trans-kom ISSN 1867-4844 https://www.trans-kom.eu
trans-kom ist eine wissenschaftliche Zeitschrift für Translation und Fachkommunikation.

trans-kom 18 [2] (2025): 656–683
Seite 656

Silvana Deilen & Christine Paasch-Kaiser & Ralph Krüger

Artificial intelligence in intra- and interlingual translation

Investigating the potential of the large language model GPT-4o in Plain Language and low-resource translation contexts

Abstract

This study sets out to investigate the potential of the large language model (LLM) GPT-4o in Plain Language and low-resource translation contexts. We do so in the wider context of an artificial intelligence (AI) literacy framework for translation, interpreting and specialised communication, which attempts to model a comprehensive set of competences required in light of new high-performing AI technologies such as LLMs. For each of the two use cases, we explore different prompting strategies and compare GPT-4o's performance with the performance of a dedicated expert machine translation (MT) system. For intralingual translation, we are unable to conclude that prompting led to a significantly better output. In addition, our analyses showed that the dedicated expert MT system, which was specifically developed, trained and fine-tuned for Plain Language translations, outperformed GPT-4o in all criteria. In terms of low-resource translation, our results point in the opposite direction. While specifically highlighting diasystematically marked elements in the prompt does not seem to have a significant impact on translation, providing GPT-4o with a detailed task description and examples significantly improved the output compared to non domain-specific MT.

1 Introduction

Large language models are recent language-oriented artificial intelligence technologies based on the transformer neural network architecture, which was originally invented in the context of neural machine translation (NMT) in 2017. LLMs are trained on large digital corpora including webpages (e. g., the *CommonCrawl* internet corpus), books (e. g., the *BookCorpus* and the *Project Gutenberg* corpus), traditional multilingual MT training corpora (e. g., the *Europarl* corpus), academic/scientific publications (e. g., *arXiv* papers), conversational texts (e. g., *Reddit* discussion threads) and, in some cases, computer code (e. g., from *GitHub* repositories) (cf. Zhao et al. 2023: 11–12). During model training, individual tokens in these training data are masked and the LLM is tasked to correctly predict these masked tokens. In this process, the models acquire a vast amount of both linguistic and world knowledge (*Massively Multi-Task Learning*, cf. Wei 2024). Models exceeding a certain size (expressed in terms of their trainable parameters) exhibit a behaviour called *in-context learning*, which means that the models can recognize and perform a wide range of tasks through natural language task descriptions and/or demonstrations (prompts) even though their original training objective was merely the prediction of masked tokens. The vast amount of linguistic and world knowledge acquired through massively multi-task learning coupled with their in-context learning ability makes LLMs so-called *general-purpose AI* (GPAI) technologies. According to the recent EU AI Act, a GPAI model "displays significant generality and is capable of competently performing a wide range of distinct tasks" (EU Artificial Intelligence Act Chapter I Article 3 Paragraph 63). For each of these tasks, the LLM output can be conditioned by using task-specific prompting strategies (see section 2.1). In addition to their general-purpose character, this ability to influence their output through adequate prompting sets current LLMs apart from previous natural language processing (NLP) technologies such as MT and at the same time opens up new areas of application together with new avenues of research.

The present article intends to explore two such avenues by investigating the performance of one of OpenAI's flagship models, GPT-4o, in two specific use cases: *intralingual* MT into Plain Language (section 3) and 'traditional' *interlingual* MT between different languages, here focusing on low-resource settings where LLM translation performance has been shown to deteriorate (cf., e. g., Robinson et al. 2023) (section 4). By examining two different use cases, we aim to investigate how GPT-4o performs in more specific translation scenarios, namely, in contexts where translation is intended to foster accessibility, and in contexts where only limited training data are available. This is particularly interesting as the performance of NLP tools in these two scenarios has traditionally been uneven (see the discussion in sections 2.2 and 2.3). For each of the two scenarios, we explore different prompting strategies and compare GPT-4o's performance with the performance of a dedicated expert system to see whether best-practice prompting GPT-4o may result in a performance that is close to, on par with or even better than the performance of such expert systems.

2 Theoretical foundations

This section lays the theoretical groundwork for the analysis. It first discusses the concept of translation-specific AI literacy, which provides the overall theoretical framework for this article. Then, the section zooms in on one central component of AI literacy, i. e., LLM prompting, which will be one important variable in our two use cases. It discusses the concept of prompt engineering and surveys two relevant prompting frameworks. The section then proceeds to discuss the theoretical foundations of our two use cases, intralingual MT into German Plain Language and interlingual low-resource MT focusing from Catalan into German.

2.1 Translation-specific artificial intelligence literacy and LLM prompting

Translation has evolved over the last decades through ongoing processes of digitalisation (i. e., the development of new and increasingly powerful translation technologies) and datafication (i. e., the accumulation and provisioning of large volumes of translation data, such as translation memories). Recent successes in the AI subfield of machine learning (ML) have led to the emergence of high-performing language-oriented AI technologies such as NMT and LLMs, which have been used to increase the degree of automation in translation beyond what was previously possible. This increasing degree of translation automation fueled by language-oriented AI requires new digital literacies on the part of relevant stakeholders in both professional and academic settings, such as data literacy, MT literacy and AI literacy. In this context, Krüger (2024) recently proposed an AI literacy framework for translation, interpreting and specialised communication, which distributes overall domain-specific AI literacy over five dimensions. A reduced version of the framework is depicted in Figure 1 below.

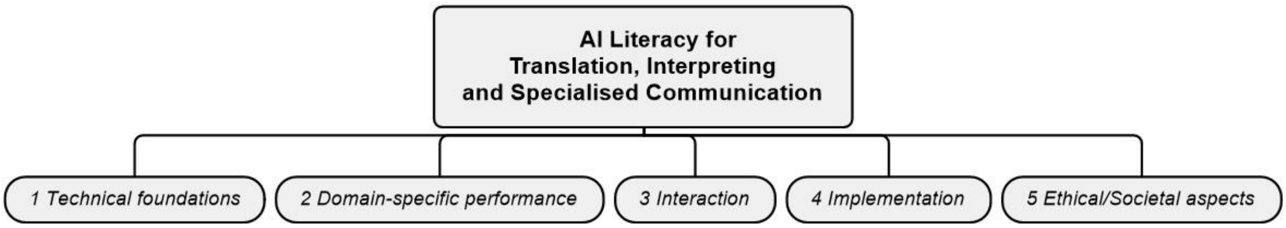

Fig. 1: Reduced version of the AI literacy framework for translation, interpreting, and specialised communication (figure taken from Krüger 2024: 15)

Dimension 1 of the framework refers to the technical foundations of current AI technologies (with a focus on LLMs) and covers aspects such as their operating principle (artificial neural networks, transformer/self-attention, embeddings, etc.), training/fine-tuning AI models, etc. Dimension 2 is concerned with the domain-specific performance of LLMs and covers aspects such as determining their current and future scope of capabilities and their task-specific performance level, identifying the added-value that human experts continue to contribute to AI-assisted tasks or processes, etc. Dimension 3 is concerned with the competences required for interacting with LLMs and covers both higher-level aspects of human cognition and agency as well as more practical aspects such as LLM prompting (see the discussion below). Dimension 4 is concerned with implementing LLMs in practical workflows in the language industry and covers aspects such as compliance with legal AI frameworks, AI risk management, process design, etc. And dimension 5 is concerned with wider ethical and societal aspects of current AI technologies and covers aspects such as AI-induced social (dis)empowerment, the risk of epistemic distortion inherent in current AI technologies, etc.

The epistemic aims of the present article are concerned with the task-specific performance level of current LLMs as covered by dimension 2 of the framework, specifically with intralingual Plain Language translation and interlingual low-resource translation. In our two use cases, we investigate different strategies for LLM prompting as part of the *interaction* dimension of the framework. At several points in our use cases, we also highlight the human added-value that needs to be contributed in intra- and interlingual AI-assisted translation, which is also subsumed under dimension 2 of the framework.

