readers.

"Audience appropriateness" is the only category in which no subcategories remain. During the first round of annotation, attempts were made to classify the errors into further subcategories (i.e., End User Suitability, Modality and Stigmatizing). This resulted in a challenge since this classification could not be verified through end user testing. From a text perspective, hazardous segments could be recognized, but not classified into definitive subcategories. Further application and development of suitable subcategories through end user testing and verification is a desideratum that remains to be addressed.

4.2. Error Category "Non-Classifiable"

Some instances of "non-classifiable errors" were found across all categories after the first two rounds of consensual annotation. The most common occurrences include:

- Unnecessary repetition of information
- Inappropriate addressing of readers (i.e., sudden change from formal "Sie" ("you") to informal "Du" ("you") or informal "Wir" ("we"))
- Complex sentence structure
- Audience appropriateness

For example, there were a couple of instances where it was not clear which target audience was addressed by the text. This made it unclear whether the target audience was readers who wanted to have general information about the pill, as illustrated in example (1-a), or if it was specifically intended to result in advice towards future actions of the readers, as in example (1-b). In some instances, the text produced by the SUMM AI model 1 switched to addressing the readers directly, resulting in a direct call to action, as illustrated in the following example:

(1) Sie sprechen mit dem Arzt. Und der Arzt untersucht Sie. (EN: You talk to the doctor. And the doctor examines you.)

These instances were put into the "Not classifiable" sub-category under the main category "Linguistic conventions" during the first round of annotation. This occurred mainly in a text about contraceptives, in which it was not clear if the text pre-supposed that readers had already decided that they would want to take the pill or if it was supposed to be a general text informing readers about the pill as a contraceptive method. After checking with the project leads, it was decided to re-classify these errors into "Audience appropriateness", since a direct call to action was not included in the source text.

4.3. Further Observations

All four models produced errors in almost every category. Notably, the only sub-categories to which this does not apply are "Collocation" and "Sub-ordinate clauses", where ChatGPT-4o was the only model that did not produce any errors.

In the category "Accuracy", the sub-category 2.1.4 Unit conversion remained unused. This, however, was to be expected as, unlike in interlingual translation, in intralingual translation unit conversion is usually not necessary. After two rounds of consensual annotation, the following categories were used less than 15 times across all 120 texts:

- 2.1.6 Date/Time (4 annotated segments).
- 2.1.7 Entity (4 annotated segments).
- 2.5.3 Incomplete procedure (10 annotated segments).

This prompted the annotators to evaluate these categories to see if the errors in these categories could be put into different categories or if they are relevant at all. After this evaluation, it became clear that these errors could also be put into more fitting categories, making the categories mentioned above obsolete. Subsequently, the sub-categories 2.1.6 Date/Time and 2.1.7 Entity remained unused. Category 2.1.6 Date/Time was not needed, as the errors in this category stemmed from the MT systems using the original publication dates from the source texts. Since the translations are to be used as stand-alone texts, they should be marked with the date the translation was published. Hence, the category was not used in the final annotation process.

In the category "Linguistic conventions", sub-categories 3.4 Sorting, 3.6 Unintelligible and 3.7 Textual conventions remained unused in the final annotation process. This shows that these subcategories are primarily relevant to interlingual translation, where differences in textual conventions or sorting are more likely to occur. In interlingual translation, for example, items that are alphabetically ordered in the source text may not retain the same order in the target text, resulting in potential sorting errors. Also, textual conventions

often differ between the source and target language, resulting in errors, whereas in Plain Language, we often do not even have well-established textual conventions. Moreover, it is likely that instances which might have been classified as "Unintelligible" were instead annotated under the category "Accuracy".

4.4. Difficulties

One of the main difficulties in the error analysis of intralingual translation is the lack of alignment. Since intralingual translation is much more liberal than interlingual translation, adequately aligning source and target texts on the sentence level is not possible [40]. This lack of adequate alignment not only makes the calculation of automatic metrics, such as the SARI score, difficult [41] but it also complicates the identification of translation errors in the target texts. For example, incomplete enumerations can be more easily overlooked, especially when the enumeration occurs within a sentence (rather than in bullet points) and only a single element is missing. The annotator has to meticulously compare source and target texts manually, which is not only very time-consuming but also prone to errors. In future studies, it would therefore be worth-investigating to investigate whether LLMs can be used for detecting and annotating errors.

