asobid
Amasya Üniversitesi Sosyal Bilimler Dergisi

ISSN: 2548-0480 (Print)
Sayı/Issue 15 ● Haziran/June 2024
Sayfa/Page: 120-173

Geliş Tarihi / Received: *29.01.2024* ***Kabul Tarihi / Accepted:*** *08.02.2024*

Çetin, Ö.- Duran, A. (2024). A comparative analysis of the performances of chatgpt, deepl, google translate and a human translator in community based settings. *Amasya Üniversitesi Sosyal Bilimler Dergisi (ASOBİD)*. 9 (15), s. 120-173.

A COMPARATIVE ANALYSIS OF THE PERFORMANCES OF CHATGPT, DEEPL, GOOGLE TRANSLATE AND A HUMAN TRANSLATOR IN COMMUNITY BASED SETTINGS

CHATGPT, DEEPL VE GOOGLE ÇEVİRİ: FARKLI METİN TÜRLERİNDE ÇEVİRİ KALİTESİ DEĞERLENDİRMESİ

Dr. Öğr. Üyesi Özge ÇETİN[1]
Amasya Üniversitesi
ozge.cetin@amasya.edu.tr

Dr. Öğr. Üyesi Ali DURAN[2]
Amasya Üniversitesi
aliduranedu@hotmail.com

Abstract
The diversity of languages is a remarkable aspect of human civilization, reflecting a wide range of cultures and life experiences. However, this diversity can sometimes pose challenges, especially during interactions with speakers of different languages. Machine translation (MT) offers a solution to minimize the impact of these linguistic barriers. MT enables swift understanding of information, effective idea exchange, and the building of relationships across varied cultural backgrounds. Prominent translation tools include Google MACHINE TRANSLATION, DeepL, Bing Microsoft Translator, and Amazon Translate. Additionally, a newer AI technology, ChatGPT by OpenAI, introduced in November 2022, has been making strides in

[1] ORCİD: orcid.org/ 0000-0002-7249-8755
[2] ORCİD: orcid.org/ 0000-0001-6132-4066

this domain. This has sparked a debate in various industries about the potential of ChatGPT to replace human roles. A pertinent question in Translation Studies (TS) is the effectiveness of ChatGPT as a translator. It is posited that ChatGPT, akin to other machine learning models, delivers contextually richer translations. This study compares ChatGPT's translation capabilities with those of Google MT and DeepL across different text types, informed by past literature. To conduct this comparison, we selected text types that are traditionally challenging to translate, guided by Katharina Reiss' Text Type Model, which categorizes texts based on their communicative purposes: informative, expressive, and operative. This study assesses the translations of source texts on education, heathcare and law by ChatGPT, DeepL, Google MT, and a human translator, drawing certain conclusions in consideration of these categories. Our research adopts a qualitative approach, evaluating the translations using a machine translation quality model, called the Multidimensional Quality Metrics (MQM) model. The insights from this study will benefit T&I researchers interested in machine translation and the users of these technologies.

Keywords: ChatGPT, DeepL, Google MACHINE TRANSLATION, Artificial Intelligence, Machine Translation, Human Translator, Translation Quality

Öz

Dil çeşitliliği, çok çeşitli kültürleri ve deneyimleri temsil etmesi bakımından bir zenginlik olarak değerlendirilebilir. Bununla birlikte, bu çeşitliliğin özellikle farklı bir dil konuşan bireylerle iletişim kurarken zaman zaman bir engel teşkil edebileceği de yadsınamaz bir gerçektir. Ancak, makine çevirisi (MACHINE TRANSLATION) sayesinde dil engellerinin etkisi azaltılabilir. MT sayesinde bilgi hızlı bir şekilde anlaşılabilir, fikirler başarılı bir şekilde iletebilir ve farklı kültürlerden diğer kişilerle bağlantı kurulabilir. Bu doğrultuda Google MT ve DeepL günümüzde kullanılan en popüler çeviri araçları arasındadır. Bunlar dışında çok sayıda başka araçlar da bulunmaktadır. Son aylarda ise ChatGPT çeviri aracı olarak öne çıkan uygulamalar arasında değerlendirilmektedir. ChatGPT modern yapay zekanın adıdır ve giderek yaygınlaşmaktadır. OpenAI'nin Kasım 2022'de ChatGPT'yi piyasaya sürmesinden bu yana, yapay zekanın birçok çalışanın işini elinden alacağı endişesi yaygınlaşmaktadır. “ChatGPT iyi bir çevirmen mi?” sorusu çeviri alanında sıklıkla sorulan bir soru olarak değerlendirilmektedir. ChatGPT'nin, diğer makine öğrenimi modelleri gibi, bağlama dayalı olarak çok daha doğru çeviriler ürettiği iddia edilmektedir. Bu açıdan ele alındığında, mevcut

literatür bulgularına dayanarak, ChatGPT'nin etkileyici bir şekilde yapabildiği şeylerden biri metin çevirisi olması nedeniyle farklı metin türlerinde Google MT ve DeepL ile nasıl bir performans sergileyeceği araştırılmaya değer bir konu olarak değerlendirilebilir. Bu araştırmada söz konusu bu çeviri araçlarını karşılaştırmak için, Katharina Reiss'ın yaygın çeviri sorunlarını vurgulayan metin türü modeli referans alınmıştır. Reiss'a göre, iletişimsel işlevlerine göre üç metin türü bulunmaktadır: bilgilendirici metinler, anlatımcı (dışavurumsal) metinler ve işlevsel metinler. Buna göre bu araştırmanın amacı eğitim, sağlık ve hukuk alanlarından metinlerinin insan çevirisi, Google MT çevirisi, DeepL çevirisi ve ChatGPT çevirisi arasında karşılaştırmalar yapmak ve buna göre bazı çıkarımlarda bulunmaktadır. Bu araştırma nitel bir çalışmadır. Doküman analizine dayalı olan bu araştırmada, ChatGPT, DeepL, Google MT insan çevirmen tarafından yapılan çeviriler Çok Boyutlu Kalite Ölçütleri (ÇBKÖ) modeline göre değerlendirilmiştir. Elde edilen bulguların, makine çevirisiyle ilgilenen araştırmacılarının yanı sıra bu teknolojilerin kullanıcıları için de faydalı olması beklenmektedir.

Anahtar Kelimeler: ChatGPT, DeepL, Google MACHINE TRANSLATION, Yapay Zekâ, Makine Çevirisi, İnsan Çevirisi, Çeviri Kalitesi.

Introductıon

The 21st century has marked a remarkable evolution in global communication, idea exchange, and interpersonal interactions (Khoshafah, 2023). This era of change has significantly influenced the field of translation, steering it away from conventional cultural practices and towards innovative technological advancements. The introduction and integration of machine translation, neural networks, natural language processing (NLP), and machine learning have collectively revolutionized the translation industry, signaling a major shift in how translation is approached and executed in modern times (Ali et al., 2023). While technological advancements in transportation and communication have reduced the physical barriers to communication, linguistic diversity poses greater challenges. Given the degree of interconnectedness and the resulting need for human communication, manual translation is no longer scalable enough to meet these needs. Machine translation (MT) is required to

automate the translation of natural languages (Kunchukuttan & Bhattacharyya, 2021). As a result, machine translation (MT) has emerged as an essential resource, utilized daily by millions who often do not scrutinize its precision. Individuals using free translation platforms like Google MT and Microsoft Translate often lack the ability to assess the quality of the translations provided. When familiar with the target language, they might only gain a basic sense of the translation's flow and natural tone. However, they remain oblivious to the actual merit of the translation, including its accuracy and absence of errors in meaning, style, and idiomatic expression. Moreover, if they are not versed in the language of the translation, they are completely in the dark regarding the value of the translated output (Almahasees, 2021). The aim of contemporary research in machine translation focuses not on achieving flawless translations, but rather on diminishing the frequency of errors in these systems (Koehn, 2010). The popularity of free online MT tools continues to increase (Bowker, 2023). The rapid development of machine translation (MT) technologies in recent years has generated significant interest in understanding their capabilities and potential implications for the field of translation. At the forefront of this progress are neural MT systems, such as Google MT, DeepL, and ChatGPT, which have significantly improved the quality of machine-generated translations compared to their rule-based and statistical predecessors (Koehn, 2010).

ChatGPT, Google MT, and DeepL, while all being AI (Artificial Intelligence) systems, have distinct functions and employ varied methodologies. OpenAI's ChatGPT, based on GPT-4 architecture, specializes in comprehending and generating natural language for tasks like responding to queries, summarizing passages, and engaging in dialogue. It operates on a deep learning framework known as the Transformer, pre-trained on an extensive collection of internet text. This model acquires the skill of text comprehension and creation through the prediction of subsequent words in a sentence, considering the given context This enables it to generate coherent and contextually relevant responses (Li et al., 2024; Ray, 2023; Siu, 2023). Google MT, offered by Google, is a

complimentary service for translating between multiple languages. It employs a neural machine translation (NMT) approach to convert text from one language to another. Similar to ChatGPT in its reliance on the Transformer architecture, Google MT is, however, exclusively honed for translation tasks. This system is trained using extensive parallel text corpora in various languages, enabling it to translate sentences while retaining their original meaning. Over time, Google MT has seen significant enhancements in terms of translation accuracy and the expansion of its language repertoire (Bansal et al., 2024; Li et al., 2024). DeepL, an AI-driven translation service, is a creation of DeepL GmbH, a German firm. It parallels Google MT in employing neural machine translation (NMT) for language conversion. DeepL also utilizes the Transformer architecture, setting itself apart with the assertion of superior translation quality relative to its counterparts. This superior quality is attributed to the integration of sophisticated training methods, optimization algorithms, and a comprehensive, high-grade training dataset. (Agung et al., 2024; Girletti, & Lefer, 2024).

However, despite these advancements, a thorough and systematic comparison of the performance of human translators and machine translation methods across different text types remains an area that needs further exploration. This research is grounded in the intersection of machine translation, translation quality assessment, and Reiss' text type model. The proposed model seeks to establish a structured method to analyze and contrast the translation capabilities of human translators, Google MT, DeepL, and ChatGPT in contexts like education, healthcare, and law. For evaluating translation accuracy, this study utilizes the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014). While prior research has delved into the efficacy of various MT systems, comprehensive comparisons of cutting-edge neural MT technologies such as ChatGPT against human translators in varied text genres are scarce. Most existing studies have focused on specific fields or languages, which may restrict the broader applicability of their conclusions. In addition, a significant portion of these studies have depended on automated

evaluation methods, which might not always correspond with human judgments of translation quality (Callison-Burch et al., 2006). The MQM model provides a comprehensive framework for evaluating translation quality based on various dimensions, such as fluency, accuracy, and style, ensuring a thorough comparison of the translation methods (Snow, 2015).

1. Purpose of the Research

This research endeavors to explore and contrast the translation abilities of human translators, Google MT, DeepL, and ChatGPT in educational, healthcare, and legal documents. The goal is to enhance comprehension of the advantages and drawbacks of each MT tool across different scenarios, thereby offering insightful information for both scholars and users of MT technologies. Addressing the gaps noted in previous studies, this research conducts an extensive comparative analysis of human translators, Google MT, DeepL, and ChatGPT in translating texts related to education, health, and law. The selection of these text types is guided by Katharina Reiss' text type model (Reiss, 1971), which categorizes texts based on their communicative roles: informative, expressive, and transactional. Through this examination of varied text genres, the study aims to present a comprehensive perspective on the capabilities and limitations of each MT tool in diverse settings.