According to White et al. (2023: 1, boldface removed), a prompt is "a set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities."[1] Since the quality of a prompt may have a significant impact on model performance, prompts should not be created ad hoc but crafted according to specific criteria. This process is called *prompt engineering*, which Zhou et al. (2023: 1) define as "optimizing the language in a prompt in order to elicit the best possible [model] performance." In recent years, various prompting frameworks have been developed, which provide criteria, patterns, and templates for structured prompt engineering. We focus on two of these:

The first framework is called *CO-STAR* (cf. Teo 2023), which structures a prompt into six elements: (1) ***C****ontext* (provides background information on the task), (2) ***O****bjective* (defines the task that the LLM should perform), (3) ***S****tyle* (specifies the writing style the LLM should use), (4) ***T****one* (defines the attitude of the LLM's response), (5) ***A****udience* (stipulates the target audience of the LLM's response) and (6) ***R****esponse* (defines the response format). The second, more complex, prompting framework is provided by OpenAI, the developer of the GPT-4o model (cf. OpenAI n. d.). OpenAI defines six strategies for prompt engineering: (1) *Write clear instructions*, (2) *Provide reference text*, (3) *Split complex tasks into simpler subtasks*, (4) *Give the model time to 'think'*, (5) *Use external tools*, and (6) *Test changes systematically*. Each of these strategies is subdivided into different tactics. For example, strategy 1 (which is of particular relevance to our two use cases) is subdivided into the six tactics (1) *Include details in your query*, (2) *Ask the model to adopt a persona*, (3) *Use delimiters to clearly indicate distinct parts of the output*, (4) *Specify the steps required to complete a task*, (5) *Provide examples* (= n-shot prompting) and (6) *Specify the desired length of the output*. Tactic 1 of strategy 6 (*Evaluate model outputs with reference to gold-standard answers*) is also relevant to our two use cases since we compare the LLM output with the output of dedicated expert systems and perform a human expert analysis of this output. The prompts engineered based on these two frameworks are presented and discussed in the methodologies of our two use cases (sections 3.1 and 4.1).

---

1 Within a prompt, a distinction can be made between task *descriptions* (e. g., *Translate this text from Catalan into German*) and task *demonstrations*, where the model is provided with one or more examples of the task to perform. Providing task demonstrations is called *n-shot prompting* (1 example = 1-shot prompting, 2 examples = 2-shot prompting, >1 examples = few-shot prompting). When no task demonstration is included in a prompt, this strategy is called *zero-shot prompting*.

2.2 Intralingual MT into Plain Language

2.2.1 Definition

According to Jakobson, the term *translation* traditionally refers to *inter*lingual translation, which he defines as "an interpretation of verbal signs by means of some other language" (Jakobson 1959: 233). While interlingual translation always involves a change in language, *intra*lingual translation, which Jakobson refers to as *rewording*, is defined as "an interpretation of verbal signs by means of other signs of the same language" (Jakobson 1959: 233). However, according to Jakobson, intralingual translation (i. e., translating between different varieties of one language) does not belong to classic, prototypical translation or "translation proper" (Jakobson 1959: 233). It was only with the emergence of *skopos*-oriented and functionalist approaches (Reiß/Vermeer 1984) that the concept of translation was expanded to also include *intralingual* forms of translation. This broader "reconceptualization of translation" (Hansen-Schirra/Maaß 2019: 2), which is now firmly established in translation studies, has also led to significant changes in the translator's professional profile and required competencies. Today, translators are no longer only required to overcome barriers between different languages (and cultures), they must also be capable of overcoming barriers within a single language, for example, those arising from different language varieties.

In Germany, there are two well-established language varieties that belong to the field of intralingual translation: *Easy Language* (EL) and *Plain Language* (PL). As EL and PL are clearly defined and distinct language varieties, we will briefly explain the difference between these two concepts. Both EL and PL are simplified, complexity-reduced varieties of German that aim to enhance the readability and comprehensibility of texts and to make standard and specialized texts more accessible (Maaß 2020). However, there are significant differences in terms of their complexity and target groups.

While EL is characterized by a maximally reduced complexity at all language levels and mainly addresses people with communication impairments, the syntactic and lexical features of PL are much more complex. Also, PL is mainly aimed at laypersons or people with limited prior knowledge in a specific field and is primarily used to provide access to expert communication (Maaß 2020). Another PL target group is people with German as a second language, with the language level required to understand a PL text typically set at B2 (cf. Maaß/Hérnandez Garrido forthc.). Another difference between these two varieties is that EL has a fixed set of rules, whereas PL is a dynamic, flexible variety that does not have a fixed set of rules but can be adapted to the needs and reading skills of the intended audience. The advantage of PL is that it is less stigmatizing and more highly regarded in society than EL. However, PL texts are often too complex and therefore not accessible for people with reading skills below B2.

In German legislation, "Easy Language currently prevails over Plain Language as the means for communicative accessibility" (Maaß 2020: 61) (for an overview of the legal situation of EL and PL in Germany see Maaß 2020: 56–64). Since 2014, EL has become a subject of academic research in Germany, with an increasing number of publications in the following years. Research in this field has taken two main directions: one focusing on text quality and potential barriers in different forms of communication (e. g., Rink 2019), and the other examining comprehensibility among various target groups (e. g., Gutermuth 2020; Deilen 2021). However, PL has gained increased recognition in recent years as well, with more and more initiatives being implemented that propose PL as a preferred means of communication (Schaeffer et al. 2018: 39; Maaß 2024). One prominent example of using PL to make information more widely accessible is the health magazine *Apotheken Umschau*, which regularly publishes health-related articles in PL, or the television news service *Tagesschau* in PL, which airs Monday to Friday at 7 p.m. CET.

However, despite the legal requirement to publish accessible information, a closer look at current text practices reveals that many public authorities still do not provide sufficient information in EL; rather, the offer is often reduced to a bare minimum. One of the reasons for the lack of accessible texts is that translating a text into PL or EL is costly and time-consuming and requires professional translators, for which public authorities often do not have the necessary resources. Also, many content providers are unaware that producing high-quality, functional translations requires professional translators. As a consequence, translations are often carried out by individuals without adequate training in intralingual translation, such as regular employees of public authorities or social service agencies (Maaß et al. 2014: 54; Hansen-Schirra et al. 2020: 121).

2.2.2 AI tools for intralingual translation

High-quality intralingual translations are time-consuming for several reasons. For example, contrary to the field of interlingual translation, there are, so far, no well-established CAT Tools in the intralingual translation workflow such as translation memory systems. However, given the legal requirements and the overwhelming volume of texts to be translated, it is evident that tools that have become an indispensable component in the interlingual translation process (including MT systems) are also urgently needed in intralingual translation.

In the last few years, various AI tools have been launched that offer MT into EL or PL. One of these tools is provided by SUMM AI, which initially only offered translations into German EL, but now also offers a tool for translating into German PL. There are various other text simplification tools, such as those from *capito.ai* or *T2K (text2knowledge)*. These tools, whose user interfaces are similar to those of typical interlingual MT systems like *DeepL*, must be distinguished from analysis tools such as *Wortliga* or *TextLab*. In addition to an MT function, these analysis tools offer text analyses as well as editing and revision features, such as rephrasing or shortening texts, suggesting improvements or synonyms. While the tool *Wortliga* also includes an analysis of the text's language level (based on the Common European Framework of Reference for Languages from A1 to C2), the tool *TextLab* calculates the Hohenheim Comprehensibility Index (HIX), which rates the comprehensibility of a text on a scale from 0 to 20. According to Rink (2019: 77) and Deilen et al. (2024a, b), EL texts should have an

HIX of at least 18 points whereas the predefined benchmark for PL texts is set at 14 points.

The general-purpose character of LLMs as discussed in section 1 means that the above-mentioned options of using AI for intralingual translation can be expanded further. For example, by prompting a LLM, it is now possible to adapt a text to (customer-)specific requirements, to automatically add explanations, or to shorten the text to a specific word, sentence, or text length, thereby creating a 'tailored' text. There is no doubt that MT is a promising technology in the field of intralingual translation. However, we still do not know whether the output really meets the predefined criteria for EL and PL texts – a question we explore in more detail in use case 1 in section 3.

2.2.3 Limitation of AI tools for intralingual translation

In the meantime, first empirical studies have been published in which the quality of MT systems for intralingual translation was evaluated and whose initial results have shown that the existing skepticism is justified. In the following, we summarize these first results, which we will then take as a starting point for an exemplary analysis (see Section 3.1).

As EL texts follow specific rules at word, sentence and text levels, EL is often regarded as a controlled language which, at least in theory, should be amenable to automation. This assumption was tested by Deilen et al. (2023). Their study, in which they translated administrative texts into EL using GPT-3.5, showed that the generated target texts (TTs) were easier than the source texts (STs) in standard German, but that the texts did not meet the EL standards since they were not rule-consistent and still too complex to be classified as EL texts. As the study also revealed that most of the TTs contained at least one serious mistake, the authors concluded that GPT-3.5 can be used as an MT tool for professional translators but that the texts cannot be used safely by the primary target groups. This was also confirmed by Anschütz et al. (2023), who found that models can be trained to adapt to the style of EL but that the output so far cannot be used by the target groups.

Deilen et al. (2024a,b) evaluated the quality of PL translations produced with different versions of the MT system SUMM AI. Using texts from the *Apotheken Umschau* corpus, their study showed that especially the model that was trained and fine-tuned with specific training data from the *Apotheken Umschau*, as well as with human gold standard translations from the same text type, showed promising results in terms of syntactic complexity and comprehensibility. However, here as well, the evaluation of the content revealed serious mistakes, which is why they concluded that, as of now, these tools cannot be used safely by primary target groups.