What stood out about the errors annotated in the ChatGPT-4o texts compared to the SUMM AI texts is the category "Incomplete list". While the SUMM AI corpus contains only two errors of this category, the ChatGPT-4o corpus contains 55 errors of this category. A reason for this could be the length of the ChatGPT-4o texts. Since these texts are considerably shorter than the SUMM AI texts, errors in lists are more easily detectable for annotators, while they might have been overlooked in longer texts.

In addition, the annotators often had a hard time reaching consensus on whether information that was left out in the target text should be classified as an error or not. In intralingual translation, information that is not absolutely necessary for text comprehension or that is irrelevant for the target reader can be left out in the target text so that missing information is not an error per se. However, omissions in the target texts that affect text comprehension should be annotated as errors, even though the error severity of "Incomplete list" errors is expected to be low, especially in comprehensibility-enhanced communication. Still, our study revealed that the question of whether an omission is an error or not is highly disputable. Annotators' decision-making could have been facilitated by providing some sort of support, for example, in the form of a decision tree that guides their decisions during the annotation process.

Another difficulty was dealing with passages or phrases that violated the guidelines for translation provided by the Apotheken Umschau. SUMM AI model 1 and model 2 have been fine-tuned for Plain Language translation and thus would, in theory, be able to follow the translation guidelines that Apotheken Umschau uses for their texts. However, the Baseline model and ChatGPT-4o were not fine-tuned, so these models were not capable of following guidelines when producing the translations. Therefore, to be still able to compare the performance of the different models, we decided not to use the tag "Textual conventions" for any errors that would be specific to the guidelines of Apotheken Umschau. Still, when only evaluating models that have been fine-tuned with the guidelines, these instances should count as errors.

We noticed that the texts translated by the SUMM AI models 1 and 2 are more detailed at the beginning than at the end. At the end, information is often omitted or only summarised. The latter appears to be selected arbitrarily and does not follow any apparent rules. We checked with SUMM AI to ask about potential character limits within the tools, but SUMM AI clarified that no character limit has been programmed for their tools. Therefore, we were not able to determine how and why omissions and summarisation occur.

5. Discussion

All in all, our data demonstrate substantial variation in error patterns between systems: although all four MT models produced errors in every category, the fine-tuned health-domain models showed substantial gains in accuracy and reduction in critical mistakes. Nevertheless, no model was free of fundamental errors—especially regarding information omissions with potential impact on comprehension, and the production of non-idiomatic or stigmatising text passages.

The error analysis revealed that accuracy and linguistic conventions are the most frequent error categories across all evaluated MT systems. In particular, fine-tuned domain-specific models significantly reduced the number of mistranslations and errors related to subordinate clauses compared to general-purpose LLMs like ChatGPT-4o. While domain-adapted models (SUMM AI variants) achieved lower rates of terminology and grammatical errors, they still exhibited non-negligible rates of omissions and issues with audience appropriateness, especially in health communication contexts where accurate advice and accessibility for lay readers are critical. The high amount of non-negligible errors in the fine-tuned models can be explained by the fact that producing Plain Language translations is a very challenging, complex task that requires much more than just implementing rules or simplifying certain words or sentence structures. For example, the text has to be adapted to the prior knowledge of the target audience, i.e., certain terms have to be explained and examples have to be added. Our study has shown that without explicitly prompting the model to do so, the model itself does not add this necessary information to the source text, leading to a higher amount of errors in the category "Audience appropriateness". The high rate of omissions in the SUMM AI translations indicated that the model fails to differentiate between important and less-important information of the text, but seems to treat all information as equally important. In Plain Language translations, it is indeed common practice to omit information; however, only information that is not relevant for the overall text comprehension is to be omitted. For human translators, it is relatively easy to differentiate between relevant and non-relevant information. The omissions made by the SUMM AI models, however, were detrimental to the text comprehension and therefore not acceptable, which led to the high error rate in the category "Ommission".

An examination of error subcategories showed that the incidence of incorrect terminology and ambiguous content was notably lower in domain-specific models, but completeness and cohesion remained problematic across all systems. Interestingly, ChatGPT-4o produced more "Incomplete list" errors, likely due to its tendency towards brevity, whereas the Baseline model struggled most with overall linguistic conventions.

The frequency and nature of such errors suggest that even advanced models are not reliably able to anticipate or reproduce target-group-specific demands without further human post-editing or iterative prompting. Models trained with specific editorial guidelines outperformed general models in both syntactic complexity and adherence to Plain Language standards.