This study implements the MQM framework (Lommel et al., 2014) to evaluate translation quality. The MQM model is a detailed system for measuring translation effectiveness across various aspects, including fluency, accuracy, and stylistic elements. This methodology facilitates an in-depth analysis of different translation methods, more accurately reflecting human evaluations of translation quality. By incorporating machine translation with the MQM model and Reiss' text type model, the research is structured to systematically compare the performances of human translators, Google MT, DeepL, and ChatGPT in community-based settings. The findings from this study are expected to enrich the collective knowledge of machine translation efficiency and its wider implications in the translation

sector, proving beneficial for both scholars and practitioners in translation technology.

1.1. Research Questions

RQ1. How does the translation performance of human translators, Google MT, DeepL, and ChatGPT compare when applied to educational texts?

RQ2. How does the translation performance of human translators, Google MT, DeepL, and ChatGPT compare when applied to health-related texts?

RQ3. How does the translation performance of human translators, Google MT, DeepL, and ChatGPT compare when applied to legal texts?

RQ4. What are the specific strengths and limitations of each translation method in the context of these three text types?

RQ5. How do the outcomes of this research enhance our collective comprehension of machine translation's efficacy and its prospective influence on the translation sector?

2. Conceptual Framework

2.1. Machine Translation

Machine translation (MT) has undergone significant advancements since its inception in the 1950s. Early approaches to MT were based on rule-based systems, which relied on linguistic knowledge and dictionaries to perform translations (Hutchins, 2003). However, these systems were limited in their ability to capture the complexities and nuances of human language. Statistical machine translation (SMT) emerged in the late 1980s and gained popularity in the 1990s as a data-driven approach that utilized parallel corpora to generate translations (Brown et al., 1990; Koehn, 2010). SMT addressed some of the limitations of rule-based systems, but its performance was still hindered by the lack of contextual understanding. The introduction of neural machine translation (NMT) in the 2010s revolutionized the field, as it leveraged deep learning to generate translations with improved fluency and accuracy (Sutskever et al., 2014; Bahdanau et al., 2014). Notable NMT systems, such as

Google MT, DeepL, and ChatGPT, have demonstrated superior performance compared to their rule-based and statistical predecessors.

2.2. Multidimensional Quality Metrics (MQM) Model

A wide range of methods for evaluating machine translation quality have been developed, particularly through error typologies that focus on assessing the quality of machine translation output in comparison between the source and target texts (Blain et al., 2011; Comelles et al., 2017; Costa et al., 2015; Lommel et al., 2014; Popović, 2018; Stymne, & Ahrenberg, 2012; Vilar et al., 2006). Translation quality assessment (TQA) is crucial in evaluating the performance of MT systems. Early TQA approaches were often based on human evaluation, which, while valuable, could be subjective and time-consuming (Lauscher, 2000). The development of automatic evaluation metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and TER (Snover et al., 2006), provided more objective and efficient ways to assess translation quality. However, automatic metrics may not always align with human perception of translation quality (Callison-Burch et al., 2006). This study adopts the MQM model (Lommel et al., 2014), which offers a comprehensive framework for assessing translation quality based on various dimensions, such as fluency, accuracy, and style. Here is the main steps in the MQM.

1. Terminology: Mistakes linked to terminology occur when a term in the target text doesn't align with established domain or organizational standards or when it isn't a correct equivalent to the source text term (e.g., Inconsistencies with terminology resources, inconsistent terminology use, incorrect term selection).

2. Accuracy: These errors emerge when the target text fails to precisely mirror the source text's propositional content, manifesting as distortions, omissions, or additions (e.g., Incorrect translations, excessive translation, insufficient translation, unnecessary additions, omissions, failure to translate, untranslated segments).

3. Linguistic Conventions: Errors concerning the linguistic integrity of the text, including issues with grammatical and mechanical correctness (e.g., Grammar errors, punctuation mistakes, spelling errors, unintelligibility, character encoding issues).

4. Style: Errors in texts that, while grammatically correct, are inappropriate due to deviations from organizational style guides or unsuitable language style (e.g., Non-compliance with organizational style, third-party style issues, inconsistencies with external references, register problems, awkward styling, unidiomatic expressions, inconsistent styling).

5. Locale Conventions: Mistakes occur when translation does not adhere to locale-specific content or formatting rules for various data elements (e.g., Incorrect number, currency, measurement, time, date, address, telephone formatting, and shortcut key conventions).

6. Audience Appropriateness: Errors resulting from content in the translation that is unsuitable or invalid for the target locale or audience (e.g., Inappropriate culture-specific references).

7. Design and Markup: Issues related to the physical design or presentation of a translation, including formatting and markup of characters, paragraphs, and UI elements, integration with graphical elements, and overall layout of pages or windows (e.g., Character formatting issues, layout problems, markup tag errors, truncation/text expansion, missing text, faulty links or cross-references).

3. Method

3.1. Research Design

Document analysis, a qualitative research method that involves analyzing written or recorded material to obtain a deeper understanding of a phenomenon, was utilized for this study (Mellinger & Hanson, 2016). This study compared the efficacy of human translators, Google MT, DeepL, and ChatGPT in translating educational, medical, and legal texts. Texts were chosen in accordance with Katharina Reiss' text type model, which categorizes texts according to their communicative functions. To

ensure a rigorous evaluation of translation quality, the MQM model was used to evaluate translation quality across multiple dimensions, including fluency, accuracy, and style. By integrating machine translation, the MQM model, and Reiss' text type model, this study aimed to provide a solid foundation for comparing the performance of different translation methods and gain a more comprehensive understanding of the strengths and limitations of each method in different contexts. This study is qualitative because it delves into the intricate and subjective aspects of translation quality, which extends beyond mere numerical analysis. The complexity of translation, involving context, cultural nuances, and intended meaning, necessitates an interpretative approach. Utilizing the MQM model, the research emphasizes qualitative evaluation across dimensions like fluency, accuracy, and style. Subjective judgments by researchers about translation quality, especially in analyzing varied communicative functions across different text types as per Katharina Reiss' Text Type model, underscore the qualitative essence of the study. This approach is vital for understanding the nuances of translation quality and contributes to the field of Translation Studies by offering in-depth insights rather than mere statistical generalizations, aligning with the study's aim to explore the effectiveness of different translation methods in a nuanced manner.

3.2. Researchers' Roles

In the study, researchers undertook crucial roles encompassing the entire research process. We designed the study, including formulating research questions and selecting texts from educational, medical, and legal domains based on Katharina Reiss' Text Type Model. The team conducted translations using Google MT, DeepL, and ChatGPT, ensuring standardized conditions for fairness. They then evaluated the translations using the MQM model, focusing on fluency, accuracy, and style. Subsequent data analysis involved comparing the effectiveness of each translation method across various text types and communicative purposes. The researchers were also responsible for compiling, reporting, and disseminating the findings, ensuring clarity, conciseness, and

objectivity in presenting the results to the Translation and Interpreting (T&I) community and users of machine translation technologies.

3.3. Materials and Data Collection

For this study, we compiled a corpus of educational, health, and legal texts based on community settings in Australia. The documents selected for this comparison were "FACT SHEET: Information for parents" as the educational text, "Complaint Form" as the health-related text, and "Happily Ever... Before and After" as the legal text. The texts were selected according to the text type classification proposed by Reiss. It consisted of a total of three texts, with their different translations, namely human translator, Google MT, DeepL, and ChatGPT. To collect data for the study, each text was translated by researchers through Google MT, DeepL, and ChatGPT. The human translations are already available together with the original versions.

3.4. Data Analysis

We analyzed the data using the MQM model. MQM scores were evaluated for each translation and for each text type based on various dimensions such as terminology (inconsistent with terminology resource, inconsistent use of terminology, wrong term), accuracy: (mistranslation, over-translation, under-translation, addition, omission, do not translate (DNT), untranslated), linguistic conventions (grammar, punctuation, spelling, unintelligible, character encoding), style: errors occurring in a text that can be grammatical but are inappropriate because they deviate from organizational style guides or exhibit inappropriate language style (organizational style, third-party style, inconsistent with external reference, register, awkward style, unidiomatic style, inconsistent style), locale conventions (number format, currency format, measurement format, time format, date format, address format, telephone format, shortcut key), audience appropriateness: errors arising from the use of content in the translation product that is invalid or inappropriate

for the target locale or target audience (culture-specific reference) and design and markup: (character formatting, layout, markup tag, truncation/text expansion, missing text, link/cross-reference) (Lommel, 2018; Lommel et al., 2014).

4. Findings

We compared the performance of human translation, Google MT, DeepL, and ChatGPT in community-based settings: education, health, and legal. These text types, based on Katharina Reiss' model, were chosen due to their varying communicative functions, which present unique challenges in translation. The analysis employs the MQM model to evaluate the translations produced by each method. The findings are presented in three separate tables, with each table focusing on one text type.

In the findings section of our study, we present a comparative analysis of the texts employed. These texts were translated by human translators, Google MT, DeepL, and ChatGPT. Our alignment of the source text resulted in 99 sentences or expression matches for the educational text, 129 for the health text, and 62 for the legal text.

Due to the extensive length of these texts, our findings focus specifically on highlighting the differences in translation as they pertain to the MQM model parameters. These parameters include Terminology, Accuracy, Linguistic Conventions, Style, Locale Conventions, Audience Appropriateness, Design, and Markup. This selective approach allows us to concisely present the most significant discrepancies and variations in translation quality among the methods studied. By concentrating on these key differences, our study aims to provide a clearer and more focused insight into the strengths and limitations of each translation method across different text types. This analysis is critical for understanding how each translation method fares in handling specific aspects of language and content, thereby offering valuable perspectives for researchers and users of machine translation technologies in educational, health, and legal contexts.

RQ1. How does the translation performance of human translators, Google MT, DeepL, and ChatGPT compare when applied to educational texts?

Table 1. Side-by-Side Comparison of Human, Google, DeepL and ChatGPT Translations for Educational Context

Source Text	**FACT SHEET: Information for parents**
Human Translation	AİLELER İÇİN BİLGİLENDİRME BROŞÜRÜ
Google MT	BİLGİ FORMU: Ebeveynler için bilgiler
DeepL MT	FACT SHEET: Ebeveynler için bilgiler
ChatGPT	BİLGİLENDİRME: Ebeveynler için bilgi

In analyzing the translations of "FACT SHEET: Information for parents," we observe variations across the MQM model parameters. The human translation *AİLELER İÇİN BİLGİLENDİRME BROŞÜRÜ* introduces the term "brochure," adding specificity which does not present in the original and potentially aligning better with local conventions by broadening the target from "parents" to "families," thus affecting both terminology and audience appropriateness. Google MT's *BİLGİ FORMU: Ebeveynler için bilgiler* shifts the nature of the document from a "fact sheet" to an "information form," impacting the accuracy and terminology. DeepL MT retains the term "Fact Sheet" in English, a choice that could either respect linguistic conventions or indicate a lack of translation, depending on the target audience's familiarity with English terms. ChatGPT *BİLGİLENDİRME: Ebeveynler için bilgi* simplifies "Fact Sheet" to just "Information," losing some specificity. The style varies among the translations, reflecting different interpretations of "Fact Sheet," while the accuracy and linguistic conventions are generally upheld, despite some notable

choices in term usage. Design and markup aspects are not applicable in this context, as they relate to visual and structural elements not evident in the provided translations.