Ahrens et al. (forthc.) also automatically translated the same STs from the *Apotheken Umschau* corpus into PL using GPT-4o. Their first results showed that the SUMM AI models outperformed GPT-4o, suggesting that a model that is fine-tuned with rules and domain-specific data and specifically trained for intralingual translation performs better than a general-purpose LLM. When testing GPT-4o, Ahrens et al. (forthc.) only employed one prompting strategy. More specifically, they used a zero-shot strategy with a detailed task description. However, as it has been shown that the generated output is highly dependent on the prompting strategy (cf. section 2.1) and that few-shot prompting tends to produce better outputs, the question arises whether different prompting strategies can raise the performance of GPT-4o to the level of a fine-tuned model that was specifically trained for intralingual MT. This question will be tentatively answered in section 3.

2.3 MT and low-resource languages

2.3.1 Low-resource languages

The terms *high-resource language* (HRL) and (very or extremely) *low-resource language* (LRL) are employed to distinguish different types of languages in the context of NLP and NMT. Languages such as English or German are considered to be HRLs (cf. Ranathunga et al. 2021: 2). Big, labeled datasets are available in these languages, covering a wide range of tasks and domains. Usually, large-scale private and public investments in HRL are made in order to expand digital resources and develop language technologies (cf. Joshi et al. 2020: 6284). For many other languages, in contrast, no or only little digital or digitally accessible data for NMT and AI training are available. This lack often results in a poor performance of neural models on LRL (cf. Robinson et al. 2023: 392).

To identify (very or extremely) LRL, the presence of labeled datasets in various repositories has been analyzed. Further, unlabeled datasets such as Wikipedia pages have been taken as indicators. By these means, Joshi et al. (2020: 6284–6285) identified six different language types, ranking from *0 – The Left-Behinds* to *5 – The Winners*.[2] While the authors classify German among the winners, they do not mention Catalan in their study.

Catalan and German clearly do not belong to the same category when comparing the amount of resources stored in repositories such as LDC (2024). They include far more resources for German than for Catalan. This also applies to unlabeled datasets: the number of Wikipedia pages currently available in Catalan amounts to only one-third of those in German Wikipedia 2024a,b). Accordingly, Catalan should be classified as type *4 – The Underdogs.* However, such a characterization is not that obvious because the Catalan communication community includes “NLP communities researching [this language]” and has a more or less “large amount of unlabeled data” as well as a “lesser, but still a significant amount of labeled data” (cf. Ranathunga et al. 2021: 3). Furthermore, it comprises many linguistically well-trained speakers, necessary for the production of suitable NLP (training) data. However, the crowd-sourcing initiative *Parla’m* (Òmnium Cultural 2024) aims at further improving the availability of oral and written data to build large text and voice databases for the implementation of Catalan in new AI technologies and devices, since there is still not sufficient training data (at least, certain domains are

[2] 0 – The Left-Behinds, 1 – The Scraping-Bys, 2 – The Hopefuls, 3 – The Rising Stars, 4 – The Underdogs, 5 – The Winners (Joshi et al. 2020: 6284).

still low-resource). The necessity of such an initiative shows that Catalan should probably be classified in the middle or more towards the bottom of type 4. In sum, Catalan still has to be considered an LRL, which quite often has to resort to pivot languages (especially English) in order to compensate for a lack of data.

2.3.2 MT for the language pair Catalan–German

Only a subset of the publicly available MT services provides translation from Catalan into German (*GT*, *MSBing Translator*, *Yandex*, *ModernMT*, *Babelfish*, *Softtraductor*). These are mainly generic (i. e., not domain-specific) NMT systems. In a very small number of qualitative studies, the systems tested to date (*GT*, *MSBing Translator*, *ModernMT*, *Softtraductor*) show severe quality deficits. In terms of the Multidimensional Quality Metrics (MQM), an established and widely-used framework for error analysis in interlingual translation (The MQM Council 2024), these deficits mostly pertain to the categories of *Accuracy*, *Linguistic Conventions* and *Style* (cf. Paasch-Kaiser/Hoberg 2020; Paasch-Kaiser/Hoberg submitted). While Paasch-Kaiser and Hoberg (2020) focus on the use of MT in formal written language, the study by Paasch-Kaiser and Hoberg (submitted) focuses on the MT of pragmatic aspects of informal communication, i. e., diaphasically marked linguistic elements characteristic of colloquial Catalan, informal language use and Computer-mediated Communication (CMC). Their study is based on a corpus of YouTube comments. Among the analyzed elements are lexemes such as Cat. *shipeig* ‘merging of two names’, phraseological units, e. g., Cat. *mengar-se la boca* ‘to kiss like devouring each other’, or vocatives, used as interjection, such as Cat. *tio* ‘bro’. The authors show that the MT of these elements is often of poor quality. The translated comments contain severe grammatical and semantic errors that (often) overshadow an (accurate) translation of pragmatic meaning. Frequently, MT is performed via a pivot language (English or Spanish), which sometimes even leads to mixed-up sentences. Such a pivot translation can bring about various problems, e. g., a mistranslation of words if a Catalan lexeme is categorized as a Spanish one, as in the case of Cat. *tio* ‘bro’, borrowed from Spanish (where *tío* —written with an accent— means ‘uncle’ or, in colloquial Spanish, ‘bro, man’). This was translated by MS Bing Translator as Ger. *Onkel* (‘uncle‘) although the Catalan word for *uncle* is *tiet* (1a). The other tools investigated, however, took Cat. *tio* as part of a proper name (1b):

(1) Es que el Joan es copia de vegetta *tio*, es que no para el cabró.

(1a) Es ist so, dass John sich von Vegetta *Onkel* kopiert, es ist, dass er den Bastard nicht aufhält. (MS Bing Translator)

(1b) Es liegt daran, dass Joan Vegeta *Tio* kopiert, der Bastard hört einfach nicht damit auf. (GT)

In general, MT from Catalan to German still lacks (broad) studies evaluating its quality. Studies using GPAI models (such as GPT-4o) for translation from Catalan to German

trans-kom 18 [2] (2025): 656–683

are also still a desideratum.[3] A translation of YouTube comments from colloquial Catalan into German is especially interesting for learners of Catalan using the comments to improve their language skills. These learners may include future translators and interpreters who need to be aware of the markedness of forms and phraseological units. The ingenuous use of diaphasically heavily marked units in an improper way, even in chat communication, could be (considered) a face-threatening act. In section 4, we therefore conduct an exploratory analysis of GPT-4o for the translation of diasystematically marked elements of Catalan from a YouTube corpus.

In the following section, we turn to our two research questions: Firstly, we aim to investigate whether different prompting strategies can bring GPT-4o's performance close to that of a fine-tuned model specifically trained for Plain Language translation and secondly, we want to explore whether GPT-4o performs better in terms of translating diaphasically marked elements of Catalan into German than GT.

3 Use case 1: Intralingual Plain Language MT

3.1 Methodology

In the following, we will tentatively answer the question of whether different prompting strategies can raise GPT-4o's performance to the level of a fine-tuned model that was trained for intralingual translation. For our analysis, we apply three prompting strategies to one text. We would like to emphasize that this is by no means a comprehensive empirical study, but merely a small-scale exploratory analysis. For conclusive and representative results, a larger-scale study is required. Still, since there are currently no results or studies addressing this question, our study can be considered a preliminary starting point. For our test, we chose a ST and the corresponding translations from the corpus presented in Deilen et al. (2024a, b). More specifically, we chose the text about skin cancer, for which we had a version in standard German (ST) as well as a human gold standard translation (produced by professional translators) and the machine translation produced with SUMM AI. Deilen et al. (2024a, b) had different translations created with different models from SUMM AI, from which we chose the translation by the model that performed best and did not contain any content-related errors. We defined the following three different prompts,[4] which adhere to the prompting strategies discussed in section 2.1 (see Table 1).

---

[3] To our knowledge, only one (unpublished) Bachelor's thesis has been written to date which compares MT of marketing texts from German into Catalan with GT and GPT-4o. The study shows that the GPT model frequently provides more convincing translation solutions than GT.

[4] In both use cases, GPT-4o's cache was deleted after each prompting session. This was done to avoid artifacts of a previous prompting session potentially impacting the output of a subsequent session.

trans-kom 18 [2] (2025): 656–683

Prompt type	Prompt
PL1: Zero-Shot	Übersetze den Text in Einfache Sprache:[5] *'Translate the text into Plain Language:'*
PL2: Zero-Shot with detailed task description (context, objective, and audience)	Du bist ein Einfache-Sprache-Übersetzer. Bitte übersetze den folgenden Text in Einfache Sprache. Der Text soll in der Online-Ausgabe der *Apotheken Umschau* erscheinen. Die Übersetzung soll für Menschen mit Leseschwierigkeiten leicht verständlich sein: 'You are a translator for Plain Language. Please translate the following text into Plain Language. The translation will be published in the online edition of the Apotheken Umschau. The translation should be easy to understand for people with reading difficulties:'
PL3: 2-Shot (with detailed task description, context, objective, and audience + two examples)	Du bist ein Einfache-Sprache-Übersetzer. Bitte übersetze den folgenden Text in Einfache Sprache. Der Text soll in der Online-Ausgabe der Apotheken Umschau erscheinen. Die Übersetzung soll für Menschen mit Leseschwierigkeiten leicht verständlich sein. Orientiere dich bei der Übersetzung an folgenden Beispielen: […] Hier ist der zu übersetzende Text: [….] 'You are a translator for Plain Language. Please translate the following text into Plain Language. The translation will be published in the online edition of the Apotheken Umschau. The translation should be easy to understand for people with reading difficulties. When translating, consider the following examples […] This is the text to be translated: […]'[6]

Table 1: Prompts used for the PL translation use case

In the next step, we translated the ST with GPT-4o using the three prompts listed in Table 1. To evaluate the quality of the texts, we used the same criteria as Deilen et al. (2023, 2024a,b). More specifically, we first calculated the readability of the texts by using the HIX (see section 2.2.2). Secondly, we automatically identified complex syntactic relations using dependency parsing that we obtained by using the R Package UDPipe.