As so far, an error scheme for evaluating intralingual translations did not exist, our study is first of all intended to provide researchers and translators with an appropriate error scheme for intralingual translations that can be used for evaluating Plain Language translations, both human and machine translations. In addition, our analysis is intended to highlight error categories that seem to be especially problematic for LLMs. The results can be used both by model developers to improve their systems and by translators and post-editors. For translators and post-editors, the error distributions identified in our study provide concrete guidance on where particular vigilance is required when working with AI-generated Plain Language translations. For instance, we found that negation is a recurrent source of errors, which can not only lead to substantial changes in meaning but

can also have fatal consequences in the field of healthcare. Therefore, our results should prompt translators and post-editors to meticulously compare the target text with the source text and to carefully verify that all negations have been transferred correctly. All in all, our study has shown that including LLMs in translation workflows is definitely useful, however, only for translators as end users, and so far not for the target groups as end users. If translators include LLMs in their professional workflow, it is essential that they have certain knowledge on how LLMs work and where their limitations and problems are, so that they are, for instance, able to identify hallucinations and biases and also able to detect and correct other severe mistakes, such as omitted negation markers. If LLMs are included in professional translation workflows, they can act as assistive components and can be used, for example, for creating drafts, reformulating phrases and passages, or providing definitions and synonyms. Still, human experts remain responsible for validating and correcting the outputs as they are the ones that are held responsible for the final translation. Therefore, human translators are still indispensable as they will remain the necessary expert-in-the-loop.

Among the limitations of our study is that we have not calculated the inter-annotator agreement. This was because we first wanted to figure out to what extent MQM is suitable for intralingual translation and what categories need to be added and deleted. For that reason, we preferred to work with a consensual annotation process, which constitutes a form of inter-annotator agreement in qualitative studies [42]. However, we are aware of the fact that calculating the inter-annotator agreement ensures transparency and strengthens the validity of the framework and the reliability of the annotation process. Calculating the inter-annotator agreement adapted to qualitative studies will therefore be part of our future studies, as it not only shows us how consistently different annotators apply the established framework but also reveals ambiguities and cases that may be analysed in more detail.

Another limitation is that so far, we have also not ranked the errors in terms of severity levels. This, however, is important when assessing the final quality of a translation because some errors, such as spelling mistakes, are less severe and harmful, whereas other errors, such as mistranslation or omission of negation markers, can have severe and even fatal consequences and may lead to wrong instructions. In future studies, we therefore aim to assess error severity and calculate a final MQM score for the different outputs. We also plan to expand this study to other LLMs and aim to test different prompting approaches to see whether the translation quality may depend on how a model is prompted.

In addition, we plan to replicate the study with a newer version of the models and compare the outcomes over the years. This will allow us to examine whether model performance improves, stagnates, or even declines over time, given the fact that since 2024, large volumes of AI-generated texts have been produced and have subsequently been used as training data, with the risk of deteriorating output quality potentially far below the high-quality standards of professional human translations.

Our study has shown that manual error classification and annotation are highly time-consuming and costly; therefore, investigating whether LLMs can be effectively applied to error annotation in intralingual translation represents another promising avenue for future research. A further learning of this study is the fact that a corpora comparison using MQM or our adapted MQM framework for intralingual translation would be more effective if the texts could be more easily aligned. Only then can errors be compared effectively. We also aim to test whether, in intralingual translation, automatic metrics align with human judgement. However, when calculating automatic metrics, it has to be kept in mind that in intralingual translation, an adequate alignment of source and target text is often not possible. This is because, due to the high amount of explanations and added examples in Easy and Plain Language and the fact that source text information can be omitted by

the translators, source and target text often differ a lot, so that a 1:1 alignment is often not possible.

In the future, we also consider prompting LLMs to automatically evaluate the outputs based on our scheme. Similar approaches were tested for interlingual translation [43–45] and could be extended to intralingual translations.