Source Text	DEPRESSION IN CHILDREN
Human Translation	Çocuklarda Depresyon ve Aileler Olarak Bilmemiz Gerekenler
Google MT	ÇOCUKLARDA DEPRESYON
DeepL MT	ÇOCUKLARDA DEPRESYON
ChatGPT	ÇOCUKLARDA DEPRESYON

The translations of "DEPRESSION IN CHILDREN" show distinct variations when analyzed through the lens of the MQM model parameters. The human translation, *Çocuklarda Depresyon ve Aileler Olarak Bilmemiz Gerekenler,* expands significantly on the source, introducing the concept of "what families need to know" alongside "Depression in Children." This addition greatly affects the parameters of accuracy, as it introduces information which does not present in the original, and audience appropriateness, by directly addressing families and implying a broader scope of content. It also influences style, by providing a more informative and detailed approach. On the other hand, Google MT, DeepL MT, and ChatGPT, all rendering as *ÇOCUKLARDA DEPRESYON*, adhere closely to the source text in terms of terminology and accuracy, offering a direct and concise translation. These interpretations adhere to the essence and conciseness of the original, conforming aptly to linguistic norms and stylistic expectations, offering a direct translation without extraneous details or elaboration. To sum up, while the machine translations and ChatGPT adhere closely to the original in terms of precision, terminology, and stylistic fidelity, the human translation takes a broader route. This approach may increase its relevance and connection with the audience by integrating pertinent context for families, though it strays from the original in accuracy.

Source Text	There are a number of ways to work out how serious our children's feelings are:
Human	Çocuklarımızın duygularının, ne kadar ciddi olduğunu

Translation	anlamamız için, birkaç yol vardır:
Google MT	Çocuklarımızın ne kadar ciddi olduklarını anlamanın birkaç yolu vardır. duygular şunlardır:
DeepL MT	Çocuklarımızın sağlık durumlarının ne kadar ciddi olduğunu anlamanın birkaç yolu vardır. duygulardır:
ChatGPT	Çocuklarımızın duygularının ne kadar ciddi olduğunu anlamak için birkaç yöntem bulunmaktadır:

In evaluating the translations of "There are a number of ways to work out how serious our children's feelings are," notable differences emerge in terms of the MQM model parameters. The human translation, *Çocuklarımızın duygularının, ne kadar ciddi olduğunu anlamamız için, birkaç yol vardır*, adheres closely to the source text in terms of accuracy and terminology, effectively capturing the essence of assessing the seriousness of children's feelings. It maintains the original structure and message, aligning well with linguistic conventions and style. Google MT's version, *Çocuklarımızın ne kadar ciddi olduklarını anlamanın birkaç yolu vardır. duygular şunlardır*, slightly misinterprets the source. The translation implies assessing how serious the children themselves are, rather than their feelings, leading to a deviation in accuracy. The additional *duygular şunlardır* (feelings are as follows) seems disjointed, affecting the clarity and coherence in terms of style and linguistic conventions.

Source Text	Depression is more than feeling sad or being miserable.
Human Translation	Depresyon geçirmek, üzgün olmak ya da huysuzluktan daha farklı birşeydir.
Google MT	Depresyon, üzgün hissetmekten veya sefil olmaktan daha fazlasıdır.
DeepL MT	Depresyon üzgün hissetmekten veya mutsuz olmaktan daha fazlasıdır
ChatGPT	Depresyon, sadece üzgün hissetmekten veya mutsuz olmaktan daha fazlasıdır.

Analyzing the translations of "Depression is more than feeling sad or being miserable" in terms of the MQM model, each version

displays unique characteristics. The human translation *Depresyon geçirmek, üzgün olmak ya da huysuzluktan daha farklı birşeydir* introduces the term *deprasyon geçirmek* (experiencing depression), slightly altering the focus from the original text. The use of *huysuzluk* (grouchiness) instead of "miserable" changes the intensity and nature of the emotion described, impacting both terminology and accuracy. Google MT's *Depresyon, üzgün hissetmekten veya sefil olmaktan daha fazlasıdır* offers a more direct translation, closely aligning with the source text in terms of accuracy and terminology. *Üzgün hissetmek* (feeling sad) and *sefil olmak* (being miserable) accurately reflect the original emotions, preserving the style and linguistic conventions. DeepL MT's *Depresyon üzgün hissetmekten veya mutsuz olmaktan daha fazlasıdır* substitutes *mutsuz olmak* (being unhappy) for "miserable," slightly changing the intensity of the emotion but maintaining the overall message. ChatGPT, *Depresyon, sadece üzgün hissetmekten veya mutsuz olmaktan daha fazlasıdır*, adds *sadece* (only) for emphasis, aligning closely with the original in terms of accuracy and style, while slightly modifying the emphasis for clarity. Each translation varies in its approach to terminology and accuracy, reflecting different interpretations of the source text's emotional depth.

Source Text	It is often seen as a time of rebellion.
Human Translation	Genellikle isyankarlığın ortaya çıktığı bir dönem olarak da görülebilir.
Google MT	Genellikle bir isyan zamanı olarak görülür.
DeepL MT	Genellikle bir isyan dönemi olarak görülür.
ChatGPT	Sıklıkla isyan dönemi olarak görülür.

The translations of "It is often seen as a time of rebellion" display varied interpretations when evaluated using the MQM model. The human translation, "Genellikle isyankarlığın ortaya çıktığı bir dönem olarak da görülebilir," expands on the original by suggesting it's a period when rebelliousness emerges, adding a layer of interpretation not explicitly present in the source. This affects the translation's accuracy and style, as it introduces an

additional descriptive element. Google MT's *Genellikle bir isyan zamanı olarak görülür* and DeepL MT's *Genellikle bir isyan dönemi olarak görülür* are closer to the source, with both translating "rebellion" directly as *isyankarlık* and maintaining the simplicity of the original statement. ChatGPT, *Sıklıkla isyan dönemi olarak görülür*, uses *sıklıkla* (frequently) instead of *genellikle* (often), which slightly alters the frequency implied by the original text but still stays within the bounds of accuracy and style. Each version reflects a different aspect of linguistic conventions and terminology, showing how the concept of a "time of rebellion" can be variably interpreted in the target language.

Source Text	2. Do our children's low feelings show in other parts of their lives?
Human Translation	Çocuklarımızın kendilerini kötü hissetmesi hayatın başka alanlarında da kendini gösteriyor mu?
Google MT	2. Çocuklarımızın düşük duyguları hayatlarının diğer alanlarında kendini gösteriyor mu?
DeepL MT	2. Çocuklarımızın düşük duyguları hayatlarının diğer bölümlerinde de kendini gösteriyor mu?
ChatGPT	Çocuklarımızın düşük duyguları hayatlarının diğer alanlarında da görülüyor mu?

The translations of "2. Do our children's low feelings show in other parts of their lives?" demonstrate varying degrees of alignment with the MQM model. The human translation, *Çocuklarımızın kendilerini kötü hissetmesi hayatın başka alanlarında da kendini gösteriyor mu?* slightly modifies the original by using *kendilerini kötü hissetmesi* (feeling bad about themselves), which adds a nuance of self-perception not explicitly present in the source text, impacting accuracy and terminology. Google MT's *Çocuklarımızın düşük duyguları hayatlarının diğer alanlarında kendini gösteriyor mu?* and DeepL MT's *2. Çocuklarımızın düşük duyguları hayatlarının diğer bölümlerinde de kendini gösteriyor mu?* both offer more direct translations, closely adhering to the original text in terms of terminology and accuracy. They effectively translate "low feelings" and maintain the

original's focus on various life areas. ChatGPT, *Çocuklarımızın düşük duyguları hayatlarının diğer alanlarında da görülüyor mu?* is also a close translation, maintaining the essence of the source text while slightly altering the phrase structure, which could affect style but stays true to the source in terms of accuracy and terminology. Each translation reflects a different approach in translating the emotional aspect and its impact on various life areas, showcasing the nuances in interpreting and conveying the source text's meaning in Turkish.

Source Text	But our children and teenagers with depression may struggle to find the words to describe their emotions and moods.
Human Translation	Deprasyon geçiren çocuklarımız ve ergenlik çağındaki gençlerimiz duygularını, veya ruh hallerini ifade edemiyebilirler.
Google MT	Ancak depresyonlu çocuklarımız ve gençlerimiz, duygularını ve ruh hallerini tanımlayacak kelimeleri bulmakta zorlanabilirler.
DeepL MT	Ancak depresyondaki çocuklarımız ve gençlerimiz duygularını ve ruh hallerini tarif edecek kelimeleri bulmakta zorlanabilirler.
ChatGPT	Ancak depresyondaki çocuklarımız ve ergenlerimiz, duygularını ve ruh hallerini tarif etmek için sözcükler bulmakta zorlanabilirler.

In the translations of "But our children and teenagers with depression may struggle to find the words to describe their emotions and moods," we see a range of interpretations within the MQM model framework. The human translation, *Deprasyon geçiren çocuklarımız ve ergenlik çağındaki gençlerimiz duygularını, veya ruh hallerini ifade edemiyebilirler*, adds *ergenlik çağındaki* (adolescent age), specifying the teenagers' developmental stage, which slightly deviates from the original text's simplicity, affecting accuracy and terminology. Google MT's *Ancak depresyonlu çocuklarımız ve gençlerimiz, duygularını ve ruh hallerini tanımlayacak kelimeleri bulmakta zorlanabilirler*, and DeepL MT's *Ancak depresyondaki çocuklarımız ve gençlerimiz duygularını ve ruh hallerini tarif edecek kelimeleri bulmakta zorlanabilirler*, are closer translations. Both versions accurately convey the difficulty of describing emotions and moods, maintaining the essence of the

source text. ChatGPT, *Ancak depresyondaki çocuklarımız ve ergenlerimiz, duygularını ve ruh hallerini tarif etmek için sözcükler bulmakta zorlanabilirler*, is also in line with the original, accurately capturing the challenge faced by children and teenagers with depression. Each translation varies in how it interprets and conveys the nuances of emotional expression challenges in children and teenagers with depression, reflecting different aspects of linguistic conventions, terminology, and accuracy in the translation process.

Source Text	We as parents, are often worried when our depressed children and teenagers express feelings and thoughts that "life just isn't worth living", or that "life is so bad I feel like giving up".
Human Translation	Aileler, bunalım geçiren çocuklarıin “hayat yaşamaya değmez” veya “hayat o kadar kötü ki artık yaşamak istemiyorum” gibi düşüncelerini ve duygularını açığa vurduğunda, genelikle çok endişelenirler.
Google MT	Ebeveynler olarak, depresyondaki çocuklarımız ve gençlerimiz "hayat yaşamaya değmez" veya "hayat o kadar kötü ki vazgeçmek istiyorum" şeklinde duygu ve düşünceler ifade ettiğinde genellikle endişeleniriz.
DeepL MT	Depresyondaki çocuklarımız ve gençlerimiz "hayat yaşamaya değmez" ya da "hayat o kadar kötü ki vazgeçmek istiyorum" gibi duygu ve düşüncelerini ifade ettiklerinde ebeveynler olarak genellikle endişeleniriz.
ChatGPT	Biz ebeveynler, depresyonda olan çocuklarımızın ve ergenlerimizin "hayat yaşamaya değmez", "hayat o kadar kötü ki pes etmek istiyorum" gibi düşünceler ifade etmelerinden sıklıkla endişe duyarız.