5 As we focus on translating into German, the prompts used in the two use cases were also written in German. Even though the OpenAI models are trained predominantly on English text (cf., e. g., Krüger 2023: 296), model performance in other high-resource languages (French, German, Spanish, etc.) is also generally high, making prompting in these languages feasible.

6 The two examples are provided in the online Appendix.

The list of analyzed relations included *acl* (adnominal clause or clausal modifier of noun), *advcl* (adverbial clause modifier), *ccomp* (clausal component), *csubj* (clausal subject), *xcomp* (open clausal element) and *parataxis* (parataxis relation), which are all categorized as Universal Dependency Relations (n. d.; see de Marneffe et al. 2021 for more details). We assume that the higher the number of these dependency relations, the more complex the text. In the last step, we manually analyzed the correctness of the content. To assess the errors in a more structured way, we based our error analysis on the MQM framework (see section 2.3.2). Rodríguez Vázquez et al. (2022) and Ahrens et al. (forthc.) also use this framework for evaluating intralingual MT. However, Ahrens et al. (forthc.) suggest only focusing on four main error categories and their respective error types as these are the error categories and types that are most relevant in intralingual translation. We follow this approach and assign the errors to one of the following four categories: “Terminology”, “Accuracy”, “Linguistic Conventions”, or “Audience Appropriateness” (for a detailed overview of the error categories and error types see The MQM Council 2024).

3.2 Results

Table 2 provides an overview of the results, which will now be discussed in more detail. The translations generated by GPT-4o are provided in the appendix.

	ST	**Human translation (HT)**	**SUMM AI**	**PL1**	**PL2**	**PL3**
Characters	10344	5545	3492	1490	1769	1775
Words	1667	958	639	256	297	302
Language level	C1	A1	A1	B1	A2	B1
HIX	11.24	19.13	19.73	18.21	17.78	19.22
csubj	1	0	0	0	0	0
ccomp	6	0	0	0	1	0
xcomp	13	2	2	2	1	5
advcl	9	0	0	4	8	3
acl	24	0	1	3	2	1
parataxis	12	10	9	0	1	3
Correctness of the content	yes	yes	yes	yes	no	yes

Table 2: Overview of the calculated metrics

Table 2 shows that all of the three GPT-4o translations (PL1, PL2, PL3) are significantly shorter than the HT and the SUMM AI translation. Even the translation produced with the 2-shot prompt (PL3), which is the longest of the GPT-4o translations (302 words), is only half as long as the SUMM AI translation (639 words) and only one-third as long as the HT (958 words). The analysis of the language proficiency level shows that the HT and the SUMM AI translations are classified as A1 texts, whereas the GPT-4o texts are classified as A2 (PL2) or B1 level (PL1, PL3). Comparing the readability of the texts reveals that the SUMM AI text had the highest HIX (19.73), closely followed by the 2-shot translation PL3 (19.22) and the HT (19.13). The zero-shot translation PL1 was the least comprehensible (17.78). However, given that the language proficiency level required to understand a PL text is set at B2 and PL texts should have an HIX of at least 14 points, all of the GPT-4o translations met these criteria.

Analyzing the distribution of dependency relations revealed that, unlike the SUMM AI and the HT, the GPT-4o translations contain many clauses modifying verbs and adjectives (*advcl*). This is true for all prompts, with the PL2 text being the most complex and the PL3 text the least complex in terms of clauses modifying verbs and adjectives. Altogether, the other syntactic relations show that in the HT and the SUMM AI translations, parataxes prevail whereas the GPT-4o texts make more use of subjectless clausal complements and clauses modifying nouns (*acl*), which are mainly absent in the HT and SUMM AI translations. The high number of parataxes in the HT and SUMM AI translations is mainly due to the fact that these independent clauses are relatively easy to process and therefore frequently used in PL texts (Ahrens et al. forthc.).

Looking at the GPT-4o translations, we can conclude that they are syntactically more complex than the HT and the SUMM AI translations but that the texts produced via 2-shot prompting (PL3) and zero-shot prompting without task description (PL1) contain less complex clauses than the text produced via zero-shot prompting with task description (PL2). Given that the PL3 text has more words and sentences than the PL1 text, we can conclude that, in terms of syntactic complexity, 2-shot prompting provides the best results.

In terms of the correctness of the content, we only found mistakes in the PL2 translation. Applying the MQM framework to the errors resulted in three mistakes from the error category “Accuracy” and one from the error category “Linguistic Conventions” (see Table 3).

Error Category	Error Type	Error	Explanation
Accuracy	**Wrong explanation**	Meistens ist zu viel Sonne Schuld. **Deswegen** nennt man es auch “Schwarzen Hautkrebs”. ‘Most of the time, too much sun is to blame. That is why it is also called black skin cancer.’	The adverb “Deswegen” (“therefore/that’s why”) is not used correctly because in this case the causal connection between these two sentences does not make sense. The causal adverb implies that the name of the disease results from the cause (the sun), which is not correct. Instead, the name “black skin cancer” comes from the dark color of the cancer on the skin.
Linguistic Conventions	**Grammar**	In Deutschland bekommen jedes Jahr 19 von 100.000 Menschen **Schwarzer Hautkrebs**. ‘In Germany, 19 out of 100,000 people get black skin cancer every year.’	This sentence contains a grammatical mistake because the phrase “Schwarzer Hautkrebs“ must appear in the accusative case, which would be “Schwarzen”.
Accuracy	**Ambiguous content**	Mehr Menschen haben diese Krankheit, weil wir heute mehr Sonne **bekommen**. ‘More people have this disease because nowadays we get more sun.’	In this context, the German verb “bekommen” (‘get/ receive’) is not appropriate and does not fit idiomatically with “sun”, because “mehr Sonne bekommen” (‘getting more sun’) refers to being exposed to (increased) sunlight, not receiving it as one would receive an object or gift. The appropriate translation would be “mehr Sonnenlicht ausgesetzt sind” (‘being exposed to increased sunlight’)

Error Category	Error Type	Error	Explanation
Accuracy	**Mistranslation**	Die beste Behandlung ist das **Herausschneiden des Tumors**. […] Ist der Hautkrebs weiter fortgeschritten und hat sich ausgebreitet, braucht man eine **andere** Therapie. **Dazu** gehören **Operationen**, Bestrahlungen oder Medikamente. ‘The best treatment is cutting out the tumor. […] If the skin cancer has progressed further and has spread, a different therapy is needed. This includes surgeries, radiation therapy or medication.’	The word “dazu” (“this”) suggests that surgeries are something different compared to the method of cutting out the tumor mentioned earlier and implies that cutting out the tumor is no surgery. This is misleading, because cutting out the tumor definitely is a surgical procedure.

Table 3: Correctness of content in the PL2 translation

Even though the PL1 and PL3 texts do not contain any errors that are immediately obvious at the textual surface, comparing their content with the ST reveals that in all of the GPT-4o translations essential ST information is missing, which, given the significantly shorter text length compared to the HT and SUMM AI translations, comes as no surprise. However, as in PL translation, it is up to the translator to decide what information they consider relevant for the reader, and since no such alignment and comparison was done for the HT and the SUMM AI translations either, we refrain from adding missing information as mistakes. However, given the short text length, the GPT-4o translations look more like a summary than a complete translation.

Summarizing the results of all evaluation criteria, we may conclude that the worst output was achieved with the second prompt. This was true for all three evaluation criteria. The first and third prompt performed equally well in terms of correctness and syntactic complexity, with the third prompt outperforming the first one in terms of readability. What is also interesting to note is that in the 2-shot prompt (PL3), we see some improvements in the register, more specifically in the way the reader is addressed. While in the first and second prompt the reader was addressed using the informal form “Du”, in the PL3 translation the model followed the example provided and adjusted the form of address accordingly, using the formal “Sie“.

However, in terms of syntactic constructions, the examples provided in the 2-shot prompt were not followed. This was especially obvious for conditional clauses which, in the examples, were translated with a question followed by a second sentence starting with the adverb “dann” (“then”) (see [1]).