Author Contributions: Conceptualization, S.D., S.H.G., E.L.-K. and C.M.; methodology, S.D., S.H.G., E.L.-K. and C.M.; formal analysis, A.W.; writing—original draft preparation, S.D., S.H.G., E.L.-K., C.M. and A.W.; visualization, A.W.; funding acquisition, S.D., S.H.G., E.L.-K. and C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Culture of Lower Saxony and its program "zukunft.niedersachsen".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study, i.e., the annotated data and the annotation guidelines, will be made openly available in a dedicated github repository. https://github.com/katjakaterina/chatgpt4easylang accessed on 1 January 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Teubner, T.; Flath, C.M.; Weinhardt, C.; Van Der Aalst, W.; Hinz, O. Welcome to the era of chatgpt et al. the prospects of large language models. *Bus. Inf. Syst. Eng.* **2023**, *65*, 95–101. [CrossRef]
2. Dale, R. A year's a long time in generative AI. *Nat. Lang. Eng.* **2024**, *30*, 201–213. [CrossRef]
3. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
4. Banerjee, S.; Lavie, A. 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, Ann Arbor, MI, USA, 29–30 June 2005; pp. 65–72.
5. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A neural framework for MT evaluation. *arXiv* **2020**, arXiv:2009.09025. [CrossRef]
6. Flesch, R. "Simplification of Flesch Reading Ease Formula": Reply. *J. Appl. Psychol.* **1952**, *36*, 54–55. [CrossRef]
7. Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; Callison-Burch, C. Optimizing statistical machine translation for text simplification. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 401–415. [CrossRef]
8. Castilho, S.; Doherty, S.; Gaspari, F.; Moorkens, J. Approaches to human and machine translation quality assessment. In *Translation Quality Assessment: From Principles to Practice*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 9–38.
9. Lommel, A.; Uszkoreit, H.; Burchardt, A. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica* **2014**, *12*, 455–463. [CrossRef]
10. McDonald, S.V. Accuracy, Readability, and Acceptability in Translation. *Appl. Transl.* **2022**, *16*, 21–29. [CrossRef]
11. Baumgart, M.; Hösel, C.; Breck, D.; Schuster, M.; Roschke, C.; Ritter, M. Development of a holistic web-based interface assistance system to support the intralingual translation process. In Proceedings of the International Conference on Human–Computer Interaction, Virtual, 24–29 July 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 505–511.
12. González-Sordé, M.; Matamala, A. Empirical evaluation of Easy Language recommendations: A systematic literature review from journal research in Catalan, English, and Spanish. *Univers. Access Inf. Soc.* **2024**, *23*, 1369–1387. [CrossRef]
13. Luque Lopéz, L. Leveraging Large Language Models to Translate into Easy Language: An Exploratory Study on University Websites. Master's Thesis, Université de Genève, Geneva, Switzerland, 2025.
14. Freyer, N.; Kempt, H.; Klöser, L. Easy-read and large language models: On the ethical dimensions of LLM-based text simplification. *Ethics Inf. Technol.* **2024**, *26*, 50. [CrossRef]
15. Jakobson, R. On linguistic aspects of translation. In *The Translation Studies Reader*; Routledge: Abingdon, UK, 1959; pp. 46–60.
16. Maaß, C. *Easy Language–Plain Language–Easy Language Plus: Balancing Comprehensibility and Acceptability*; Frank & Timme: Berlin, Germany, 2020.
17. Maaß, C.; Hernández Garrido, S. Einfache Sprache: Einfach, leicht, verständlich? In *Einfache Sprache mit KI-Tools: Ein Leitfaden für die Redaktionelle Praxis*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 17–36.