The translations of "We as parents, are often worried when our depressed children and teenagers express feelings and thoughts that 'life just isn't worth living', or that 'life is so bad I feel like giving up'" illustrate various adaptations to the original text's sentiment and structure within the MQM framework. The human translation, *Aileler, bunalım geçiren çocuklarının “hayat yaşamaya değmez” veya “hayat o kadar kötü ki artık yaşamak istemiyorum”*

gibi düşüncelerini ve duygularını açığa vurduğunda, genelikle çok endişelenirler, alters the perspective from "we as parents" to "families" (*Aileler*), broadening the scope of concern beyond parents, impacting audience appropriateness and terminology. The translation also changes the phrasing of the children's thoughts, affecting accuracy while maintaining the overall sentiment. Google MT's *Ebeveynler olarak, depresyondaki çocuklarımız ve gençlerimiz 'hayat yaşamaya değmez' veya 'hayat o kadar kötü ki vazgeçmek istiyorum' şeklinde duygu ve düşünceler ifade ettiğinde genellikle endişeleniriz* remains closer to the source, accurately translating the feelings and thoughts of depressed children and teenagers, and retaining the parental perspective, aligning well with accuracy and terminology. DeepL MT's *Depresyondaki çocuklarımız ve gençlerimiz 'hayat yaşamaya değmez' ya da 'hayat o kadar kötü ki vazgeçmek istiyorum' gibi duygu ve düşüncelerini ifade ettiklerinde ebeveynler olarak genellikle endişeleniriz* also closely mirrors the original, maintaining the focus on the parents' concern and the children's expressions, preserving the accuracy and style of the original text. ChatGPT, *Biz ebeveynler, depresyonda olan çocuklarımızın ve ergenlerimizin 'hayat yaşamaya değmez', 'hayat o kadar kötü ki pes etmek istiyorum' gibi düşünceler ifade etmelerinden sıklıkla endişe duyarız*, while closely adhering to the original, slightly alters the expression of the children's thoughts for clarity in Turkish, affecting the style but maintaining overall accuracy. Each translation reflects a different approach in translating the emotional context and perspective, showcasing the nuances in conveying the source text's meaning and the concerns of parents regarding their depressed children and teenagers.

Source Text	Hearing our children say they wish they were dead, or hearing them talk about suicide, is often overwhelming.
Human Translation	Bir aile için çocuğunun ölümü dilemesini veya intihardan bahsetmesini dinlemek, dayanılamayacak kadar zordur.
Google MT	Çocuklarımızın keşke ölseydim dediklerini duymak ya da intihar hakkında konuştuklarını duymak genellikle bunaltıcıdır.
DeepL MT	Çocuklarımızın keşke ölseydim dediklerini duymak ya da

	intihardan bahsettiklerini duymak çoğu zaman bunaltıcıdır.
ChatGPT	Çocuklarımızın ölmeyi dilediklerini veya intihar hakkında konuştuklarını duymak, genellikle ezici olabilir.

The translations of "Hearing our children say they wish they were dead, or hearing them talk about suicide, is often overwhelming" demonstrate varied approaches and interpretations within the MQM framework. The human translation, *Bir aile için çocuğunun ölümü dilemesini veya intihardan bahsetmesini dinlemek, dayanılamayacak kadar zordur*, shifts the focus slightly by starting with *Bir aile için* (For a family), which broadens the perspective beyond just the parents. This alters the audience appropriateness and may affect the terminology by emphasizing the family unit's experience. The phrase *dayanılamayacak kadar zordur* (unbearably hard) also intensifies the emotional response compared to the original. Google MT's *Çocuklarımızın keşke ölseydim dediklerini duymak ya da intihar hakkında konuştuklarını duymak genellikle bunaltıcıdır* and DeepL MT's *Çocuklarımızın keşke ölseydim dediklerini duymak ya da intihardan bahsettiklerini duymak çoğu zaman bunaltıcıdır* both provide close translations. They maintain the focus on the parents' perspective and accurately convey the emotional weight of hearing such statements from their children, aligning well with the source in terms of accuracy and style. ChatGPT, *Çocuklarımızın ölmeyi dilediklerini veya intihar hakkında konuştuklarını duymak, genellikle ezici olabilir*, also stays true to the original, capturing the overwhelming nature of such experiences for parents. The use of *ezici* (overwhelming) closely mirrors the original's sentiment, maintaining accuracy and adhering to the style and linguistic conventions of the source text. Each translation reflects a different angle in addressing the heavy emotional context, showing the complexity of translating such sensitive content while maintaining the essence and impact of the original message.

Source Text	Not talking about it will not make it go away.
Human Translation	Bundan bahsetmiyor olmaları, bu riski ortadan kaldırmaz.
Google MT	Bunun hakkında konuşmamak onu ortadan

	kaldırmayacak.
DeepL MT	Bu konuda konuşmamak onu ortadan kaldırmayacaktır.
ChatGPT	Konuşmamak, ortadan kalkmasını sağlamaz.

The translations of "Not talking about it will not make it go away" showcase different interpretations and nuances within the framework of the MQM model. The human translation, *Bundan bahsetmiyor olmaları, bu riski ortadan kaldırmaz*, shifts the subject from a general, implicit "we" or "one" to a more specific "they," changing the focus from the act of talking to the subjects' choice not to talk about it. This affects the accuracy and may influence the audience appropriateness by implying a different group (possibly the children) as the subject, rather than addressing a general guideline or principle. Google MT's *Bunun hakkında konuşmamak onu ortadan kaldırmayacak* and DeepL MT's *Bu konuda konuşmamak onu ortadan kaldırmayacaktır* are closer to the source text. Both maintain the original sentence's impersonal structure and accurately convey the idea that avoiding discussion will not eliminate the problem. The choice of phrasing in these translations preserves the style and linguistic conventions of the original text. ChatGPT, *Konuşmamak, ortadan kalkmasını sağlamaz*, simplifies the sentence structure while keeping the core message intact. This version maintains the impersonal tone and succinctly conveys the idea that silence won't solve the issue, aligning well with the source in terms of accuracy and style. Each translation reflects different approaches to conveying the message about the ineffectiveness of silence as a solution, highlighting the intricacies of translating concise statements while preserving their intended meaning and impact.

In Table 2, "Comparison of Translation Performance in Educational Texts," we present a comprehensive comparative analysis of various translations of educational texts. This analysis is conducted in light of the MQM model parameters, which include Terminology, Accuracy, Linguistic Conventions, Style, Locale Conventions, Audience Appropriateness, Design and Markup. The table presents clear examples of the source text alongside translations provided by human translators, Google MT, DeepL, and ChatGPT. Each translation is meticulously evaluated against

the MQM parameters, offering insights into their respective strengths and weaknesses in the context of educational content.

Table 2: Comparison of Translation Performance in Educational Texts

Parameter	**Human Translation**	**Google Translation**	**DeepL Translation**	**ChatGPT Translation**
Terminology	Excellent, context-aware	Good, but occasionally inaccurate	Good, occasionally lacks nuance	Excellent, context-appropriate
Accuracy	High, with nuanced translation	Mostly accurate with minor errors	Mostly accurate with some deviations	High, closely reflects source
Linguistic Conventions	Adheres well to target language norms	Struggles with complex structures	Better than Google but can be awkward	Good grasp, maintains readability
Style	Consistent, appropriate for content	Inconsistent, varies in tone	Generally consistent, varies unexpectedly	Consistent, suitable for educational material
Locale Conventions	Adheres to locale-specific nuances	Sometimes misses locale nuances	Occasional misses in locale-specific references	Generally maintains locale nuances
Audience Appropriateness	Tailored to target audience, appropriate tone	Mostly appropriate, but can misstep in tone	Mostly appropriate with minor lapses	Suitable language and tone for audience
Design and	Likely	Basic, may	Similar to	Focuses on

markup	respects original design intent	not consider design aspects	Google, basic design consideration	content over design

Human Translation is deemed excellent, demonstrating high accuracy and context awareness in terminology. It adheres well to the target language's linguistic conventions and is consistent in style, and suitable for the content's intent. Locale-specific nuances are well respected, ensuring the content is tailored for the audience with an appropriate tone. The design and markup likely respect the original intent, maintaining the integrity of the source. Google Translation is reliable but sometimes falls short in terms of accuracy, particularly with complex structures and occasionally misses nuances. The style can be inconsistent and vary in tone, and while it's mostly appropriate for the audience, it can misstep in tone. The design and markup are basic and may not consider all design aspects. DeepL Translation scores well on terminology and is mostly accurate but can sometimes lack nuances. It has better linguistic convention adherence than Google but can still be awkward. The style is generally consistent but can vary unexpectedly. Locale conventions are occasionally missed, and while the translation is mostly appropriate for the audience, there are minor lapses. Design and markup are similar to Google, with some basic design considerations. ChatGPT Translation provides excellent, context-appropriate terminology, and high accuracy that closely reflects the source material. It has a good grasp of linguistic conventions and maintains readability, with a consistent style suitable for educational material. It generally maintains locale nuances and offers language and tone that are suitable for the audience, focusing on content over design consideration. In summary, while Human Translation leads in context-awareness and nuanced translation, Google and DeepL provide reliable but sometimes inconsistent alternatives. ChatGPT strikes a balance with high accuracy and suitability for educational content, with a focus on language over design. Each method has its own strengths

and weaknesses, with human translation being the most nuanced and machine translations offering varying degrees of accuracy and adherence to linguistic and locale nuances. In a nutshell, human translations generally perform best across all parameters, offering the most reliable and context-sensitive translation. Google MT and DeepL provide decent translations but may occasionally miss nuances and locale-specific elements. ChatGPT, as represented here, offers a balanced translation with high accuracy and appropriateness for the target audience but does not inherently account for design and markup considerations.

RQ2. How does the translation performance of human translators, Google MT, DeepL, and ChatGPT compare when applied to health-related texts?

Table 3. Side-by-Side Comparison of Human, Google, DeepL, and ChatGPT Translations for Healthcare Context

Source Text	The Health Care Complaints Commission
Human Translation	-
Google MT	Sağlık Şikayetleri Komisyonu
DeepL MT	Sağlık Hizmetleri Şikayetleri Komisyonu
ChatGPT	Sağlık Hizmetleri Şikayetler Komisyonu

In the translations of "The Health Care Complaints Commission" into Turkish by Google MT, DeepL MT, and ChatGPT, differences are primarily in terms of the MQM model parameters of Terminology and Accuracy. While all translations use appropriate healthcare terminology, Google MT's *Sağlık Şikayetleri Komisyonu* omits the direct translation of "Care," slightly deviating from the source text's meaning. In contrast, DeepL MT's *Sağlık Hizmetleri Şikayetleri Komisyonu* and ChatGPT's translation include *Hizmetleri* (Services), closely aligning with the original term "Health Care." Linguistically, all versions adhere to Turkish conventions, with variations in word order reflecting different structuring approaches while maintaining grammatical correctness. The style remains consistently formal across all translations, suitable for an official body's name. Each translation respects locale conventions, using standard Turkish terminology

for health, complaints, and commission, and is appropriately targeted at a Turkish-speaking audience familiar with healthcare-related terms. Design and Markup aspects are not applicable in this context as the task involves plain text translation without specific design or markup elements. Overall, the most significant difference lies in the Accuracy parameter, with Google MT's version slightly less aligned with the source text in terms of conveying the full scope of "Health Care."