[1] Standard German: Wenn der Blutfluss nicht schnell wiederhergestellt werden kann, sind die Herzmuskelzellen im Versorgungsgebiet der verschlossenen Arterie nach spätestens zwei bis vier Stunden abgestorben.

‘If blood flow cannot be restored quickly, the heart muscle cells in the area supplied by the blocked artery will die within two to four hours at the latest.’

[1] PL: Wird die Verstopfung im Blutgefäß nicht gelöst? Dann sterben Muskelzellen im Herz ab.

‘If the blockage in the blood vessel is not unclogged, muscle cells in the heart will die.’

This was not implemented in the GPT-4o translations, which instead used the syntactically (more) complex hypotaxes (see [2]).

[2] PL: Wenn Sie Veränderungen bemerken, gehen Sie bitte zum Arzt.

‘If you notice any changes, please see a doctor.’

Based on these tentative results, we are unable to conclude that prompting led to a significantly better output.

Before proceeding to use case 2, we briefly want to discuss some other major problems that users have to deal with when using GPT-4o for intralingual translation tasks, such as context dependency or hallucinations.

LLMs generate text by using linguistic context. When using ambiguous words, this can sometimes cause misunderstandings. While humans typically use the global context and the situation in which a sentence is uttered to correctly disambiguate the sentence, LLMs do not have access to this global context and therefore require sufficient linguistic contextual information to disambiguate words correctly (cf. also LR1 Scenario 3 of use case 2). In these cases, it is especially crucial to check whether the LLMs chose the right meaning.

When we humans read a text, we also quickly recognize what information is important and what is less relevant. However, in an LLM, every proposition is initially treated as equally important. As a consequence, the model sometimes struggles to extract important information or to weight the information appropriately, which may result in less important details being overemphasized. This highlights both the need for and the added value of human experts, as they can critically assess and refine the model’s outputs, e. g., by distinguishing between essential and non-essential information (see section 2.1).

One of the major issues when evaluating the correctness of the content are hallucinations. Hallucinations are instances where the LLM generates information that deviates from the input and is factually incorrect. For lay people without sufficient expert knowledge, it is hard to determine which information is made up by the model. Hallucinations occur, for instance, when the LLM does not have sufficient training data for a specific task, field, or language or when the data with which the LLM is trained was aligned incorrectly or incompletely. It goes without saying that taking hallucinations as factual information can have severe consequences.

However, the following should be kept in mind: It is not possible to generalize the result that in our study prompting did not lead to a significantly better output, because GPT-4o is non-deterministic and conducting the same study again using exactly the same text and prompts may lead to a completely different output. For empirical studies, this non-determinism also makes it more difficult to replicate results.

4 Use case 2: Interlingual low-resource MT

4.1 Methodology

In use case 2, we conduct an exploratory analysis of GPT-4o for translating diaphasically marked elements of Catalan that are present in a corpus of YouTube comments collected in 2023 (cf. Paasch-Kaiser/Hoberg submitted). We aim to investigate whether GPT-4o produces better results with regard to the translation of these aspects than GT for Catalan-German, as LLM prompting permits a more fine-grained approach to this task. For the experiment, we chose ten comments which are also analyzed by Paasch-Kaiser and Hoberg (submitted). We apply three different prompting strategies (see Table 4) following the best-practice prompting strategies outlined in section 2.1 and 3.1, in order to determine whether these have an impact on GPT-4o's performance. Unlike in use case 1, we use delimiters before and after the diaphasically marked elements, e. g., <<<element of colloquial Catalan>>> or ===element of CMC===. We do so in order to indicate to GPT-4o elements in the comment which are of particular importance in translation, as delimiters help the LLM to interpret a certain part of a prompt as a unit of (special) meaning (cf. Teo 2023). The examples provided in the few-shot prompt vary. They are chosen in order to illustrate to GPT-4o what kind of element (lexical or phraseological unit, or vocative/interjection) it should pay special attention to.

Prompt Type	**Prompt**
LR1: Zero-Shot	Übersetze den folgenden Text aus dem Katalanischen ins Deutsche: "**…**" (*Translate the following text from Catalan into German*)
LR2: Zero-Shot with detailed task description (context, objective, style and delimiters)	Du bist ein Katalanisch-Deutsch-Übersetzer. Bitte übersetze den folgenden Text aus dem Katalanischen ins Deutsche und erhalte dabei im Zieltext das sprachliche Register des Ausgangstextes. Es handelt sich um einen Kommentar zu einem YouTube-Video. Der Kommentar beinhaltet lexikalische oder phraseologische Einheiten aus der katalanischen Umgangssprache (català col·loquial). Diese Einheiten sind mit <<<Umgangssprache>>> markiert. Außerdem enthält der Kommentar Elemente, die typisch für die Computervermittelte Kommunikation (CMC) bzw. Internet-Slang sind. Diese sind mit ===CMC=== markiert. Text: … 'You are a translator for Catalan-German translation. Please translate the following text from Catalan into German preserving the language register of the source text in the target text. The text is a comment published in the comment section under a YouTube video. The comment contains lexical or phraseological units of colloquial Catalan (català col·loquial). These units are marked as follows <<<colloquial>>>. Furthermore, the comment contains elements of computer-mediated communication (CMC) or internet-slang, marked as ===CMC===. Text: …'

Prompt Type	Prompt
LR3: Few-shot (with detailed task description context, objective, style and delimiters + two examples)	Du bist ein Katalanisch-Deutsch-Übersetzer. Bitte übersetze den folgenden Text aus dem Katalanischen ins Deutsche und erhalte dabei im Zieltext das sprachliche Register des Ausgangstextes. Es handelt sich um einen Kommentar zu einem YouTube-Video. Der Kommentar beinhaltet lexikalische oder phraseologische Einheiten aus der katalanischen Umgangssprache (català col·loquial). Diese Einheiten sind mit <<<Umgangssprache>>> markiert. Außerdem enthält der Kommentar Elemente, die typisch für die Computer-vermittelte Kommunikation (CMC) bzw. Internet-Slang sind. Diese sind mit ===CMC=== markiert: Orientiere Dich bei der Übersetzung an folgenden Beispielen: […] Hier ist der zu übersetzende Text: […] ‘You are a translator for Catalan-German translation. Please translate the following text from Catalan into German preserving the language register of the source text in the target text. The text is a comment published in the comment section under a YouTube video. The comment contains lexical or phraseological units of colloquial Catalan (català col·loquial). These units are marked as follows <<<colloquial>>>. Furthermore, the comment contains elements characteristic for computer mediated communication (CMC) or internet-slang. These are marke with ===CMC===. When translating, consider the following examples […] This is the text to be translated: […]’

Table 4: Prompts used for the low-resource translation use case

As for use case 1, it is important to underscore that no definite conclusions can be drawn from this first experiment. However, it might be seen as a starting point for future studies on the use of GPAI models for the translation of texts in informal Catalan into the same register in German, e. g., for literature or video game localisation. In the next step, we will evaluate some results of the GPT-4o translations based on the MQM error typology.

4.2 Results

In this section, we focus on errors of the MQM category *Style_Language register* (St_LR*)* and have a closer look at the category *Accuracy* (Acc) whenever it is relevant with regard to the translation's register. We use the description of several unexpected aspects in responses (referred to as LR1, LR2**,** LR3) related to the prompting as a starting point to discuss some of the results of the GPT-4o translations. The respective error category will be indicated in brackets.

In only three out of 10 LR1 scenarios, GPT-4o provides solely a translation of the comment in its response the same way a standard MT system such as GT would do. In four other scenarios, the translation is preceded by an introductory phrase such as Ger. *Der Text kann ins Deutsche übersetzt werden als:* 'The text can be translated into German as:', followed by the comment's translation in quotation marks. This occurs also in one LR2 response. In three other responses (LR2: 1; LR3: 2), GPT-4o repeats the prompt's indication *Text:* before the translation. It is not obvious what causes or inhibits these introductory phrases; it might be related to the non-determinism of the LLM discussed in section 3.2.

In the three remaining LR1 scenarios, GPT-4o provides a reflection about particularities and problems that arise from the ST (see Table 5). In two of these scenarios (1 and 2), GPT-4o provides a translation of the comment. In scenario 2, it also highlights stylistic aspects, indicating that the Catalan comment contains 'colloquial expressions' (Ger. *umgangssprachliche Ausdrücke*) and 'slang elements' (Ger. *Slang*), even though it does not specify what these elements are, i. e., it refers explicitly to language register. In scenario 3, GPT-4o reflects about atypical or incorrectly written forms, 'typos' (Ger. *Tippfehler*) and 'dialect forms' (Ger. *Dialektformen*) and indicates that it cannot translate the comment without having more context.