18. *DIN ISO 24495-1:2024-03*; Einfache Sprache—Teil 1: Grundsätze und Leitlinien. Deutsches Institut für Normung: Berlin, Germany, 2024.
19. *DIN 8581-1*; Einfache Sprache—Anwendung für das Deutsche—Teil 1: Sprachspezifische Festlegungen. Deutsches Institut für Normung: Berlin, Germany, 2024.
20. *ISO 24495-1:2023*; Plain Language—Part 1: Governing Principles and Guidelines. International Organization for Standardization (ISO): Geneva, Switzerland, 2023
21. Deilen, S.; Lapshinova-Koltunski, E.; Garrido, S.; Hörner, J.; Maaß, C.; Theel, V.; Ziemer, S. Evaluation of intralingual machine translation for health communication. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation, Sheffield, UK, 24–27 June 2024; Volume 1, pp. 469–479.
22. Maddela, M.; Dou, Y.; Heineman, D.; Xu, W. LENS: A Learnable Evaluation Metric for Text Simplification. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 16383–16408. [CrossRef]
23. Dale, E.; Chall, J.S. A formula for predicting readability: Instructions. *Educ. Res. Bull.* **1948**, *27*, 37–54.
24. Spache, G. A new readability formula for primary-grade reading materials. *Elem. Sch. J.* **1953**, *53*, 410–413. [CrossRef]
25. Coleman, M.; Liau, T.L. A computer readability formula designed for machine scoring. *J. Appl. Psychol.* **1975**, *60*, 283–284. [CrossRef]
26. Isnaeni, N.R. Readability of English written materials. *Elite Engl. Lit. J.* **2017**, *1*, 179–191.
27. Cachola, I.; Khashabi, D.; Dredze, M. Evaluating the Evaluators: Are readability metrics good measures of readability? In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025; Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 24022–24038.
28. Saggion, H. Artificial Intelligence and Natural Language Processing for Easy-to-Read Texts. *J. Lang. Law* **2024**, *82*, 84–103. [CrossRef]
29. Guo, Y.; August, T.; Leroy, G.; Cohen, T.A.; Wang, L.L. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; Volume 2024, pp. 9194–9211.
30. Gao, M.; Ruan, J.; Sun, R.; Yin, X.; Yang, S.; Wan, X. Human-like Summarization Evaluation with ChatGPT. *arXiv* **2023**, arXiv:2304.02554. [CrossRef]
31. Stodden, R.; Momen, O.; Kallmeyer, L. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. *arXiv* **2023**, arXiv:2305.18939. [CrossRef]
32. Grabar, N.; Saggion, H. Evaluation of automatic text simplification: Where are we now, where should we go from here. In Proceedings of the Traitement Automatique des Langues Naturelles, ATALA, Avignon, France, 27 June–1 July 2022; pp. 453–463.
33. Patil, U.; Calvillo, J.; Lago, S.; Schumann, A.K. Quantifying word complexity for Leichte Sprache: A computational metric and its psycholinguistic validation. In Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI & EL/PL), Geneva, Switzerland, 23 June 2025; pp. 94–107.
34. Bredel, U.; Maaß, C. *Leichte Sprache: Theoretische Grundlagen? Orientierung für die Praxis*; Duden: Berlin/Mannheim, Germany, 2016.
35. Anschütz, M.; Oehms, J.; Wimmer, T.; Jezierski, B.; Groh, G. Language models for German text simplification: Overcoming parallel data scarcity through style-specific pre-training. *arXiv* **2023**, arXiv:2305.12908. [CrossRef]
36. Elmakias, I.; Vilenchik, D. An oblivious approach to machine translation quality estimation. *Mathematics* **2021**, *9*, 2090. [CrossRef]
37. Deilen, S.; Garrido, S.H.; Lapshinova-Koltunski, E.; Maaß, C. Using ChatGPT as a CAT tool in Easy Language translation. *arXiv* **2023**, arXiv:2308.11563. [CrossRef]
38. Ahrens, S.; Deilen, S.; Garrido, S.H.; Lapshinova-Koltunski, E.; Maaß, C. Evaluation of Machine Translation Errors in German Plain Language Texts in the Domain of Health Information. In Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops, Hildesheim, Germany, 9–12 September 2025; Wartena, C., Heid, U., Eds.; HsH Applied Academics: Hannover, Germany, 2025; pp. 176–185.
39. Ahrens, S.; Deilen, S.; Lapshinova-Koltunski, E.; Garrido, S.H.; Maaß, C. Evaluation of translations into plain german produced by humans and mt systems including chatgpt. *SKASE J. Transl. Interpret.* **2025**, *18*, 38–54.
40. Hansen-Schirra, S.; Nitzke, J.; Gutermuth, S. Language (Geasy Corpus): What Sentence Alignments Can Tell Us About Translation Strategies in Intralingual. In *New Perspectives on Corpus Translation Studies*; Springer: Singapore, 2021; p. 281.
41. Deilen, S.; Lapshinova-Koltunski, E.; Garrido, S.H.; Maaß, C.; Hörner, J.; Theel, V.; Ziemer, S. 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) at LREC-COLING 2024, Torino, Italy, 20 May 2024; pp. 44–53.
42. Kuckartz, U. *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*; Beltz Juventa: Weinheim, Germany, 2018.

43. Lu, Q.; Qiu, B.; Ding, L.; Zhang, K.; Kocmi, T.; Tao, D. Error analysis prompting enables human-like translation evaluation in large language models. *arXiv* **2023**, arXiv:2303.13809.
44. Fernandes, P.; Deutsch, D.; Finkelstein, M.; Riley, P.; Martins, A.F.; Neubig, G.; Garg, A.; Clark, J.H.; Freitag, M.; Firat, O. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv* **2023**, arXiv:2308.07286. [CrossRef]
45. Kocmi, T.; Federmann, C. GEMBA-MQM: Detecting translation quality error spans with GPT-4. *arXiv* **2023**, arXiv:2310.13988. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.