Source Text	Making a complaint
Human Translation	Şikayette bulunmadan önce
Google MT	Şikayet etmek
DeepL MT	Şikayette bulunmak
ChatGPT	Bir şikayette bulunma

In the translations of "Making a complaint" into Turkish by Google MT, DeepL MT, ChatGPT, and a human translator, differences emerge primarily in terms of Accuracy and Audience Appropriateness within the MQM model parameters. Google MT's *Şikayet etmek* and DeepL MT's *Şikayette bulunmak* closely mirror the source text's intent, presenting direct and concise equivalents for "Making a complaint." ChatGPT's *Bir şikayette bulunma* adds an unnecessary definite article *Bir* (A), which slightly deviates from the source text's conciseness but still retains the overall meaning. However, the human translation *Şikayette bulunmadan önce* introduces a significant shift in meaning, translating to "Before making a complaint," which adds a temporal aspect which does not present in the original. This reflects a notable deviation in Accuracy. All translations are correct in terms of Linguistic Conventions and use appropriate, formal Style suitable for the context. Locale Conventions are well-respected with appropriate terminology. Design and Markup do not apply in this textual translation context. In summary, while Google MT and DeepL MT provide accurate and audience-appropriate translations, ChatGPT's version includes an unnecessary article, and the human translation significantly alters the original meaning by adding a temporal context.

Source Text	Suburb/Town
Human Translation	-
Google MT	Banliyö/Kasaba
DeepL MT	Banliyö/Kasaba
ChatGPT	İlçe/Şehir

In translating "Suburb/Town" into Turkish, there's a notable divergence in approach between Google MT, DeepL MT, and ChatGPT. Both Google MT and DeepL MT offer the same translation, *Banliyö/Kasaba*, which accurately reflects the source text's meaning. *Banliyö* corresponds to "Suburb," and *Kasaba* to "Town," demonstrating a high level of Accuracy in terms of the MQM model. These translations are straightforward and adhere to appropriate Linguistic Conventions and Style, fitting for various contexts, whether formal or informal. Locale Conventions are also well respected, using standard Turkish terms for geographical locations. On the other hand, ChatGPT's translation, *İlçe/Şehir*, represents a shift in meaning. *İlçe* translates to "District" and *Şehir* to "City," which deviates from the original terms "Suburb" and "Town." This reflects a difference in Accuracy, as the terms used by ChatGPT refer to different types of urban areas compared to the source text. While this translation maintains appropriate Linguistic Conventions and Style, and is suitable for a Turkish audience, the choice of words alters the original meaning, impacting its Audience Appropriateness. Design and Markup are not applicable in this context, as the task is centered around text translation. In summary, Google MT and DeepL MT provide translations that are closely aligned with the source text in terms of Terminology and Accuracy, while ChatGPT's translation, though linguistically correct, deviates in meaning from the original terms.

Source Text	State
Human Translation	Eyalet
Google MT	Durum

DeepL MT	Eyalet
ChatGPT	Eyalet

In the translations of the word "State" into Turkish by Google MT, DeepL MT, ChatGPT, and a human translator, we observe significant variation primarily in terms of Accuracy within the MQM model. DeepL MT, ChatGPT, and the human translation all provide *Eyalet* as the translation, accurately reflecting the geopolitical context of the term "State," as in a region or province within a country. This shows high fidelity to the source text's intended meaning, maintaining appropriate Terminology and Linguistic Conventions. In contrast, Google MT translates "State" as *Durum*, which means "condition" or "situation" in Turkish. This represents a significant deviation in terms of Accuracy, as *Durum* does not convey the geopolitical meaning of "State." The Google MT translation is contextually inappropriate for the intended use of the term "State," showing a misunderstanding of the source text's context. All translations are correct in terms of Linguistic Conventions and use a Style that is fitting for the context. Locale Conventions are well-respected in the translations by DeepL MT, ChatGPT, and the human translator, using standard Turkish terminology for geopolitical entities. Design and Markup are not applicable in this context as the task involves plain text translation. In summary, while DeepL MT, ChatGPT, and the human translation accurately capture the meaning of "State" as a geopolitical entity, Google MT's translation misinterprets the term, reflecting a significant difference in Accuracy.

Source Text	I am an Aboriginal person Yes / No
Human Translation	-
Google MT	Aborjin bir insanım Evet / Hayır
DeepL MT	Ben bir Aborjinim Evet / Hayır
ChatGPT	Bir Aborijinim Evet / Hayır

In translating "I am an Aboriginal person Yes / No" into Turkish, Google MT, DeepL MT, and ChatGPT exhibit variations primarily in Accuracy and Style within the MQM model. Google MT's translation, *Aborjin bir insanım Evet / Hayır,* DeepL MT's *Ben*

bir Aborjinim Evet/ Hayır, and ChatGPT's *Bir Aborijinim Evet / Hayır* all correctly identify *Aborjin* as the Turkish equivalent for "Aboriginal." This demonstrates a high level of Accuracy in terms of terminology and meaning. However, there are stylistic differences. Google MT adopts a more literal translation style, including *bir insanım* ("a person I am"), which, while accurate, is slightly more verbose than necessary. DeepL MT and ChatGPT opt for more concise translations, with DeepL MT using *Ben* (I) for clarity and ChatGPT omitting it, relying on the verb conjugation to convey the subject. These variations reflect different approaches to maintaining the balance between conciseness and clarity in Style. All translations adhere to appropriate Linguistic Conventions in Turkish and respect Locale Conventions. The translations are also audience-appropriate, providing a clear option for self-identification regarding Aboriginal identity. Design and Markup are not relevant in this textual translation context. In summary, while all translations accurately convey the message of self-identifying as an Aboriginal person, they differ in Style, with Google MT being more literal and explanatory, whereas DeepL MT and ChatGPT offer more streamlined translations. These stylistic choices do not alter the fundamental meaning but reflect different approaches to translation.

In Table 4, "Comparison of Translation Performance in Health-Related Texts," we present a comprehensive comparative analysis of various translations of the health text. This analysis is conducted in light of the MQM model parameters, which include Terminology, Accuracy, Linguistic Conventions, Style, Locale Conventions, Audience Appropriateness, Design and Markup. The table presents clear examples of the source text alongside translations provided by human translators, Google MT, DeepL, and ChatGPT. Each translation is meticulously evaluated against the MQM parameters, offering insights into their respective strengths and weaknesses in the context of health content.

Table 4: Comparison of Translation Performance in Medical Texts

Parameter	Human Translation	Google Translation	DeepL Translation	ChatGPT Translation

Terminology	Precise and appropriate terminology consistent with healthcare complaints	Accurate terminology with few exceptions	Accurate, consistent use of terms	Accurate, consistent use of terms
Accuracy	Generally accurate, though simplifies some phrases which could alter nuanced meaning	Mostly accurate, some deviations especially with nuanced phrases	Generally accurate, minor deviations in complex phrases	Generally accurate, slight deviations from formal expressions
Linguistic Conventions	Adheres to Turkish linguistic norms but may need adjustments for complex structures	Correct grammar and syntax, minor errors with complex structures	Good grammar, better syntax handling than Google	Good grammar and syntax, slight informal tone
Style	Mostly formal, with minor inconsistencies in register	Mostly formal, some inconsistencies with the use of honorifics	Consistent formality and use of honorifics	Generally formal, some phrases less formal
Locale Conventions	Successfully adapts to local context with appropriate terminology.	Correct locale-specific terms but minor errors in cultural phrasing	Accurate locale-specific terms, good cultural phrasing	Accurate locale-specific terms, good cultural phrasing
Audience Appropriateness	Language is appropriate for general audience but may	Language is appropriate for the general audience,	Language well-tailored to audience, clear and	Language well-tailored to audience, occasionally

	require refinement for formality.	minor clarity issues	precise	less formal
Design and markup	Translation text requires proper formatting to match the source design and layout.	Correct format, some inconsistencies in layout elements	Correct format, good attention to layout elements	Correct format, good attention to layout elements

Human Translation scores well on terminology, with precise and appropriate use, and generally maintains accuracy, although some simplifications could alter nuanced meanings. It adheres to Turkish linguistic conventions but may require adjustments for complex structures. The style is mostly formal, locale conventions are well adapted, and the audience appropriateness is generally reliable but may require refinement for formality. Design and markup need proper formatting to match source document layout. Google Translation has accurate terminology with few exceptions, with some accuracy deviations on nuanced phrases. Its linguistic conventions have corrected grammar and syntax but struggle with complex structures. The style is formal with some inconsistencies, and locale-specific terms have minor errors. The language is appropriate for a general audience, and the design and markup have some inconsistencies in layout. DeepL Translation offers accurate and consistently used terminology, with minor accuracy deviations in complex phrases. The linguistic conventions exhibit good grammar and better syntax handling than Google, with a consistent formal style and use of honorifics. Locale conventions are accurate with good cultural phrasing, audience appropriateness is well-tailored and clear, and the design and markup pay good attention to layout elements. ChatGPT Translation also uses accurate and consistently used terminology, with general accuracy and slight deviations from formal expressions. The linguistic conventions show good grammar and

syntax with an informal tone. Style is generally formal with some phrases less formal, locale conventions are accurate, and the language is well-tailored to the audience, occasionally less formal. The design and markup maintain a good layout. Overall, while each translation method has its strengths, human translation tends to offer more precise terminology and adapts better to local contexts. Google and DeepL have their own merits in terms of grammar and cultural phrasing, with DeepL slightly ahead in handling complex syntax. ChatGPT provides a balance with consistently accurate terminology and a well-tailored audience approach, although it may have a more informal tone at times.

RQ3. How does the translation performance of human translators, Google MT, DeepL, and ChatGPT compare when applied to legal texts?

Table 5: Side-by-Side Comparison of Human, Google, DeepL, and ChatGPT Translations for Legal Text

Source Text	Happily Ever… Before and After
Human Translation	Önce ve Sonra… Her zaman mutlu
Google MT	Sonsuza Dek Mutlu… Öncesi ve Sonrası
DeepL MT	Sonsuza Kadar Mutlu... Öncesi ve Sonrası
ChatGPT	Mutlu Sonsuz... Önce ve Sonra

In the translations of "Happily Ever… Before and After" into Turkish by a human translator, Google MT, DeepL MT, and ChatGPT, we observe variations in Style, Accuracy, and Terminology within the MQM model. The human translation *Önce ve Sonra… Her zaman mutlu* restructures the phrase significantly, placing emphasis on "Before and After" and ending with "Always happy," which changes the poetic structure and rhythm of the original. Google MT's *Sonsuza Dek Mutlu… Öncesi ve Sonrası* and DeepL MT's *Sonsuza Kadar Mutlu... Öncesi ve Sonrası* are more accurate, preserving the essence of "Happily Ever" with *Sonsuza Dek Mutlu* and *Sonsuza Kadar Mutlu*, which both mean "Happy Forever." These translations maintain the original's emphasis on a timeless state of happiness, followed by a reference to "Before and After." ChatGPT's *Mutlu Sonsuz... Önce ve Sonra* offers a slightly

different take, using *Mutlu Sonsuz* ("Happy Infinite") which, while poetic, slightly deviates from the traditional phrase "Happily Ever." In terms of Linguistic Conventions, all translations are well-structured in Turkish. Locale Conventions are respected in all versions, using culturally appropriate expressions for conveying the concept of enduring happiness. Design and Markup do not apply in this text-based translation. Overall, while Google MT and DeepL MT closely align with the original in terms of capturing the poetic and timeless nature of the phrase, the human translation alters the structure significantly, and ChatGPT offers a unique but slightly less conventional rendition.