	Scenario 1	**Scenario 2**	**Scenario 3**
YouTube comments included in LR1-Prompts	"Que es mengin la boca JA"	"Magrada com dos babies com vosaltrs aneu al 2017 com si dos la puta pre-historia y fa dos dies 😂 😂 😂 jajajaja bebeeeeesss"	"Si us agradeu pos lieuvos"
Response	Der Satz "Que es mengin la boca JA" lässt sich auf Deutsch etwa mit "Dass sie sich sofort küssen" übersetzen. Der Ausdruck kann je nach Kontext leicht variieren, aber diese Übersetzung gibt die grundlegende Bedeutung wieder.	Der Text auf Katalanisch enthält umgangssprach-liche Ausdrücke und Slang. Hier ist die Überset-zung ins Deutsche: "Ich finde es lustig, wie ihr zwei Babys über 2017 redet, als ob es die verdammte Ur-geschichte wäre, und dabei ist es erst zwei Tage her 😂 😂 😂 hahaha Babys."	Der Text "Si us agradeu pos lieuvos" scheint im Katalanischen keine klare Bedeutung zu haben, da einige Wörter wie "agredeu" und "lieuvos" untypisch oder möglicher-weise falsch geschrieben sind. Es könnte sich um Tippfehler oder Dialekt-formen handeln. Möglicher-weise wollten Sie etwas anderes ausdrücken. Könnten Sie den Satz bitte überprüfen oder zusätz-lichen Kontext bereit-stellen?

Table 5: Three YouTube comments included in LR1-prompts and GPT-4o's responses to them

In scenario 1, GPT-4o indicates that its translation renders the *grundlegende Bedeutung* 'fundamental meaning' of the given comment. This holds true for the meaning of the phraseological unit embedded in the comment, *menjar-se la boca*, which is rendered semantically correctly as *küssen* '(to) kiss', but not with regard to the register [St_LR]. The German verb *küssen* is diasystematically unmarked while *menjar-se la boca* is considered as informal (cf. Salvanyà 2020). Even though GPT-4o indicates that the translation could vary in regard to its context, the solution provided does not satisfy our demand (a translation preserving the ST register), whereas it certainly accomplishes what has been asked for in LR1: a translation from Catalan into German. Nevertheless, this translation is far better than the one provided by GT (*Lassen Sie sie BEREITS ihren Mund fressen* 'Let [3. P. Sg. or Pl. polite form] them ALREADY feed their mouth'), which did not recognize the meaning of the phraseological unit at all. Comparing the results of the LR1 with those of LR2 and LR3, they differ from the LR1 response in terms of language register: the LR2 response uses the German verb *knutschen* (informal) 'to snog', whereas the LR3 response renders the phraseological unit as *abschlabbern* 'to lick', used more often for dogs and, in a rather pejorative way, for two humans that do not stop kissing each other fervently. While we consider both to be adequate in terms of language register, the LR3 response conveys an additional connotation not present in the ST [Acc].

In scenario 2, the translation provided in the LR1 response is semantically and grammatically adequate. The typos in the ST were recognized and interpreted correctly (e. g., *com si dos* 'as if two' instead of *com si fos* 'as if it was'). The LR1 response is also adequate in terms of register, rendering the diaphasically marked element *puta* 'damn' with the equally marked German lexeme *verdammt* 'damn' (also in LR2 and LR3). The translations do not use word contractions, a typical feature of informal written language use. Word contraction is used in the ST (e. g. Cat. *Magrada* 'its[7] great' instead of *M'agrada* 'it's great') and would also have been expected and acceptable in a register equivalent, i. e., informal written German TT (e. g. *ich finds* [finde es] *lustig*) [St_LR]. The translation by GT is similar to that of GPT-4o in terms of language register (Cat. *puta* > Ger. *verdammt*, normative punctuation, no contraction). However, GT's translation contains several errors, caused by the typos in the ST, that lead to a semantically incorrect and unintelligible German output.

Word contractions are one of the main problems GPT-4o comes up against in scenario 3. Here, the model does not recognize the colloquial pronominal verb *liar-se* ('to get involved with s. o. in a sexual way', cf. Salvanyà 2005: 34) in the form *lieuvos* (2nd p. pl.). Together with the misinterpretation of the Catalan verb form *agradeu* (2nd p. pl. of *agradar* 'to please (s. o.)'), this makes it impossible for GPT-4o to provide a translation in the case of LR1. In contrast, LR2 (Ger. *<<<macht>>> <<<euch ans Werk* 'start doing the job') and LR3 (Ger. <<<dann>>> <<<haut euch aufs Ohr>>> 'then hit the hay') yielded two different translations—an informal register variant in LR3 and an unmarked variant in LR2 [ST_LR]—, which both differ semantically from the ST content [Acc]. However, one might interpret LR2 as a face-saving informal variant, as it could be understood as an (informal) euphemism, i. e. a disguise for 'start hooking up'.

Finally, we look at the translation of vocatives proper, and vocatives used as interjections. GPT-4o translated these adequately in most cases, e. g., Cat. *home* 'man, bro' was translated as Ger. *Mann* 'man' (LR1) or *Alter* 'dude' (or literally 'old man') (LR2, LR3). By contrast, Cat. *tio*, a vocative used as interjection, was translated as such only in two cases (LR1, LR3). The LR2 response provides an erroneous translation: a determined nominal group Ger. *dieser Typ* 'this guy' [Acc]. GT did not recognize the interjection *tio* either (cf. 2.3.2).

Summing up, we found several shortcomings with regard to *Style* in all three prompting scenarios. In several cases, these are related to other types of errors (i. e., *Accuracy* but also *Linguistic Conventions*, not discussed further in this article). In several cases, the responses to LR2 provide a better translation of the language register (e. g., *knutschen*) than the alternative promptings. The LR2 responses also contain the lowest

[7] We refer to *its* instead of *it's*, i. e., we omitted the apostrophe as in the Catalan case of *magrada* instead of *m'agrada*, which might lead to a confusion with the English genitive of the neutral possessive pronoun.

number of errors in the category of *Accuracy*. Although we described a series of shortcomings in the LR1 responses, these are still globally more adequate regarding the language register than those of GT. Consequently, in this exploratory analysis GPT-4o performed better than the NMT with regard to the translation of diasystematically marked texts into the same language register. Despite this result, users still need to know how and where to cross-check the correctness of the MT output in LRL settings. This kind of research competence is crucial in interlingual translation from a LRL into a HRL since the data set available to the LLM is limited. This is even more important if the ST is written in usually not well-documented informal language and/or contains features of CMC. The largely better result as compared to an expert NMT system is probably due to the possibility of providing contextual information. With respect to the use of delimiters in our prompts, it seems that they helped to improve the translation with respect to the chosen register in some cases (e. g. *knutschen, abschlabbern* instead of *küssen*), but we cannot confirm a significant improvement related to their use in general.

5 Conclusion and outlook

In this article, we investigated the potential of the large language model GPT-4o in an intralingual Plain Language translation setting and in an interlingual low-resource translation setting. The analysis was conducted through the lens of translation-specific artificial intelligence literacy, focusing on the domain-specific performance of GPT-4o and investigating LLM interaction through different prompting strategies. At various points of the article, we also highlighted the human added-value that needs to be involved in AI-assisted translation scenarios.

Concerning use case 1, the analysis clearly underscored the finding from Deilen et al. (2023) that GPT-4o cannot be used as a stand-alone intralingual translation tool and that certain human competencies are necessary to effectively and safely integrate the tool into the workflow and translation process. For the use of AI, the core competencies for intralingual translation defined by Maaß (2020) could be expanded by the domain-specific AI literacy framework discussed in section 2.1. However, more conceptual work may be required in order to fully align the competences defined by Maaß with this framework.

In the intralingual translation process, it is crucial to first decide which tool is best suited for the specific task and language variety (in the AI literacy framework, this is covered by the *implementation* dimension, which is concerned, among other things, with AI model/tool selection). For example, tools like the one from SUMM AI are developed, trained and fine-tuned specifically for PL translations, but do not allow for writing customized prompts or to adapt the text to (customer-)specific requirements. GPAI models like GPT-4o, on the other hand, are not specifically trained for intralingual translation, but they allow users to adapt the output to client-specific guidelines, to shorten the text, etc.

Concerning use case 2, GPT-4o outperformed the expert NMT system GT the translation of diasystematically marked elements in YouTube comments written in informal Catalan, i. e., from a LRL into a HRL. This result is in contrast to the overall poor LRL translation performance of LLMs observed in Robinson et al. (2023). This performance was probably achieved by providing the LLM with a specific task description (cf. table 4). However, the analysis of the results has shown that the assessment of the LLM's output requires, in particular, research, error detection and correction skills combined with a certain level of proficiency in the LRL. This is especially true when users have to evaluate the translation of elements not included in monolingual or bilingual dictionaries, but it is also necessary because the translations exhibited more semantic and pragmatic rather than grammatical shortcomings. GPT-4o should therefore not be used as the only tool for interlingual translation from a LRL into a HRL but needs critical assessment and human verification based on careful research in (online) dictionaries, glossaries, corpora, etc.

In both use cases, it became evident that a basic understanding of how AI tools function is a core competence that is required when working with these tools (cf. dimension 1 of the AI literacy framework). Only if the user knows how and on what data these systems are trained (HRL vs. LRL, etc.), how texts are generated and what the actual performance characteristics of the systems are (e. g., the systems' tendency to favour fluency over adequacy and the fact that LLMs learn from the input they receive, cf. Krüger 2023: 287–296, 311) is it possible to develop adequate prompting strategies and to detect potential risks and sources of errors.