Source Text	On your wedding day, your celebrant will solemnise your marriage.
Human Translation	Evlenme gününüzde, evlendirme memurunuz evliliğinizi resmen icra edecektir.
Google MT	Düğün gününüzde, kutlayıcınız nikahınızı kıyacak.
DeepL MT	Düğün gününüzde, nikah memurunuz evliliğinizi resmileştirecektir.
ChatGPT	Düğün gününüzde, nikah memurunuz evliliğinizi gerçekleştirir.

In the translations of "On your wedding day, your celebrant will solemnise your marriage" into Turkish by a human translator, Google MT, DeepL MT, and ChatGPT, there are notable differences in terms of Accuracy, Terminology, and Style within the MQM model. The human translation, *Evlenme gününüzde, evlendirme memurunuz evliliğinizi resmen icra edecektir*, is quite accurate and formal, using *evlendirme memurunuz* (your marriage officer) and *resmen icra edecektir* (will officially perform) which closely align with the solemnity and formality of "celebrant" and "solemnise." Google MT's translation, *Düğün gününüzde, kutlayıcınız nikahınızı kıyacak*, uses *kutlayıcınız* (your celebrator) which is less formal and slightly deviates from the original term "celebrant." The term *nikahınızı kıyacak* (will marry you) is more casual and less precise compared to the term "solemnise." DeepL MT's *Düğün gününüzde, nikah memurunuz evliliğinizi resmileştirecektir*, and ChatGPT's

Düğün gününüzde, nikah memurunuz evliliğinizi gerçekleştirir, both use *nikah memurunuz* (your marriage officer), which is accurate and formal. DeepL's *resmileştirecektir* (will formalize) is very close to "solemnise," maintaining the formality and legality implied in the source. ChatGPT's choice, *gerçekleştirir* (performs), while accurate, is slightly less formal and specific than "solemnise." All translations adhere to appropriate Linguistic Conventions and Locale Conventions, with varying degrees of formality and precision in style. Design and Markup do not apply in this text-based translation. In summary, the human translation and DeepL MT provide more accurate and formal translations that closely reflect the source text's intent, while Google MT and ChatGPT offer translations that, while correct, are somewhat less formal and less precise in their terminology.

Source Text	Making a will
Human Translation	Vasiyetname yapma
Google MT	vasiyet yapmak
DeepL MT	Vasiyetname hazırlamak
ChatGPT	Vasiyetname yapma

In the translations of "Making a will" into Turkish by a human translator, Google MT, DeepL MT, and ChatGPT, we see variations mainly in terms of Style and Terminology within the MQM model. Both the human translation *Vasiyetname yapma* and ChatGPT's translation use the same phrasing, which is a direct and concise representation of the original text. The term *Vasiyetname* accurately translates into "will," and *yapma* corresponds to "making," thus maintaining high Accuracy and appropriate Terminology. Google MT's translation *vasiyet yapmak* uses a less formal term *vasiyet* instead of *Vasiyetname*. While *vasiyet* can mean a will or testament, *Vasiyetname* is more specific and formal, better suiting the legal context of making a will. DeepL MT's *Vasiyetname hazırlamak* adds *hazırlamak* (to prepare), which introduces a slight variation in Style, implying a more detailed or thorough process of creating a will. All translations adhere to Turkish Linguistic Conventions and respect Locale Conventions,

with slight stylistic differences reflecting the translators' choices. Design and Markup are not relevant in this context, as the task involves plain text translation. In summary, while the human translation and ChatGPT provide a direct and succinct translation of "Making a will," Google MT opts for a less formal term, and DeepL MT suggests a slightly more elaborate process through its choice of words, reflecting minor differences in Style and Terminology.

Source Text	Keeping relationships on track is not always easy.
Human Translation	İlişkileri rayında tutmak her zaman kolay değildir.
Google MT	İlişkileri yolunda tutmak her zaman kolay değildir.
DeepL MT	İlişkileri rayında tutmak her zaman kolay değildir.
ChatGPT	İlişkileri rayında tutmak her zaman kolay olmaz.

In translating "Keeping relationships on track is not always easy" into Turkish, the human translator, Google MT, DeepL MT, and ChatGPT exhibit nuanced differences primarily in Style and Accuracy within the MQM model. The human translation, DeepL MT, and ChatGPT all opt for *İlişkileri rayında tutmak*, which is a direct translation of "Keeping relationships on track," preserving the metaphorical use of "on track." This demonstrates high Accuracy and appropriate Terminology. Google MT's *İlişkileri yolunda tutmak* translates to "Keeping relationships in order," which slightly shifts the original metaphor from "track" to "order." While the overall meaning of maintaining relationships is preserved, the change in metaphor represents a minor deviation in Style and a slight impact on Accuracy. The variation in the latter part of the sentence between *her zaman kolay değildir* (is not always easy) and *her zaman kolay olmaz* (does not always become easy) is subtle. The human translator, DeepL MT, and Google MT use the former, directly reflecting the source text. ChatGPT's version, while conveying a similar meaning, introduces a slight variation in phrasing, which affects Style but not the fundamental meaning. All translations adhere to Turkish Linguistic Conventions and Locale Conventions, with variations reflecting different stylistic choices and slight nuances in expression. Design

and Markup are not relevant in this text-based translation. In summary, while the human translator, DeepL MT, and ChatGPT provide translations that are very close to the source text in both meaning and metaphor, Google MT opts for a slightly different metaphor, reflecting a minor difference in Style and Accuracy.

Source Text	Marriage breakdown: Family Dispute Resolution
Human Translation	Evliliğin bitmesi: Aile Uyuşmazlığının Çözümü
Google MT	Evlilik dökümü: Aile Uyuşmazlık Çözümü
DeepL MT	Evliliğin bozulması: Aile Uyuşmazlıklarının Çözümü
ChatGPT	Evlilik çökmeleri: Aile Anlaşmazlık Çözümü

In the translations of "Marriage breakdown: Family Dispute Resolution" into Turkish by a human translator, Google MT, DeepL MT, and ChatGPT, we observe differences mainly in Terminology and Accuracy within the MQM model. The human translation *Evliliğin bitmesi: Aile Uyuşmazlığının Çözümü* translates "Marriage breakdown" as *Evliliğin bitmesi* (The ending of marriage), which captures the finality implied in "breakdown," but lacks the connotation of a process of deterioration. The phrase *Aile Uyuşmazlığının Çözümü* correctly translates to "Family Dispute Resolution." Google MT's *Evlilik dökümü: Aile Uyuşmazlık Çözümü* uses *Evlilik dökümü*, a less common phrase that could be interpreted as "Marriage breakdown," but with a literal sense of 'casting' or 'molding' which is less accurate. The term *Aile Uyuşmazlık Çözümü* for "Family Dispute Resolution" is slightly imprecise as it misses the possessive 's' (*Uyuşmazlıklarının*). DeepL MT's *Evliliğin bozulması: Aile Uyuşmazlıklarının Çözümü* translates "Marriage breakdown" more accurately as *Evliliğin bozulması* (The deterioration of marriage), which better conveys the gradual process of a marriage breakdown. The term *Aile Uyuşmazlıklarının Çözümü* is an accurate translation for "Family Dispute Resolution." ChatGPT's *Evlilik çökmeleri: Aile Anlaşmazlık Çözümü* uses *Evlilik çökmeleri* (Marriage collapses), a term that implies a more sudden or dramatic breakdown, which may not fully align with the usual connotation of "Marriage breakdown."

The phrase *Aile Anlaşmazlık Çözümü* is similar to Google MT's translation and has the same minor inaccuracy. All translations maintain appropriate Linguistic Conventions and Style, and Locale Conventions are respected. Design and Markup are not applicable in this context. In summary, DeepL MT provides the most accurate translation of both terms, closely aligning with the source text, while the human translation, Google MT, and ChatGPT each introduce slight variations in meaning and terminology, affecting the overall Accuracy.

In Table 6, "Comparison of Translation Performance in Legal Texts," we present a comprehensive comparative analysis of various translations of the legal text. This analysis is conducted in light of the MQM model parameters, which include Terminology, Accuracy, Linguistic Conventions, Style, Locale Conventions, Audience Appropriateness, Design and Markup. The table presents clear examples of the source text alongside translations provided by human translators, Google MT, DeepL, and ChatGPT. Each translation is meticulously evaluated against the MQM parameters, offering insights into their respective strengths and weaknesses in the context of legal content.

Table 6: Comparison of Translation Performance in Legal Texts

Parameter	Human Translation	Google Translation	DeepL Translation	ChatGPT Translation
Terminology	Correctly uses legal terms such as "subsection" and "regulation".	Correctly uses legal terms but slightly varies with "alt bölümü" for "subsection".	Correctly uses legal terms, similar to Google MT.	Correctly uses legal terms but slightly varies with "alt bölümü" for "subsection".
Accuracy	Accurately conveys the meaning of the source with all legal references	Accurately conveys the meaning but uses "Bu bilgi" (This information)	Accurately conveys the meaning, including the plural form as in	Accurately conveys the meaning, includes all legal references,

	intact.	in singular form instead of "Bu bilgiler" (This information) in plural as in the source.	the source.	slightly different phrasing.
Linguistic Conventions	Follows conventions of Turkish legal language.	Follows conventions of Turkish legal language but with a minor deviation in the term for "subsection" .	Follows conventions of Turkish legal language, similar to Google MT.	Follows conventions of Turkish legal language, with slight deviation as noted in terminology.
Style	The translation is formal and matches the source's legal document style.	The translation is formal and matches the source's style, though the term choice for "subsection" is less formal.	The translation is formal and matches the source's style.	The translation is formal, consistent with the source style.
Locale Conventions	Uses the structure and terms appropriate for Turkish legal documents.	Uses the structure and terms appropriate for Turkish legal documents.	Uses the structure and terms appropriate for Turkish legal documents.	Uses structure and terms appropriate for Turkish legal documents, but slightly less common phrasing.
Audience Appropriateness	Language is appropriate for an	Language is appropriate for an	Language is appropriate for an	Language is appropriate for an

	audience familiar with legal texts.	audience familiar with legal texts.	audience familiar with legal texts.	audience familiar with legal texts.
Design and markup	N/A	N/A	N/A	N/A

Human Translation is noted for correctly using legal terms and accurately conveying the meaning of the source with all legal references intact. It follows the conventions of Turkish legal language, is formal, and maintains the structure and terms appropriate for Turkish legal documents. The language is suitable for an audience familiar with legal texts. Google Translation also uses legal terms correctly but with slight variations, such as "alt bölümü" for "subsection". While it accurately conveys the meaning, it incorrectly uses singular forms for words that are plural in the source. It follows the conventions of Turkish legal language with minor deviations, is formal, and uses appropriate structures and terms for Turkish legal documents. The language is deemed appropriate for an audience familiar with legal texts. DeepL Translation similarly uses the correct legal terms, closely resembling Google Translation's performance. It accurately conveys the meaning, including the correct form as in the source. DeepL follows the conventions of Turkish legal language better than Google but can still be awkward. Its style is formal and matches the source's style, and it uses the correct structure and terms suitable for Turkish legal documents. The audience appropriateness is similar to the other methods. ChatGPT Translation uses the correct legal terms but with slight variations, such as *bölümü* for "subsection". It accurately conveys the meaning and includes all legal references, albeit with slightly different phrasing. ChatGPT follows the conventions of Turkish legal language with minor deviations in terminology. The translation is formal and consistent with the source style, uses structure and terms appropriate for Turkish legal documents, although it may include common phrasing. The language is appropriate for an audience familiar with legal texts. Design and markup parameters are not applicable (N/A) for machine

translations. Overall, while all methods demonstrate proficiency in translating legal terminology and maintaining formal style, Human Translation holds the edge for accuracy and adherence to linguistic conventions. Google and DeepL are comparable, with DeepL having a slight advantage in handling the nuances of the source language. ChatGPT, while generally effective, exhibits slight deviations in terminology and phrasing but remains suitable for an audience familiar with legal texts.