We can conclude that, if used correctly, AI tools such as LLMs can definitely assist the translation process, but their use is tied to specific conditions, skills, and prior knowledge, both pertaining to AI literacy and to more 'traditional' core translation competences. If the translation is to be a functionally adequate, high-quality and rule-consistent text, a human review of the output remains indispensable. It is crucial to keep in mind that the unreflected use of tools like GPT-4o can have severe consequences not only for readers but also for translators, as they are the ones who are made liable for the published translation.[8] Therefore, we strongly advise against solely trusting an MT tool to generate a functionally adequate, high-quality translation in both EL/PL and LRL translation contexts. In today's fast-paced digital age, it is more important than ever to publish information quickly (as was the case during the COVID-19 pandemic, for example). This is associated with a currently reduced willingness to invest in high-quality translations. Therefore, it can be assumed that AI tools such as GPT-4o and other LLMs will increasingly be integrated into intra- and interlingual translation workflows. Still, as of now, the output produced by these tools may predominantly serve merely as a draft for professional translators, meaning that these tools may well assist but will not replace human experts in intra- and interlingual communication.

---

8 In this context, aspects such as data ethics and risk assessments skills are also required (for more details see Nitzke et al. 2019; Krüger 2023 and dimensions 4 and 5 of the AI literacy framework).

Appendix

The appendix is available at
https://th-koeln.sciebo.de/s/EdQOXMUohqk4UT6 (29 September 2025)

References

Ahrens, Sarah; Silvana Deilen, Sergio Garrido, Ekaterina, Lapshinova-Koltunski, Christiane Maaß (forthc.): “Evaluation of translations into Plain German produced by humans and MT-systems Including ChatGPT.” *SKASE Journal of Translation and Interpretation*

Anschütz, Miriam; Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, Georg Groh (2023): “Language models for German text simplification: Overcoming parallel data scarcity through style-specific pre-training.” Anna Rogers, Jordan Boyd-Graber, Naoaki Okazaki (eds): *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto: Association for Computational Linguistics, 1147–1158 – https://doi.org/10.18653/v1/2023.findings-acl.74

De Marneffe, Marie-Catherine; Christopher Manning, Joakim Nivre, Daniel Zeman (2021): “Universal dependencies.” *Computational Linguistics* 47 [2]: 255–308 – https://doi.org/10.1162/coli_a_00402

Deilen, Silvana (2021): *Optische Gliederung von Komposita in Leichter Sprache. Blickbewegungsstudien zum Einfluss visueller, morphologischer und semantischer Faktoren auf die Verarbeitung deutscher Substantivkomposita*. Berlin: Frank & Timme

Deilen, Silvana; Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, Christiane Maaß (2023): “Using ChatGPT as a CAT tool in Easy Language translation.” Sanja Štajner, Horacio Saggio, Matthew Shardlow, Fernando Alva-Manchego (eds): *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*. Varna: INCOMA, 1–10 – https://aclanthology.org/2023.tsar-1.1/ (14 March 2025)

Deilen, Silvana; Ekaterina Lapshinova-Koltunski, Sergio Garrido, Julian Hörner, Christiane Maaß, Vanessa Theel, Sophie Ziemer (2024a): “Evaluation of intralingual machine translation for health communication.” Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, Helena Moniz (eds): *Proceedings of the 25th Annual Conference of the European Association for Machine Translation.* Vol. 1. Sheffield: European Association for Machine Translation, 469–479 – https://aclanthology.org/2024.eamt-1.39/ (14 March 2025)

Deilen, Silvana; Ekaterina Lapshinova-Koltunski, Sergio Garrido, Julian Hörner, Christiane Maaß, Vanessa Theel, Sophie Ziemer (2024b): “Towards AI-supported health communication in Plain Language: Evaluating intralingual machine translation of medical texts.” Dina Demner-Fushman, Sophia Ananiadou, Paul Thompson, Brian Ondov (eds): *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*. Torino: ELRA/ICCL, 44–53 – https://aclanthology.org/2024.cl4health-1.6/ (14 March 2025)

EU Artificial Intelligence Act (2024): “EU Artificial Intelligence Act chapter I (general provisions) article 3 (definitions). Paragraph 63.” *European Union* – https://artificialintelligenceact.eu/article/4/ (6 December 2024)

Gutermuth, Silke (2020): *Leichte Sprache für alle?: eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache*. Berlin: Frank & Timme

Hansen-Schirra, Silvia; Christiane Maaß (2019): Translation proper: Kommunikationsbarrieren überwinden. – https://d-nb.info/1196794170/34

Hansen-Schirra, Silvia; Jean Nitzke, Silke Gutermuth, Christiane Maaß, Isabel Rink (2020): "Technologies for translation of specialised texts into easy language. Easy Language research: Text and user perspectives." Silvia Hansen-Schirra, Christiane Maaß (eds.): *Easy Language research: Text and user perspectives*. Berlin: Frank & Timme, 99–127

Jakobson, Roman (1959): "On linguistic aspects of translation." Reuben Arthur Brower (ed.): *On translation*. Cambridge: Harvard University Press, 232–239

Joshi, Pratik; Sebastian Santy, Amar Budhiraja, Kalika Bali, Monojit Choudhury (2020): "The state and fate of linguistic diversity and inclusion in the NLP world." Dan Jurafsky, Joyce Chai, Natalie Schluter, Joel Tetreault (eds): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 6282–6293 – https://aclanthology.org/2020.acl-main.560.pdf (14 March 2025)

Krüger, Ralph (2023): "Artificial intelligence literacy for the language industry – With particular emphasis on recent large language models such as GPT-4." *Lebende Sprachen* 68 [2]: 283–330 – https://doi.org/10.1515/les-2023-0024

Krüger, Ralph (2024): "Outline of an artificial intelligence literacy framework for translation, interpreting and specialised communication." *Lublin Studies in Modern Languages and Literature* 48 [3]: 11–23 – https://doi.org/10.17951/lsmll.2024.48.3.11-23

Linguistic Data Consortium (LDC) (2024): "Search the LDC catalog." *LDC* – https://catalog.ldc.upenn.edu/search (15 January 2025)

trans-kom **ISSN 1867-4844**

trans-kom ist eine wissenschaftliche Zeitschrift für Translation und Fachkommunikation.

trans-kom veröffentlicht Forschungsergebnisse und wissenschaftliche Diskussionsbeiträge zu Themen des Übersetzens und Dolmetschens, der Fachkommunikation, der Technikkommunikation, der Fachsprachen, der Terminologie und verwandter Gebiete.

Beiträge können in deutscher, englischer, französischer oder spanischer Sprache eingereicht werden. Sie müssen nach den Publikationsrichtlinien der Zeitschrift gestaltet sein. Diese Richtlinien können von der **trans-kom**-Website heruntergeladen werden. Alle Beiträge werden vor der Veröffentlichung anonym begutachtet.

trans-kom wird ausschließlich im Internet publiziert: https://www.trans-kom.eu

Maaß, Christiane (2020): *Easy language – plain language – easy language plus: Balancing comprehensibility and acceptability*. Berlin: Frank & Timme

Maaß, Christiane (2024): "Intralingual translation in Easy Language and in Plain Language." Linda Pillière, Özlem Berk Albachten (eds): *The Routledge handbook of intralingual translation*. Oxon/New York: Routledge, 234–251 – https://www.taylorfrancis.com/chapters/edit/10.4324/9781003188872-18/intralingual-translation-easy-language-plain-language-christiane-maa%C3%9F (14 March 2025)

Maaß, Christiane; Sergio Hérnandez Garrido (forthc.): "Was ist Einfache Sprache?" Christiane Maaß, Ekaterina Lapshinova-Koltunski, Julian Hörner (eds): *Einfache Sprache mit KI-Tools: Ein Leitfaden für die redaktionelle Praxis und Praxisleitfaden.* Wiesbaden: Springer

Maaß, Christiane; Isabel Rink, Christiane Zehrer (2014): "Leichte Sprache in der Sprach- und Übersetzungswissenschaft." Susanne J. Jekat, Heike Elisabeth Jüngst, Klaus Schubert, Claudia Villiger (eds): *Sprache barrierefrei gestalten. Perspektiven aus der Angewandten Linguistik*. Berlin: Frank & Timme, 53–85

The MQM Council (2024): "MQM (Multidimensional Quality Metrics)." – https://themqm.org/error-types-2/typology/ (12 March 2025)

Nitzke, Jean; Silvia Hansen-Schirra, Carmen Canfora (2019): "Risk management and post-editing competence." *The Journal of Specialised Translation* 31 [1]: 239–259

OpenAI (n. d.): "Prompt engineering. Enhance results with prompt engineering strategies." *OpenAI* – https://platform.openai.com/docs/guides/prompt-engineering (13 March 2025)