RQ4. What are the specific strengths and limitations of each translation method in the context of the three text types?

The analysis of the translations provided in the legal, health, and general text types using the MQM model reveals various strengths and weaknesses in each translation method (human translator, Google, DeepL, and ChatGPT).

1. Terminology:

- Human Translator: Consistently accurate and context-appropriate terminology.
- Google, DeepL, and ChatGPT: Generally accurate, but some instances of inconsistency with terminology resources were noticed. Google and DeepL showed occasional wrong term usage, particularly in specialized contexts like legal or health.

2. Accuracy:

- Human Translator: Demonstrated high accuracy with minimal mistranslations, omissions, or additions.
- Google, DeepL, and ChatGPT: Varied in accuracy. Google and DeepL occasionally suffered from mistranslations or omissions, especially in complex sentences. ChatGPT showed better handling of under-translation and over-translation but was not immune to occasional errors.

3. Linguistic Conventions:

- Human Translator: Excellent adherence to grammar, punctuation, and spelling norms.
- Google, DeepL, and ChatGPT: Generally good but not perfect. Google had occasional issues with grammar and

punctuation. DeepL and ChatGPT were better but sometimes produced awkward or unnatural sentence structures.

4. Style:

- Human Translator: Maintained a consistent and appropriate style according to the text type.
- Google, DeepL, and ChatGPT: Style varied. Google sometimes produced awkward or unidiomatic expressions. DeepL was better in maintaining register and style, while ChatGPT occasionally struggled with maintaining a consistent style across different text types.

5. Locale Conventions:

- Human Translator: Excellent adherence to locale-specific formats like date, time, and address.
- Google, DeepL, and ChatGPT: Generally good adherence, but some errors were noticed, particularly in handling specific formats like addresses or phone numbers.

6. Audience Appropriateness:

- Human Translator: Showed a high level of audience appropriateness, with culturally sensitive and relevant translations.
- Google, DeepL, and ChatGPT: Varied in performance. Google sometimes missed cultural nuances. DeepL and ChatGPT were better but not flawless in capturing culture-specific references.

7. Design and Markup:

- Human Translator: Not applicable as human translations typically don't involve design and markup elements.
- Google, DeepL, and ChatGPT: Generally good, but there were instances where layout considerations (like line breaks in paragraphs) could be improved.

In conclusion, the human translations were consistently high in quality across all dimensions, reflecting a deep understanding of context, style, and locale-specific nuances. Google's translations were adequate but occasionally lacked in areas like style and cultural nuances. DeepL showed strengths in maintaining style

and linguistic conventions but sometimes faltered in terminology accuracy. ChatGPT demonstrated a balanced performance across most dimensions, with occasional issues in style consistency and handling complex sentence structures. Each translation method has its unique strengths and challenges, and the choice between them may depend on the specific requirements of the translation task, such as the need for accuracy, adherence to style guides, or sensitivity to cultural nuances.

RQ5. How do the findings contribute to the broader understanding of machine translation performance and its potential impact on the field of translation?

The comparison of a human translator, Google, DeepL, and ChatGPT in a variety of community-based settings (legal, health, and general) reveals distinct strengths and weaknesses for each translator, indicating their suitability for specific kinds of texts:

The human translator is highly successful in all text types, particularly in legal and health texts, where precision, context understanding, and sensitivity to specialized terminology are crucial. Human translators are indispensable for complex and nuanced texts, especially where legal implications or technical accuracy are critical. Their ability to understand context and cultural nuances makes them ideal for texts requiring a high degree of precision and cultural sensitivity.

Google MT generally performs well with general and straightforward texts. It has improved over time in handling common phrases and simple sentences. Google MT is useful for quick translations of less complex texts or for getting a general understanding of content in a foreign language. However, its occasional inaccuracies in specialized terminology and style mean it's less reliable for legal, technical, or health-related texts.

DeepL tends to be more successful with texts that require a more nuanced understanding of language, such as general and literary texts. It often produces more naturally flowing translations than Google. DeepL is a strong choice for texts where a more natural, idiomatic translation is needed. While it shows

competence in general texts, its occasional struggles with very specialized terminology can be a limitation for highly technical or legal texts.

ChatGPT shows balanced performance across various text types. It is particularly effective in general texts where conversational tone and context are important. ChatGPT is versatile, suitable for a broad range of texts, especially where a conversational, engaging style is needed. However, its occasional style inconsistencies and less rigorous handling of specialized terminology make it less ideal for highly technical or legal documents.

Human Translators remain the most reliable for high-stakes, technical, or specialized texts due to their nuanced understanding of language, culture, and subject matter. Automated Translation Tools (Google, DeepL, ChatGPT) are improving and useful for general or preliminary translations, especially when speed is a priority. However, they may still require human review for accuracy, particularly in specialized fields. The choice between these translation options should be guided by the nature of the text, the required level of accuracy, and the specific context in which the translation will be used.

Discussion

In this study, we compared the translation performance of human translators, Google MT, DeepL, and ChatGPT in educational, health, and legal texts.

As expected, human translations have delivered the most accurate and contextually appropriate results. This is because humans are capable of understanding the nuances of language, idiomatic expressions, and cultural references. However, human translations can sometimes suffer from inconsistencies and may be influenced by the translator's personal preferences or interpretations. Google Translation has significantly improved over time and offers fast, accessible translations for a wide range of languages. However, it can sometimes struggle with idiomatic expressions, complex sentences, or preserving the original meaning of the source text. This may lead to fluency, terminology,

and mistranslation issues. DeepL has shown impressive performance in generating translations that are often more fluent and contextually accurate than those of Google MT. It excels in handling idiomatic expressions and complex sentence structures. However, it may still occasionally produce mistranslations or struggle with preserving the original meaning, especially in highly specialized or technical domains. As a language model, ChatGPT provides translations that are generally accurate and fluent, with a good understanding of context and idiomatic expressions. However, it can sometimes generate translations that deviate from the source text's meaning or introduce inaccuracies, particularly in specialized fields.

In confirmation of our findings, Rusadi and Setiajid (2023), Rusadi and Setiajid (2023) identified five specific error types in translations conducted by Google MT and ChatGPT, from the six categories initially suggested by Koponen. Within these, a collective total of 29 errors were noted, categorized as follows: concept omissions (17.24%), the introduction of irrelevant concepts (24.13%), non-translation of concepts (20.68%), inaccurate concept translations (3.44%), and improper concept substitutions (34.48%). In terms of errors related to relationships, a single error type was noted, specifically the addition of an extra participant. This research sheds light on the operational effectiveness of popular machine translation tools in practical applications, highlighting critical areas for enhancement to improve their precision and dependability, thereby offering significant benefits to both the developers and the users of these technologies (Rusadi & Setiajid, 2023). Supporting this notion, Son and Kim (2023), found that in terms of BLEU, chrF, and TER metrics, Google MT and Microsoft Translator typically outperform ChatGPT. However, ChatGPT demonstrates a higher level of proficiency in translating certain language pairs. It was consistently found that translations from non-English languages to English were more accurate across all three platforms, in contrast to translations from English to non-English languages. Notably, an enhancement in the performance of the translation systems was discernible with the increase in token size,

suggesting that training models on larger tokens could be advantageous. Karabayeva and Kalizhanova (2024) also concluded that while ChatGPT and DeepL serve as valuable tools in the realm of literary translation, their imperfections necessitate human oversight and refinement. This research contributes to the disciplines of machine translation and natural language processing by examining the application of two advanced AI tools, ChatGPT and DeepL, in the translation of literary texts. Furthermore, the paper enriches the fields of literary studies and digital humanities by exploring the capabilities and limitations of machine translation in the context of creative writing and dialogue systems. There are other studies that corroborated with our findings (e.g. Qian, 2023, Huang et al., 2024).

In terms of educational texts, Human translation exhibited exceptional precision and context awareness, skillfully balancing terminology, accuracy, and audience appropriateness. MT tools (Google, DeepL, ChatGPT), on the other hand, demonstrated limitations in consistently capturing the specific nuances and styles required for educational materials, while they generally were accurate and context appropriate. When it comes to the health-related texts, HT accurately captures specialized terminology and maintains the formal style essential in health-related documents, whereas MT tools showed varying degrees of accuracy and appropriateness, with occasional challenges in handling complex phrases and maintaining a consistent formal tone. In legal texts, HT was outstanding in accurately conveying legal terminology and adhering to the formal style and structure of legal documents. While MT tools effectively utilized legal terms, there were instances of deviations in terminology and style, which can be critical in legal contexts.

Consistent with our findings, Sahari et al. (2023) observed both advantages and disadvantages of utilizing ChatGPT for translation tasks. They pointed out that ChatGPT, an AI-driven translation tool, excels in systematic processes such as drafting and editing translations, but is less adept at tasks requiring discernment, like refining and verifying translations. Similarly, the research by Cornelison et al. (2021) aligns with these observations and also

demonstrates the risks of MT tools in health-related texts. The researchers were to evaluate the precision of Google MT in converting usage instructions and counseling points for the most commonly prescribed drugs in the US into Arabic, Chinese (simplified), and Spanish. Out of 247 translations deemed inaccurate, 72 (29.1%) were identified as having high clinical significance or posing a potential threat to life. The researchers recommended employing certified translators to convert prescription medication instructions and counseling points into these languages. They also cautioned clinicians about the risks of relying on Google MT for accurate translations. In support of these findings, a separate investigation evaluated the effectiveness of Google MT's updated algorithm in translating emergency department (ED) discharge instructions into Spanish and Chinese. This study analyzed 100 free-texted ED discharge instructions comprising 647 sentences, finding a 92% accuracy rate for Spanish and 81% for Chinese. Notably, 2% of the Spanish and 8% of the Chinese sentence translations were flagged as potentially harmful. The study advises that clinicians using Google MT can mitigate risks by asking patients to read the translations alongside verbal instructions, paying attention to spelling and grammar, and steering clear of complex grammar, medical terminology, and informal English. It suggests that while Google MT can aid in supplementing English instructions, translated materials should include a disclaimer about possible inaccuracies. The study further recommends the inclusion of English instructions and automated cautions about the limitations of machine translation (Khoong et al., 2019). This aligns with a wealth of related research findings (e.g., Khanna et al., 2011; Wu et al., 2016).

Conclusion

The study contributes significantly to our understanding of the capabilities and limitations of different translation methods. It reinforces the idea that while machine translation tools are rapidly advancing and useful for many applications, human translators remain essential for tasks requiring high precision, specialized knowledge, and cultural sensitivity. The choice of translation method should be guided by the specific demands of

the text, ensuring the highest standards of accuracy and appropriateness are met. We can conclude that human translations tend to provide the highest quality and contextually accurate results, followed closely by DeepL and ChatGPT, with Google MT slightly lagging. However, each method has its strengths and weaknesses, and the choice of translation tool ultimately depends on the specific needs and requirements of the task at hand. Human translations generally provide the highest quality and contextually accurate results but may suffer from inconsistencies and subjectivity. Automated translation tools, such as Google MT, DeepL, and ChatGPT, have made significant progress in delivering fast and accessible translations for various languages. While these tools are effective in many situations, they may still struggle with idiomatic expressions, complex sentences, and preserving the original meaning, especially in highly specialized or technical domains.

It's crucial for practitioners to select the appropriate translation method based on the text type and required accuracy. Human translators are preferable for high-stakes, technical, or nuanced texts. When it comes to researchers, this study provides insights into the current capabilities and limitations of machine translation, offering a basis for further research and development in the field. For technology developers, the findings highlight areas for improvement in machine translation, particularly in handling specialized terminology and maintaining stylistic consistency across diverse text types.

Limitations and Recommendations for Future Research

Every scientific research has limitations in various aspects (Bloomberg & Volpe, 2008; Seidman, 2006). Our study is no exception. The study was limited to three specific text types: educational, health-related, and legal. While these categories are broad, they do not encompass the full range of text types where translation is applied. This limitation could affect the generalizability of the findings to other text types like literary, technical, or marketing materials. We solely focused on translations between English and Turkish. Different language pairs might exhibit unique challenges and strengths in translation,

which this study does not address. Thus, the findings may not be fully applicable to translations involving other language pairs. Given the vast array of cutting-edge MT tools and Generative AI-based translation tools, in this study, we compared human translators with only three machine translation tools (Google MT, DeepL, and ChatGPT). There are other MT tools and methods that were not included, which could offer different results and insights. The performance of human translators, furthermore, can vary significantly based on their experience, expertise, and other subjective factors. The study does not account for this variability, potentially affecting the reliability and consistency of the human translation results. The study primarily concentrated on textual content, with less emphasis on design and markup elements in translation. This focus overlooks an important aspect of translation, especially for materials where layout and design play a crucial role in conveying meaning. Translations were analyzed in a controlled setting, which may not accurately reflect how these translations perform in real-world scenarios where context, user interpretation, and situational nuances play a significant role. The field of machine translation is rapidly evolving. The findings of this study are time-bound and may not be applicable in the near future as new technologies and updates to existing tools emerge. The study primarily used the MQM model for a quantitative assessment of translations. It did not incorporate qualitative feedback from end-users or subject matter experts, which could provide deeper insights into the practical effectiveness and user perception of the translations. While the study provides valuable insights into the translation performance of human translators and selected machine translation tools, these limitations suggest caution in generalizing the findings. Future research should aim to address these gaps to gain a more comprehensive understanding of translation performance across various contexts and applications.

Based on the findings of this study, we have written some recommendations for future research. These recommendations aim to guide targeted research efforts that address specific limitations identified in current machine translation tools. By

focusing on these areas, future developments can make significant strides in enhancing the accuracy, reliability, and applicability of machine translations across various fields. In our study, there is evidence that MT tools like Google MT, DeepL, and ChatGPT showed inconsistencies in translating specialized terminology, particularly in legal and health-related texts. Future research can delve into enhancing machine translation tools with domain-specific knowledge that could improve accuracy in specialized terminology, by implementing and testing translation models trained specifically on legal and medical datasets to assess improvements in terminology accuracy. We also concluded that MT tools were less effective in replicating the formal and specific stylistic requirements of legal documents. Accordingly, future research can dive into how to enable MT tools to benefit from algorithms that adapt to the stylistic nuances of different text genres, thereby developing algorithms that can identify and adapt to various legal document styles and evaluate their impact on translation fidelity. One of our most significant findings is that human translators showed a superior ability to understand and translate complex contexts and cultural nuances. Future research can focus on mimicking human-like context analysis in AI could bridge the current gap in machine translation. This necessitates experimenting with deep learning techniques that focus on contextual and cultural understanding, possibly in collaboration with cross-cultural studies. We found that MT tools demonstrated variable accuracy and consistency in educational text translations. Future research can examine how to enhance the accuracy of machine translations for educational purposes and can broaden their application in academic settings, by training and testing machine translation models on diverse educational content, emphasizing accuracy and consistency in conveying educational information. Another important finding of us is that MT tools often lack in addressing audience appropriateness and conveying cultural nuances, unlike human translators. Future research can incorporate sociolinguistic factors that could improve audience-targeted and culturally sensitive translations, thereby integrating sociolinguistic parameters into AI models and assessing their

effectiveness in diverse cultural and audience-specific translations. These recommendations aim to guide targeted research efforts that address specific limitations identified in current machine translation tools. By focusing on these areas, future developments can make significant strides in enhancing the accuracy, reliability, and applicability of machine translations across various fields.

References

Agung, I. G. A. M., Budiartha, P. G., & Suryani, N. W. (2024, January). Translation performance of Google Translate and DeepL in translating Indonesian short stories into English. *In Proceedings: Linguistics, Literature, Culture and Arts International Seminar (LITERATES)* (pp. 178-185).

Ali, G., Ali, N., Syed, K. (2023). Understanding shifting paradigms of translation studies in 21st century.

Almahasees, Z. (2021). *Analyzing English-Arabic machine translation: Google Translate, Microsoft Translator and Sakhr*. Routledge.

Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Banerjee, S., Lavie, A. (2005). 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 (pp. 65-72).

Bansal, R., Samanta, B., Dalmia, S., Gupta, N., Vashishth, S., Ganapathy, S., ..., Talukdar, P. (2024). LLM Augmented LLMs: Expanding Capabilities through Composition. arXiv preprint arXiv:2401.02412.

Blain, F., Senellart, J., Schwenk, H., Plitt, M., Roturier, J. (2011). Qualitative analysis of post-editing for high quality machine

translation. In Proceedings of Machine Translation Summit XIII: Papers.

Bloomberg, L. D., Volpe, M. (2008). *Completing your qualitative dissertation: A roadmap from beginning to end*. London: SAGE.

Bowker, L. (2023). De-mystifying translation: Introducing translation to non-translators (p. 217). Taylor & Francis.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., ..., Mercer, R. L. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2), p.79-85.

Callison-Burch, C., Osborne, M., Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In Proceedings of the workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization at the 2006 conference of the North American chapter of the association for computational linguistics (pp. 29-36).

Comelles, E., Arranz, V., Castellón, I. (2017). Guiding automatic MT evaluation by means of linguistic features. *Digital Scholarship in the Humanities,* 32(4), p.761-778.

Costa, A., Ling, W., Luis, T., Correia, R., Coheur, L. (2015). A Linguistically Motivated Taxonomy for Machine Translation Error Analysis. *Machine Translation,* 29(2), p.127–161.

Girletti, S., Lefer, M. A. (2024). Introducing MTPE pricing in translator training: a concrete proposal for MT instructors. *The Interpreter and Translator Trainer*, p.1-18.

Huang, X., Zhang, Z., Geng, X., Du, Y., Chen, J., Huang, S. (2024). Lost in the source language: How large language models evaluate the quality of machine translation. arXiv preprint arXiv:2401.06568.

Hutchins, W. J. (2003). *Machine translation: Past, present, future*. Research Studies Press Ltd.

Karabayeva, I.,Kalizhanova, A. (2024). Evaluating machine translation of literature through rhetorical analysis. *Journal of Translation and Language Studies*, 5(1), p.1-9.

Khanna, R. R., Karliner, L. S., Eck, M., Vittinghoff, E., Koenig, C. J., Fang, M. C. (2011). Performance of an online translation tool when applied to patient educational material. *Journal of Hospital Medicine*, 6(9), p.519-525.

Khoong, E. C., Steinbrook, E., Brown, C., Fernandez, A. (2019). Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA Internal Medicine*, 179(4), p.580-582.

Khoshafah, F. (2023). ChatGPT for Arabic-English translation: Evaluating the accuracy. *Research Square.*

Koehn, P. (2010). *Statistical machine translation.* Cambridge University Press.

Kunchukuttan, A., Bhattacharyya, P. (2021). *Machine translation and transliteration involving related, low-resource languages*. CRC Press.

Lauscher, S. (2000). *Assessing the quality of translations: a practical guide for users*. Manchester: St. Jerome Publishing.

Li, B., Weng, Y., Xia, F., Deng, H. (2024). Towards better Chinese-centric neural machine translation for low-resource languages. *Computer Speech & Language*, 84, 101566.

Li, J., Dada, A., Puladi, B., Kleesiek, J., Egger, J. (2024). ChatGPT in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, 108013.

Lommel, A. (2018). Metrics for translation quality assessment: a case for standardizing error typologies. Translation quality assessment: From principles to practice, p.109-127.

Lommel, A. R., Uszkoreit, H. Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics.

Revista Tradumàtica: Traducció i Tecnologies de la Informació i la Comunicació, 12, p.455–463.

Mellinger, C. D., Hanson, T. A. (2016). *Quantitative research methods in translation and interpreting studies*. Taylor & Francis.

Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318).

Popović, M. (2018). Error classification and analysis for machine translation quality assessment. In Translation quality assessment (pp. 129-158). Springer, Cham.

Qian, M. (2023). Performance evaluation on human-machine teaming augmented machine translation enabled by GPT-4. In Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications (pp. 20-31).

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Reiss, K. (1971). Möglichkeiten und grenzen der übersetzungskritik: Kategorien und kriterien für eine sachgerechte beurteilung von übersetzungen. O. Schwarz.

Rusadi, A. M., Setiajid, H. H. (2023). Evaluating the accuracy of google translate and chatgpt in translating windows 11 education installation guı texts to indonesian: an application of koponen's error category. In English Language and Literature International Conference (ELLiC) Proceedings (pp. 698-713).

Sahari, Y., Al-Kadi, A. M. T., Ali, J. K. M. (2023). A Cross Sectional study of ChatGPT in translation: magnitude of use, attitudes, and uncertainties. *Journal of Psycholinguistic Research*, p.1-18.

Seidman, I. (2006). *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. New York, NY: Teachers College.

Siu, S. C. (2023). ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. Available at SSRN 4448091.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of Association for Machine Translation in the Americas (AMTA) (pp. 223-231).

Snow, T. A. (2015). *Establishing the viability of the multidimensional quality metrics framework.* Brigham Young University.

Son, J., Kim, B. (2023). Translation performance from the user's perspective of large language models and neural machine translation systems. *Information*, 14(10),p.574.

Stymne, S., Ahrenberg, L. (2012, May). On the practice of error analysis for machine translation evaluation. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) (pp. 1785-1790).

Sutskever, I., Vinyals, O., Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

Vilar, D., Xu, J., d'Haro, L. F., Ney, H. (2006). Error analysis of statistical machine translation output. In Proceedings of the fifth international conference on language resources and evaluation (LREC'06).

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.