Òmnium Cultural (2024): "Parla'm." *Òmnium* – https://parlam.omnium.cat/ (15 January 2025)

Paasch-Kaiser, Christine; Felix Hoberg (2020): "La traducció automàtica en l'ensenyament de la traducció del Català vers l'alemany: Anàlisi dels resultats de les eines Google Translate i Bing Microsoft Translator." *RESERCLE. Revista de La Societat d'Ensenyament i Recerca Del Català Com a Llengua Estrangera* 1 [1]: 66–89 – https://doi.org/10.31009/reservle.2020.i1.05

Paasch-Kaiser, Christine; Felix Hoberg (submitted): "Pragmatik und maschinelle Übersetzung: ein Beitrag zur Translationsdidaktik im Bereich der Klein- und Minderheitensprachen." *Tagungsband Sektion I "Romanische Minderheitensprachen"*, 38. Romanistentag 2023

Ranathunga, Surangika; En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, Rishemjit Kaur (2021): "Neural machine translation for low-resource languages: A survey." *ACM Computing Surveys* 55 [11]: 1–37 – https://doi.org/10.1145/3567592

Reiß, Katharina; Hans J. Vermeer (1984): *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: Niemeyer

Rink, Isabel (2019): *Rechtskommunikation und Barrierefreiheit: Zur Übersetzung juristischer Informations- und Interaktionstexte in Leichte Sprache*. Berlin: Frank & Timme

Robinson, Nathaniel; Perez Ogayo, David R. Mortensen, Graham Neubig (2023): "ChatGPT MT: Competitive for high- (but not low-) resource languages." Philipp Koehn, Barry Haddow, Tom Kocmi, Christof Mon (eds): *Proceedings of the Eighth Conference on Machine Translation*. Singapore: Association for Computational Linguistics, 392–418 – https://doi.org/10.18653/v1/2023.wmt-1.40

Rodríguez Vázquez, Silvia; Abigail Kaplan, Pierrette Bouillon, Cornelia Griebel, Razieh Azari (2022): "La traduction automatique des textes faciles à lire et à comprendre (falc): une étude comparative." *Meta* 67 [1]: 18–48 – https://doi.org/10.7202/1092189ar

Salvanyà, Jaume (2005): "Expressions col·loquials incorrecte." *Llengua Nacional* 50: 34 – https://llenguanacional.cat/pdf/LN50_13.pdf (14 March 2025)

Salvanyà, Jaume (2020): "Silencis del DIEC." *Blog lingüístic* – https://silencisdeldiec.wordpress.com/ (14 March 2025)

Schaeffer, Doris; Klaus Hurrelmann, Ullrich Bauer, Kai Kolpatzik (2018): *National Action Plan Health Literacy: Promoting health literacy in Germany*. Berlin: KomPart

Teo, Sheila (2023): “How I won Singapore’s GPT-4 prompt engineering competition.” *Medium* – https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41/ (12 March 2025)

Universal Dependency Relations – https://universaldependencies.org/u/dep/ (29 September 2025)

Wei, Jason (2024): “Intuitions on language models.” Stanford CS25 2024 guest lecture – https://youtu.be/3gb-ZkVRemQ?si=GRUi-0ogZJuojFKc (12 March 2025)

White, Jules; Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, Douglas C. Schmidt (2023): “A prompt pattern catalog to enhance prompt engineering with ChatGPT.” *arXiv* – https://doi.org/10.48550/arXiv.2302.11382

Wikipedia (2024a): “Estadístiques.” *Wikipedia* – https://ca.wikipedia.org/wiki/Especial:Estad%C3%ADstiques (15 January 2025)

Wikipedia (2024b): “Statistik.” *Wikipedia* – https://de.wikipedia.org/wiki/Spezial:Statistik (15 January 2025)

Zhao, Wayne Xin; Kun Zhou, Junyi Li, Tiany Tangi, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen (2023): “A survey of large language models.” *arXiv* – https://doi.org/10.48550/arXiv.2303.18223

Authors

Silvana Deilen is Assistant Professor of Translation Studies with a focus on Translation Processes at the Institute for Translation Studies and Interpreting at Heidelberg University. She received her PhD from the University of Mainz in 2021. In her PhD thesis, she conducted an eye tracking study on the visual segmentation of compounds in Easy Language. Her main research interests are machine translation, Easy and Plain Language, specialized translation and eye-tracking.
E-mail: silvana.deilen@iued.uni-heidelberg.de

Christine Paasch-Kaiser is a postdoctoral research fellow at the Institute for Applied Linguistics and Translatology at Leipzig University. She received her PhD from Leipzig University in 2013. In her PhD thesis, she conducted an empirical study in the area of language varieties on Spanish spoken in the Spanish Basque Country. Her current research focuses on historical legal French, Romance minority languages, machine translation and translation didactics.
E-mail: christine.paasch-kaiser@uni-leipzig.de

Ralph Krüger is Full Professor of Language and Translation Technology at the Institute of Translation and Multilingual Communication at TH Köln – University of Applied Sciences, Cologne. He received his PhD from the University of Salford in 2014 and completed his habilitation at Johannes Gutenberg University Mainz in 2024. Ralph is co-editor of the journal Lebende Sprachen and his current research focuses on developing and operationalising an artificial intelligence literacy framework for translation, interpreting and specialised communication.
E-mail: ralph.krueger@th-koeln.de

Empfehlungen

TRANSÜD. Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens

Herausgegeben von Hartwig Kalverkämper und Sylvia Reinart

Hervé Delplanque / Stephanie Schwerter (dir.): **L'incompréhension culturelle à l'épreuve du droit.** ISBN 978-3-7329-1071-7

Easy – Plain – Accessible

Herausgegeben von Silvia Hansen-Schirra und Christiane Maaß

Laura Marie Maaß: **Erwartungen, Einstellungen, Erfahrungen. Zur Interaktion zwischen hörenden Gebärdensprachdolmetschenden und ihrer tauben Kundschaft.** ISBN 978-3-7329-1161-5

Sarah Ahrens: **Einfache Sprache in der Gesundheitskommunikation. Patientinnenaufklärung für Frauen mit Deutsch als Zweitsprache.** ISBN 978-3-7329-1132-5

Anke Radinger: **Researching Subtitling Processes. Methodological considerations for the investigation of AI-assisted subtitling workflows.** ISBN 978-3-7329-1029-8

Forum für Fachsprachen-Forschung

Herausgegeben von Hartwig Kalverkämper

Claudio Di Meola / Joachim Gerdes / Livia Tonelli (Hg.): **Sprachvariation im Deutschen zwischen Theorie und Praxis. Fachsprachlichkeit, Inklusion, Didaktik, Übersetzung, Kontrastivität.** ISBN 978-3-7329-1042-7

Anna Malena Pichler / Barbara von der Lühe / Felicitas Margarete Tesch (Hg.): **Deutsch als Fremd- und Fachsprache im internationalen Kontext.** ISBN 978-3-7329-1047-2

Daniel Green (Hg.): **The Future of Teaching Law and Language.** ISBN 978-3-7329-1090-8

Audiovisual Translation Studies (AVTS)

Herausgegeben von Marco Agnetta und Alexander Künzli

Marco Agnetta / Astrid Schmidhofer / Alena Petrova (Hg.): **Bild – Ton – Sprachtransfer. Neue Perspektiven auf Audiovisuelle Translation und Media Accessibility.** ISBN 978-3-7329-0921-6

Alexander Künzli / Klaus Kaindl (Hg.): **Handbuch Audiovisuelle Translation. Arbeitsmittel für Wissenschaft, Studium, Praxis.** ISBN 978-3-7329-0981-0

Transkulturalität – Translation – Transfer

Herausgegeben von Martina Behr, Larisa Schippel und Julia Richter

Michael Schreiber: **Die Übersetzungspolitik der Französischen Revolution und der napoleonischen Zeit. Studien zu Übersetzungen ins Deutsche, Italienische, Niederländische und Kreolische.** ISBN 978-3-7329-1141-7

Leipzig Middle East Studies

Herausgegeben von Sebastian Maisel

Ahmed Ibrahim Abdallah Mohmmed: **Übersetzung als interkultureller Dialog. Sudanesische Romane in deutscher Übersetzung.** ISBN 978-3-7329-1087-8

Malek Al Refaai: **Challenges in Interpretation Services for Syrian Refugees.** ISBN 978-3-7329-1193-6

Magdalena Zehetgruber / Bernadette Hofer-Bonfim / Elisabeth Peters / Johannes Schnitzer (Hg.): **Linguistic Diversity in Professional Settings.** ISBN 978-3-7329-1089-2

Lola Debüser / Jekatherina Lebedewa: **Ein deutsch-russisches Leben.** ISBN 978-3-7329-1181-3

Alle Bücher sind auch als E-Books erhältlich.

Frank & Timme

Wittelsbacherstraße 27a, 10707 Berlin • Telefon +49 (0)30 88 66 79 11 • info@frank-timme.de • www.frank-timme.de